Learning to Decode Cognitive States from Brain Images

Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, Francisco Pereira and Xuerui Wang
Computer Science Department, Carnegie Mellon University

Marcel Just and Sharlene Newman
Psychology Department, Carnegie Mellon University

April 29, 2003

Abstract.
Over the past decade, functional Magnetic Resonance Imaging (fMRI) has emerged as a powerful new instrument to collect vast quantities of data about activity in the human brain. A typical fMRI experiment can produce a three-dimensional image related to the human subject's brain activity every half second, at a spatial resolution of a few millimeters. As in other modern empirical sciences, this new instrumentation has led to a flood of new data, and a corresponding need for new data analysis methods. We describe recent research applying machine learning methods to the problem of classifying the cognitive state of a human subject based on fRMI data observed over a single time interval. In particular, we present case studies in which we have successfully trained classifiers to distinguish cognitive states such as (1) whether the human subject is looking at a picture or a sentence, (2) whether the subject is reading an ambiguous or non-ambiguous sentence, (3) whether the word the subject is viewing is a noun or a verb, and (4) whether the noun the subject is viewing is a word describing food, people, buildings, etc. This learning problem provides an interesting case study of classifier learning from extremely high dimensional (10^5 features), extremely sparse (tens of training examples), noisy data. This paper summarizes the results obtained in these four case studies, as well as lessons learned about how to successfully apply machine learning methods to train classifiers in such settings.

Keywords: Scientific data analysis, functional Magnetic Resonance Imaging, High dimensional data, Feature selection, Bayesian classifier, Support Vector Machine, Nearest neighbor, Brain image analysis

1. Introduction

The study of human brain function has received a tremendous boost in recent years from the advent of functional Magnetic Resonance Imaging (fMRI), a brain imaging method that dramatically improves our ability to observe correlates of neural brain activity in human subjects at high spatial resolution (several millimeters), across the entire brain. This fMRI technology offers the promise of revolutionary new approaches to studying human cognitive processes, provided we can develop appropriate data analysis methods to make sense of this huge volume of data. A twenty-minute fMRI session with a single human subject

produces a series of three dimensional brain images each containing approximately 15,000 voxels, collected once per second, yielding tens of millions of data observations.

Since its advent, fMRI has been used to conduct hundreds of studies that identify specific regions of the brain that are activated on average when a human performs a particular cognitive function (e.g., reading, mental imagery). The vast majority of this published work reports descriptive statistics of brain activity, calculated by *averaging together* fMRI data collected over multiple time intervals, in which the subject responds to repeated stimuli of some type (e.g., reading a variety of words).

In this paper we consider a different goal: training machine learning classifiers to automatically decode the subject's cognitive state, given just his/her fMRI activity at a single time instant or time interval. We describe here several case studies, such as training the system to distinguish whether the word a subject is currently processing is a noun or a verb.

This goal of training classifiers to detect cognitive states is important because such classifiers could provide the basis for new approaches to studying human reasoning processes in both normal and abnormal populations. Put succinctly, such classifiers would constitute *virtual sensors* of the subject's cognitive state, which could be useful to scientists and clinicians across a range of cognitive science research and diagnostic medical applications.

This problem is also quite interesting from the perspective of machine learning, because it provides a case study of classifier learning from extremely high dimensional, sparse, and noisy data. In our case studies we encounter problems where the examples are described by 100,000 features, and where we have less than a dozen, very noisy, training examples per class. Although conventional wisdom might suggest classifier learning would be impossible in such extreme settings, in fact we have found it is possible in this case, by design of appropriate feature abstraction and classifier training methods tuned to these problem characteristics.

In this paper we first provide a brief introduction to fMRI, then describe several fMRI data sets we have analyzed, the machine learning approaches we explored, and lessons learned about how best to apply machine learning approaches to the problem of classifying cognitive states based on single interval fMRI data.

2. Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a technique for obtaining three-dimensional images related to activity in the brain through time. More precisely, fMRI measures the ratio of oxygenated hemoglobin to deoxygenated hemoglobin in the blood with respect to a control baseline, at many individual locations within the brain. It is widely believed that blood oxygen level is influenced by local neural activity, and hence this blood oxygen level dependent (BOLD) response is generally taken as an indicator of neural activity.

An fMRI scanner measures the value of the fMRI signal (BOLD response) at all the points in a three dimensional grid, or *image*), covering part of the brain. In the studies described in this paper, a three dimensional image is captured every 1, 1.5, or 0.5 seconds. We refer to the cells within an image as *voxels* (volume elements). The voxels in a typical fMRI study have a volume of a few tens of cubic millimeters, and a typical three dimensional image typically contains tens of thousands of voxels, 10,000 to 15,000 of which contain cortical matter and are thus of interest. While the spatial resolution of fMRI is dramatically better than that provided by earlier brain imaging methods, each voxel nevertheless contains on the order of hundreds of thousands of neurons.

The temporal response of the fMRI BOLD signal is smeared over several seconds. Given an impulse stimulus such as a flash of patterned light, the fMRI BOLD response increases to a maximum after approximately four to five seconds, typically returning to baseline levels after another five to ten seconds. Despite this prolonged temporal response, researchers have found that the relative timing of events can be resolved to within a few tens of milliseconds (e.g. to distinguish the relative timing of two flashes of light - one in the left eye and one in the right eye - as in (Menon et al., 1998)).

A small portion of fMRI data is illustrated in Figure 1. This figure shows data collected over a fifteen second interval during which the subject was read a word, decided whether it was a noun or verb (in this case, it was a verb), then waited for another word. This data was sampled once per second for fifteen seconds, over sixteen planar slices, one of which is shown in the figure.

3. Related Work Analyzing fMRI Data

Over recent years there has been a growing interest within the computer science community in data processing for fMRI. One popular style of processing involves using a Generalized Linear Model (GLM) approach

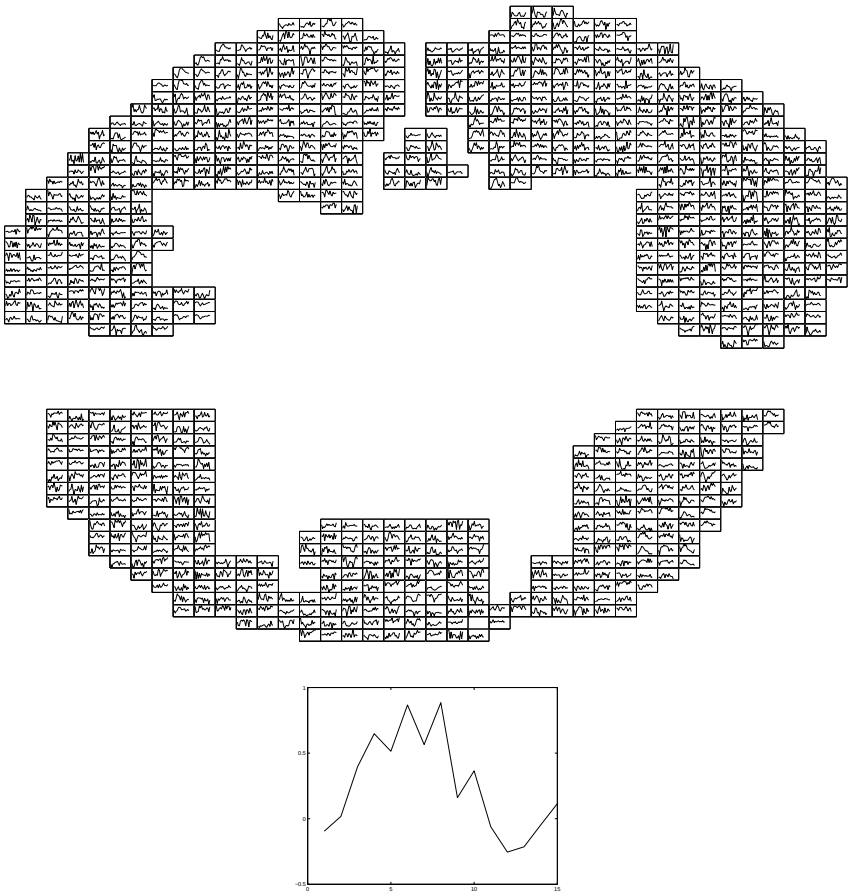


Figure 1. Typical fMRI data. The top portion of the figure shows fMRI data for a selected set of voxels in the cortex, from a two-dimensional image plane through the brain. A fifteen second interval of fMRI data is plotted at each voxel location. The anterior portion of the brain is at the top of the figure, posterior at bottom. The left side of the brain is shown on the right, according to standard radiological convention. The full three-dimensional brain image consists of sixteen such image planes. The bottom portion of the figure shows one of these plots in greater detail. During this interval the subject was presented a word, answered whether the word was a noun or verb, then waited for another word.

(Friston et al., 1995)(Bly, 2001), in which a regression is performed on the signal value at a voxel with respect to some stimulus property in order to determine whether the voxel's activity is related to the stimulus. Others have used t-statistics to determine relevant active voxels, and yet others have used more complex statistical methods to estimate parameters of the BOLD response in the presence of noise (Genovese, 1999).

Various methods for modelling time series data have been used for fMRI data. For example, (Hojen-Sorensen et al., 1999) used Hidden Markov Models (HMM) to learn a model of activity in the visual cortex resulting from a flashing light stimulus. Although the program was not told the stimulus, the on-off stimulus was recovered as the hidden state by the HMM.

A variety of unsupervised learning methods have also been used for exploratory analysis of fMRI data. For example, (Goutte et al., 1998) discussed the use of clustering methods for fMRI data. One particular approach (Penny, 2001) involved the application of Expectation Maximization to estimate mixture models to cluster the data. Others have used Principle Components Analysis and Independent Components Analysis (McKeown et al., 1998) to determine spatial-temporal factors that can be linearly combined to reconstruct the fMRI signal.

While there has been little work on our specific problem of training classifiers to decode cognitive states, there are several papers describing work with closely related goals. For example, (Haxby et al., 2001) showed that different patterns of fMRI activity are generated when a human subject views a photograph of a face versus a house, versus a shoe, versus a chair. While they did not specifically use these discovered patterns to classify subsequent single-event data, they did report that by dividing the fMRI data for each photograph category into two samples, they could automatically match the sample means related to the same category. Others (Wagner et al., 1998) reported that they have been able to make better-than-random predictions regarding whether a visually presented word will be remembered later, based on the magnitude of activity within certain parts of left prefrontal and temporal cortices during that presentation.

In addition to work on fMRI, there has been related recent work applying machine learning methods to data from other devices measuring brain activity. For example, (Blankertz et al., 2002) describe experiments training classifiers for single trial EEG data.

4. Approach

This section briefly describes our approach to data preprocessing, training classifiers, and evaluating them.

4.1. Data Acquisition and Preprocessing

In the fMRI studies considered here, data were collected from normal students from the university community. Typical studies involved between five and fifteen subjects, and we generally selected a subset of

these subjects with the strongest, least noisy fMRI signal to train our classifiers. Data were preprocessed to remove artifacts due to head motion, signal drift, and other sources, using the FIASCO program (Eddy et al., 1998)[1]. All voxel activity values were represented by the percent difference from their mean value during fixation (rest) conditions. These preprocessed images were used as input to our classifiers.

In several cases, we found it useful to identify specific anatomically defined regions of interest (ROIs) within the brain of each subject. To achieve this, two types of brain images were collected for each subject. One type of image, capturing brain activation via the BOLD response, is referred to as a *functional image*. The second type of image, called a *structural image*, reveals the static physical brain structure at higher resolution. For each subject, this structural image was used to identify the anatomical regions of interest, using the parcellation scheme of (Caviness et al., 1996) and (Rademacher et al., 1992). For each subject, the mean of their functional images was then co-registered to the structural image, so that individual voxels in the functional images could be associated with the ROIs identified in the structural image.

4.2. Learning Methods

In this paper we explore the use of machine learning methods to approximate classification functions of the following form

$$f : \text{fMRI-sequence}(t_1, t_2) \rightarrow \text{CognitiveState}$$

where fMRI-sequence(t_1, t_2) is the sequence of fMRI images collected during the contiguous time interval $[t_1, t_2]$, and where CognitiveState is the set of cognitive states to be discriminated.

We explored a variety of methods for encoding fMRI-sequence(t_1, t_2) as input to the classifier. In some cases, we encoded it as a vector of features, one for each voxel at each time in the interval $[t_1, t_2]$. This can be an extremely high dimensional feature vector, consisting of hundreds of thousands of features given that a typical image contains 10,000 to 15,000 voxels, and a training example can include dozens of images. Therefore, we explored a variety of approaches to reducing the dimension of this feature vector, including methods for feature selection, as well as methods that replace multiple feature values by their mean. These feature selection and feature abstraction methods are described in detail in Section 6.3.

We explored a number of classifier training methods, including:

[1] FIASCO is available at http://www.stat.cmu.edu/~fiasco.

– *Gaussian Naive Bayes (GNB).* The GNB classifier uses the training data to estimate the probability distribution over fMRI observations, conditioned on the subject's cognitive state. It then classifies a new example $X = \langle x_1 \ldots x_n \rangle$ by estimating the probability $P(c_i|X)$ of cognitive state c_i, using Bayes rule along with the assumption that the features x_i are conditionally independent given the class:

$$\hat{P}(c_i|X) = \frac{\hat{P}(c_i) \prod_j \hat{P}(x_j|c_i)}{\sum_k [\hat{P}(c_k) \prod_j \hat{P}(x_j|c_k)]}$$

where $\hat{P}$ denotes distributions estimated by GNB. Each distribution of the form $\hat{P}(x_j|c_i)$ is modelled as a Gaussian, using maximum likelihood estimates of the mean and variance derived from the training data. Distributions of the form $\hat{P}(c_i)$ are modelled as Bernoulli, again using maximum likelihood estimates based on training data. Given a new example to be classified, the GNB outputs posterior probabilities for each cognitive state, calculated using the above formula.

– *Support Vector Machine (SVM).* We used a linear kernel Support Vector Machine (see, for instance, (Burges, 1998)).

– *k Nearest Neighbor(kNN).* We use k Nearest Neighbor with a Euclidean distance metric, considering values of 1, 3, 5, 7, and 9 for k (see, for instance (Mitchell, 1997)).

4.3. Evaluating Results

Trained classifiers are evaluated by their cross-validated classification error when learning boolean-valued classification functions. When more than two classes are involved, the classifier outputs a rank-ordered list of the potential classes from most to least likely. In this case, we score the success of each prediction by the normalized rank of the correct class in this sorted list. Thus, the normalized rank error ranges from 0 when the correct class is ranked most likely, to 1 when it is ranked least likely. Note that random guessing yields an expected normalized rank error of 0.5.

To evaluate classifiers, we generally employ k-fold cross-validation, leaving out one example per class on each fold. In the data sets considered in this paper, the competing classes are balanced (i.e., the number of available examples is the same for each competing class). Thus, by leaving out one example per class we retain a balanced training

set for each fold, which correctly reflects the class priors. In fact, we found that when training data was especially sparse, this leave-out-one-example-per-class approach sometimes significantly outperformed a leave-out-one-example approach, and that the latter generally yielded pessimistic error estimates.

Because the fMRI BOLD response lasts for several seconds, a strict leave-out-one-example-per-class evaluation can sometimes produce optimistic estimates of the true classifier error. The reason is straightforward: when holding out a test image occurring at time t, the training images at times $t+1$ and $t-1$ will be highly correlated with this test image. Therefore, if the images at $t-1$ and $t+1$ belong to the same class as the image at t, this leads to optimistically biased error estimates for the held out example. When faced with this situation (i.e., in the Semantic Categories study described below), we avoid the optimistic bias by removing from the training set all images that occur within 5 seconds of the held out test image. In this case, our cross validation procedure involves holding out one test example per class, and also removing temporally proximate images from the training set.

5. Case Studies

This section describes four distinct fMRI studies, the data collected in each, and the classifiers trained for each. In this section we summarize the success of the best classifier obtained for each of these studies. The subsequent section discusses more generally the lessons learned across these four case studies.

5.1. Picture versus Sentence Study

In this fMRI study (Keller et al., 2001), subjects went through a sequence of trials, during which they were first shown a sentence and a simple picture, then answered whether the sentence correctly described the picture. We used this data to explore the feasibility of training classifiers to distinguish whether the subject is examining a sentence or a picture during a particular time interval.

In half of the trials the picture was presented first, followed by the sentence. In the remaining trials, the sentence was presented first, followed by the picture. In either case, the first stimulus (sentence or picture) was presented for 4 seconds, followed by a blank screen for 4 seconds. The second stimulus was then presented for up to 4 seconds, ending when the subject pressed the mouse button to indicate whether the sentence correctly described the picture. Finally, a rest or fixation

period of 15 seconds was inserted before the next trial began. Thus, each trial lasted approximately 27 seconds. Pictures were geometric arrangements of the symbols +, * and/or $, such as

$$\begin{array}{c} + \\ \hline * \end{array}$$

Sentences were descriptions such as "It is true that the plus is below the dollar." Half of the sentences were negated (e.g., "It is not true that the star is above the plus.") and the other half were affirmative sentences.

Each subject was presented a total of 40 trials as described above, interspersed with ten fixation periods. In each fixation period the subject simply stared at a fixed point on the screen. fMRI images were collected every 500 msec.

The learning task we consider for this study is to train a classifier to determine, given a particular 8-second interval of fMRI data, whether the subject is viewing a sentence or a picture during this interval. In other words, we wish to learn a separate classifier for each subject, of the following form

$$f : \text{fMRI-sequence}(t_0, t_0 + 8) \rightarrow \{\text{Picture, Sentence}\}$$

where t_0 is the time of stimulus (picture or sentence) onset. The fMRI-sequence was described by the activities of all voxels appearing in 7 distinct ROIs. These 7 ROIs were selected as most likely to be relevant by a domain expert, and contained a total of 1397 to 2864 voxels per subject, varying due to differences in brain structure from one subject to another. Note that the eight second interval considered by the classifier contains 16 images (captured twice per second), yielding an input feature vector containing from 22,352 to 45,824 features, depending on the human subject.

The expected classification error for the default classifier (guessing the most common class) is 0.50 in this case. The average error obtained for the most successful trained classifier, using the most successful feature selection strategy, was 0.09, over 13 subjects, with the best subject reaching 0.01 (refer to Section 6.2 for more details). These results are statistically highly significant, and indicate that it is indeed possible to train classifiers to distinguish these two cognitive states reliably.

In addition to these single-subject classifiers, we also experimented with training classifiers that operate across multiple subjects. In this case, we evaluated the classification error using a leave-one-subject-out regime in which we held out each of the 13 subjects in turn while training on the other 12. The mean error over the held out subject

for the most successful combination of feature selection and classifier was 0.25. Again, this is significantly better than the expected 0.5 error from random guessing, indicating that it is possible to train classifiers for this task that operate on human subjects who where not part of the training set. These results are described in detail in Section 6.4.

5.2. Syntactic Ambiguity Study

In this fMRI study (see (Mason et al., in press)) subjects were presented with two types of ambiguous sentences and two types of unambiguous sentences, and were asked to respond to a yes-no question about the content of each sentence. The questions were designed to ensure that the subject was in fact processing the sentence. The learning task for this study was to distinguish whether the subject was currently reading the least ambiguous or the most ambiguous type of sentence. An example of the most ambiguous type of sentence is "The experienced soldiers warned about the dangers conducted the midnight raid." An example of the least ambiguous type of sentence is "The experienced soldiers spoke about the dangers before the midnight raid."

Ten sentences of each of type were presented to each subject. Each sentence was presented for 10 seconds. Next a question was presented, and the subject was given 4 seconds to answer. After the subject answered the question, or 4 seconds elapsed, an "X" appeared on the screen for a 12 second rest period. The scanner collected one image every 1.5 seconds.

We are interested here in learning a classifier that takes as input an interval of fMRI activity, and determines which of the two types of sentence the subject is reading. Using our earlier notation, for each subject we trained classifiers of the form

$$f : \text{fMRI-sequence}(t_0 + 4.5, t_0 + 15) \rightarrow \text{SentenceType}$$

where SentenceType = {Ambiguous, Unambiguous}, and where t_0 is the time at which the sentence is first presented to the subject. Note the classifier input describes fMRI activity during the interval from 4.5 to 15 seconds following initial presentation of the sentence. This is the interval during which the fMRI activity is most intense. In this case we also reduced the set of voxels considered to those in 4 ROIs considered to be most relevant by a domain expert. These 4 ROIs contained a total of 1500 to 3508 voxels, depending on the subject.

The expected classification error from random guessing in this case is 0.50, given the equal number of examples from both classes. The average error obtained by the most successful combination of feature

selection and classifier is 0.21, over 5 subjects, with the best subject reaching 0.10 (refer to 6.2 for more details).

5.3. Noun versus Verb Study

In this study, subjects were shown single words, and answered for each whether it was a verb or a noun. We used this data to explore the feasibility of training classifiers to distinguish whether the subject is considering a verb or noun.

In each trial, the word was presented for 1 second, followed by a blank screen so that the total trial length was 16 seconds. A total of 16 trials, 8 of each kind of word, were presented to each subject (trials 9 through 16 used the same words as trials 1 through 8, presented in a different randomized sequence). fMRI images were captured once per second.

In this case, we trained classifiers to discriminate whether the subject is considering a noun or a verb, based on observations during a subinterval of the trial:

$$f : \text{fMRI-sequence}(t_0 + 5, t_0 + 13) \rightarrow \text{WordType}$$

where WordType = {Noun, Verb}, and where t_0 is the time when the word is first presented. This time interval was selected to capture the peak fMRI response for each word. In this case, all voxels from 30 ROIs were included, yielding a total of 8,586 to 10,899 voxels, depending on the subject. Given the nine images collected during each example, this forms a classifier whose input contains between 77,274 and 98,091 features, and which is trained from 16 examples.

Given the equal number of noun and verb trials, the expected error from random guessing is again 0.50. The average error obtained by the most successful combination of feature selection and classifier is 0.23, over four subjects, with the best subject reaching 0.19 (refer to 6.2 for details).

5.4. Semantic Categories Study

In this study, 10 subjects were presented with individual nouns belonging to twelve distinct semantic categories (e.g., Fruits, Tools), and asked to determine whether the word belonged to a particular category. We used this data to explore the feasibility of training classifiers to detect which of the semantic categories of word the subject was examining.

The trials in this study were divided into twelve blocks. In each block, the name of a semantic category was first displayed for 2 seconds. Following this, the subject was shown a succession of 20 words, each

presented for 400 msec and followed by 1200 msec of blank screen. After each word was presented, the subject clicked a mouse button to indicate whether the word belonged to the semantic category named at the beginning of the block. In fact, nearly all words belonged to the named category (half the blocks contained no out-of-category words, and the remaining blocks contained just one out-of-category word). A multi-second blank screen rest period was inserted between each of the twelve blocks. The twelve semantic categories of words presented were "fish," "four-legged animals," "trees," "flowers," "fruits," "vegetables," "family members," "occupations," "tools," "kitchen items," "dwellings," and "building parts." Words were chosen from lists of high frequency words of each category, as given in (Battig and Montague, 1968), in order to avoid obscure or multiple-meaning words. fMRI images were acquired once per second.

The learning task we considered for this study is to distinguish which of the twelve semantic categories the subject is considering, based on a single observed fMRI image. Following our earlier notation, we wish to learn a classifier of the form:

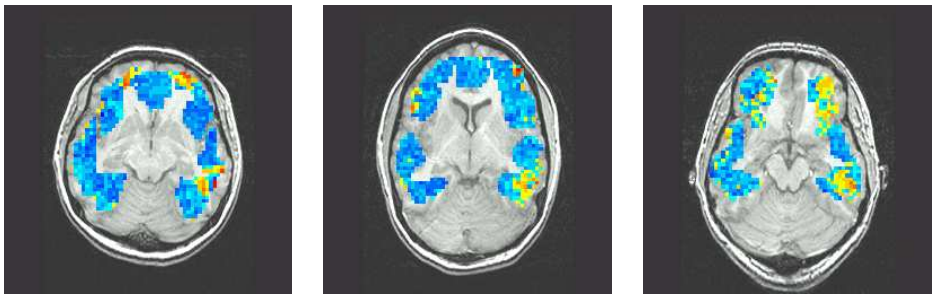
$$f : \text{fMRI}(t) \rightarrow \text{WordCategory}$$

where fMRI(t) is a single fMRI image, and where WordCategory is the set of 12 semantic categories described above. All voxels from 30 ROIs were used, yielding a total of 8,470 to 11,136 voxels, depending on the subject.

The trained classifier outputs a rank-ordered list of the 12 categories, ranked from most to least probable. We therefore evaluate classifier error using the normalized rank error described in Section 4, where random guessing gives an expected normalized rank error of 0.50. The normalized rank error for the most successful combination of feature selection and classifier is 0.08 (i.e. on average the correct word category was ranked first or second out of the twelve categories), over 10 subjects, with the best subject reaching 0.04. (please refer to 6.2 for more details).

One reasonable question that can be raised regarding these classifier results is whether the classifier is indeed learning the pattern of brain activity predictive of semantic categories, or whether it is instead learning patterns related to some other time-varying phenomenon that influences fMRI activation. One unfortunate property of the experimental protocol for collecting data, from this point of view, is that all of the words belonging to a single category are presented within a single time interval (i.e., a single experiment block). In fact we do believe this temporal adjacency may be influencing our results, but we also believe the classifier is indeed capturing regularities primarily related to semantic categories. One strong piece of supporting evidence is that

classifiers trained for different human subjects tend to rely on the same brain locations to make their predictions, and that these regions have been reported by others as related to semantic categorization. Figure 2 illustrates the brain regions containing the most informative fMRI signal for classification, across three subjects. Note the highly discriminating voxels are clustered together, in similar regions across these subjects. These locations for discriminability match those reported in earlier work on semantic categorization by (Chao et al., 1999),(Chao et al., 2002), (Ishai et al., 1999) and (Aguirre et al., 1998), as well as some novel areas that are currently under investigation.



a. b. c.

Figure 2. Color plots show locations of voxels that best predict the word semantic category, for three different subjects. For each voxel, the color indicates the normalized rank error over the test set, for a GNB classifier based on this single voxel. Note the spatial clustering of highly predictive voxels, and the similar regions of predictability across these three subjects. The range of normalized rank errors is [Red$\approx$ 0.25,Dark Blue$\approx$ 0.6], with other colors intermediate between these two extremes.

6. Lessons Learned

6.1. Can One Learn to Decode Mental States from fMRI?

The primary goal leading to this research was to determine whether it is feasible to use machine learning methods to decode mental states from single interval fMRI data. The successful results reported above for all four data sets indicate that this is indeed feasible in a variety of interesting cases. However, it is important to note that while our empirical results demonstrate the ability to successfully distinguish among a predefined set of states occurring at specific times while the subject performs specific tasks, they do not yet demonstrate that trained classifiers can reliably detect cognitive states occurring at arbitrary times while the subject performs arbitrary tasks. While our current results

may already be of use in cognitive science research, we intend to pursue this more general goal in future work.

We also attempted but failed to train successful classifiers for several other classification functions over these same data sets. For example, we were unable to train an accurate classifier to distinguish the processing of negated versus affirmative sentences in the Picture versus Sentence study, or to distinguish the processing of true versus false sentences. We were also unable to train classifiers to distinguish the exact word the subject was reading in the Noun versus Verb study. It may be that these failures could be reversed given larger training sets or more effective learning algorithms. Alternatively, it may be the case that the fMRI data simply lacks the information needed to make these distinctions. This line of research is still very new, and while the above results demonstrate the feasibility of discriminating a variety of cognitive states based on fMRI, at this point the question of exactly which cognitive states can be reliably decoded remains an open empirical question. However, given our initial successes, likely advances in brain imaging technology, and likely progress in developing machine learning methods specifically for this type of application, we are optimistic that over time we will be able to decode an increasingly useful collection of cognitive states in an increasingly open ended set of experimental settings.

6.2. Which Classifier Works Best?

As discussed earlier, we experimented with three classifier learning methods: a Gaussian Naive Bayes (GNB) classifier, k-nearest neighbor (kNN), and linear Support Vector Machines (SVM). These classifiers were selected because they have been used successfully in other applications involving high dimensional data. For example, Naive Bayes classifiers, kNN, and SVM have all been used for text classification problems (Nigam et al., 2000; Joachims, 2001; Yang, 1999), where the dimension of the data is approximately 10^5, corresponding to the size of the natural language vocabulary.

In considering these classifiers, one interesting and relevant relationship to consider is that GNB is a *generative classifier* (i.e., it learns a function $f : X \rightarrow Y$ by instead directly modelling $P(X|Y)$ and $P(Y)$), whereas SVM is a *discriminative classifier* (i.e., it learns f by estimating parameters that map X directly to Y). As discussed in (Ng and Jordan, 2002), the relative performance of generative and discriminative classifiers can depend strongly on the number of training examples available and the dimension of the data. More specifically, they consider the relationship between the generative classifier GNB

Table I. Comparison of Classification Methods. The numbers in the table are test errors averaged over all single-subject classifiers trained for each study. The rows with Feature Selection "All" show the result of using all voxels within the available ROIs. The results in the second row for each study are obtained using the feature selection method that performed best with GNB for that study. This same method performed best with SVM's as well, except in the case of Active feature selection in the Syntactic Ambiguity study, where the ROIActiveAvg feature selection method achieved a small but statistically insignificant improvement for the SVM.

Study	Feature Selection	GNB	SVM	1NN	3NN	5NN	7NN	9NN
Picture vs	All	0.29	0.32	0.43	0.41	0.37	0.37	0.33
Sentence	Active	0.16	0.09	0.20	0.18	0.19	0.18	0.17
Semantic	All	0.10	N/A	0.40	0.40	0.40	0.40	0.25
Categories	Active	0.08	N/A	0.30	0.20	0.16	0.14	0.13
Syntactic	All	0.43	0.38	0.50	0.46	0.47	0.39	0.43
Ambiguity	Active	0.25	0.23	0.29	0.29	0.28	0.29	0.26
Noun vs	All	0.36	0.39	0.44	0.45	0.39	0.44	0.41
Verb	ROIActiveAvg	0.23	0.28	0.38	0.38	0.33	0.28	0.31

and its discriminative counterpart, Logistic regression. The discriminative Logistic regression is usually preferred over GNB when training data are plentiful, because it does not make the restrictive conditional independence assumptions of GNB and therefore asymptotically it can better model the target function. However, (Ng and Jordan, 2002) show that given a feature space of dimension n, the number of examples needed to reach the asymptotic error rate for the generative GNB is $O(log(n))$, whereas for the discriminative Logistic regression it is $O(n)$. Because of this fact, one might expect that even if the discriminative classifier outperforms the generative classifier asymptotically (in the number of training examples), the generative classifier might outperform the discriminative classifier when training data is very sparse. Ng and Jordan show exactly this empirical result over a number of data sets. Furthermore, one might expect feature selection methods, which reduce the dimension n, to potentially benefit the discriminative classifier more than the generative classifier.

While our linear SVM classifier is different from Logistic regression, it is also a linear discriminative classifier that converges toward its asymptotic error rate in $O(n)$ examples, and hence we expect the number of training examples and dimension of the data to have a similar influence on the relative performance of GNB and SVM.

To test the relative performance of our classifiers, we performed two sets of experiments. First, for each study we analyzed the performance of GNB, linear SVM, and kNN (with $k \in \{1, 3, 5, 7, 9\}$) using as input to the classifier *all* voxels in the ROIs selected for those studies. Here the performance metric is classification error, except for the Semantic Categories study where the metric is average normalized rank error. The performance reported for a specific study is the average over all single-subject classifiers trained for that study, as obtained by leave-one-example out from each class. Because the Semantic Categories study is not a binary classification task, we did not experiment with SVMs on this specific study.

The results are shown in Table I (first line for each study). As can be seen here, the GNB and SVM classifiers consistently outperform kNN. Examining the performance of kNN, one can also see a trend that performance generally improves with increasing values of k.

Our second set of experiments examined the performance of the classifiers when used in conjunction with feature selection. The specific feature selection methods we considered are described in detail in the next subsection. For each study we first determined the feature selection method that yielded the best results for the GNB classifier, and the feature selection method that yielded the best results for the SVM classifier. As it turned out, this best method was identical for GNB and SVM in all cases except for one: the Syntactic Ambiguity study, where there was a small, statistically insignificant difference.[2] Therefore, we report here the performance of all classifiers using the feature selection method that optimizes GNB (and usually SVM) performance. These results are summarized in Table I, on the second line corresponding to each study. As in the first experiment, GNB and SVM outperform kNN, and the performance of kNN improves as k increases.

6.2.1. *Analysis*

One clear trend in this data is that kNN fared less well than GNB or SVMs across all studies and conditions. In retrospect, this is not too surprising given the high dimensional, sparse training data sets. It is well known that the kNN classifier is sensitive to irrelevant features, as these features add in irrelevant ways to the distance between train and test examples (Mitchell, 1997). This explanation for the poor performance of kNN is also consistent with the dramatic improvement in kNN performance resulting from feature selection. As the table results indicate, feature selection often reduces kNN error by a factor of two

[2] The "Active" feature selection method which optimized GNB performance in the Syntactic Ambiguity study yielded an SVM error of 0.23, compared to an SVM error of 0.21 using the "ROIActiveAvg" method.

or more, presumably by removing many of these irrelevant, misleading features.

The relationship between the performance of GNB and SVM is less clear cut, as each outperforms the other in three of the six rows where they are compared in Table I. Given the small number of examples available in both the Syntactic Ambiguity study, and the Noun vs. Verb study, the differences reported between GNB and SVM for these two studies are not statistically significant. However, for the Picture vs. Sentence study, which has more examples, the difference between GNB and SVM using no feature selection (the row labeled "All") is significant at the p=0.15 level, and the difference when using feature selection (the row labeled "Active) is significant at a p value extremely close to zero.

These two statistically significant results comparing GNB and SVM for the Picture vs. Sentence study are consistent with the general observations in the beginning of this subsection: the SVM performed more poorly than GNB when no feature selection was used but outperformed GNB when using feature selection. In fact, the stronger performance of SVM when using feature selection held whether the number of features selected was 20, 100, 200, 400, or 800, and held not only on average across subjects, but also held for individual subject accuracies on at least 12 of the 13 subjects. The number of selected active voxels for which GNB and SVM reached their optimal accuracies were similar.

In summary, we found when training fMRI classifiers across a variety of data sets and target functions that GNB and SVM outperformed kNN quite consistently. Furthermore, in choosing between GNB and SVMs, we found trends consistent with the analysis in (Ng and Jordan, 2002), namely that GNB worked better when the data was especially sparse and high dimensional, and that SVM performed relatively better when data sets were larger, or feature selection was used to reduce the data dimension.

6.3. Which Feature Abstraction Method Works Best?

Given that our classification problem involves very high dimensional, noisy, sparse training data, it is natural to consider feature selection methods to reduce the dimensionality of the data before training the classifier. As we discussed in the previous section, and as summarized in Table I, feature selection leads to large and statistically significant improvements in classification error across all four of our case studies. In this section we discuss in detail the feature selection methods explored in our work, and some surprising lessons learned regarding which feature selection methods worked best.

6.3.1. *Approach*
The most common approach to feature selection when training classifiers is to select a subset of the available features based on some statistical test that scores each feature by its ability to *discriminate* the target classes. For example, given the goal of learning a target classification function $f : X \rightarrow Y$, one common approach to feature selection is to rank order the features of X by their mutual information with the class variable Y, then to greedily select the n highest scoring features (e.g., (Cover and Thomas, 1991)).

Given the nature of classification problems in the fMRI domain (and many other domains as well), a second general approach to feature selection is also possible. To illustrate, consider the problem of learning a Boolean classifier $f : X \rightarrow Y$ where $Y = \{1, 2\}$, given training examples labeled as belonging to either class 1 or class 2 (e.g., learning to distinguish whether the subject is viewing a noun or verb). In fMRI studies, we naturally obtain *three* classes of data rather than two. In addition to data representing class 1 and class 2, we also obtain data corresponding to a third "fixation" or "rest" condition. This fixation condition contains data observed during the time intervals between trials, during which the subject is generally at rest (e.g., they are examining neither a noun nor a verb, but are instead staring at a fixation point). Thus, we can view the data associated with class 1 and class 2 as containing some signal conditioned on the class variable Y, whereas the data associated with fixation contains no such signal, and instead contains only background noise relative to our classification problem. In this setting, we can consider a second general approach to feature selection: score each feature by how well it discriminates the class 1 or class 2 data from the zero signal data. In the terminology of fMRI, we score each feature based on how active it is during the class 1 or class 2 intervals, relative to the fixation intervals. The intuition behind this feature selection method is that it emphasizes choosing voxels with large signal-to-noise ratios, though it ignores whether the feature distinguishes the target classes.

We refer to this general setting as the "zero signal" learning setting, summarized in Figure 3. Notice many classification problems involving sensor data can be modeled in terms of this zero signal learning setting (e.g., classifying speakers based on voice data, where the zero signal condition corresponds to background noise when neither person is speaking). Therefore, understanding how to perform feature selection and classification within this setting has relevance beyond the domain of fMRI. In fact, within the fMRI literature it is common to use activity to select a subset of relevant voxels, and then to compare the behavior of this selected subset over various conditions.

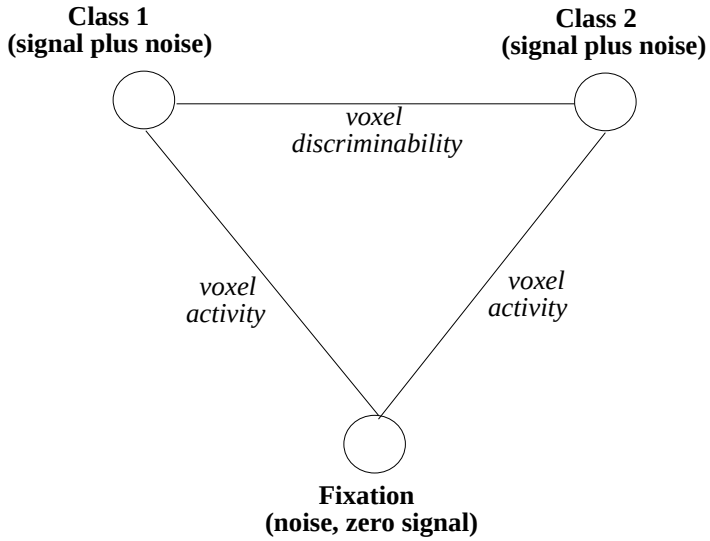


Figure 3. The "zero signal" learning setting. Boolean classification problems in the fMRI domain naturally give rise to three types of data: data corresponding to the two target classes plus data collected when the subject is in the "fixation" or "rest" condition. We assume the data from class 1 and class 2 are composed of some underlying signal plus noise, whereas data from the fixation condition contains no relevant signal but only noise. In such settings, feature selection methods can consider both *voxel discriminability* (how well the feature distinguishes class 1 from class 2), and *voxel activity* (how well the feature distinguishes class 1 or class 2 from the zero signal class).

In the experiments summarized below, we consider feature selection methods that select voxels based on both their ability to distinguish the target classes from one another (which we call *discriminability*), and on their ability to distinguish the target classes from the fixation condition (which we call *activity*). Although each feature consists of the value of a single voxel at a single time, we group the features involving the same voxel together for the purpose of feature selection, and thus focus on selecting a subset of voxels. In greater detail, the feature selection (voxel selection) methods we consider here are:

– *Select the n most discriminating voxels (Discrim).* In this method, a separate classifier is trained for each voxel, using only the observed fMRI data associated with that voxel. The accuracy of this single-voxel classifier over the training data is taken as a measure of the discriminating power of the voxel, and the n voxels that score highest according to this measure are selected. Note when reporting cross validation errors on final classifiers using this feature selection method, features are selected separately for each

cross-validation fold in order to avoid using data from the test fold during the feature selection process. Thus, the voxels selected may vary from fold to fold.

- *Select the n most active voxels (Active).* In this method, voxels are selected based on their ability to distinguish either target class from the fixation condition. More specifically, for each voxel, v, and each target class y_i, a t-test is applied to compare the voxel's fMRI activity in examples belonging to class y_i to its activity in examples belonging to fixation periods. The first voxels are then selected by choosing for each target class y_i the voxel with the greatest t statistic. The next voxels are selected by picking the second most active remaining voxel for each class, and so on, until n voxels are chosen. Notice the selected voxels may distinguish just one target class from fixation, or may distinguish both target classes from fixation.

- *Select the n most active voxels per Region of Interest (roiActive).* This is similar to the Active method above, but ensures that voxels are selected uniformly from all regions of interest (ROIs) within the brain. More precisely, given m prespecified ROIs, this method applies the Active method to each ROI, selecting n/m voxels from each. The union of these voxels are returned as the n selected voxels.

The approaches above for selecting voxels can be combined with methods for averaging the values of multiple features (in space or time), and with methods that select data over a sub-interval in time. We experimented with various combinations of such approaches, and report here on the above three methods (Discrim, Active, roiActive) as well as a fourth method derived from roiActive:

- *Calculate the mean of active voxels per ROI (roiActiveAvg).* This method first selects n/m voxels for each of the m ROIs using the roiActive method. It then creates a single "supervoxel" for each ROI, whose activity at time t is the mean activity of the selected ROI voxels at time t.

6.3.2. *Results*

We experimented with each of these four feature selection methods, over each of the four case study data sets. For comparison purposes we also report errors obtained using all features (denoted as *"All"*). In each experiment we considered a range of numbers of voxels to keep [3].

[3] The numbers of voxels considered were $\{20, 100, 200, 400, 800\}$ for Picture vs Sentence, $\{4, 20, 40, 80, 160\}$ for Syntactic Ambiguity, $\{100, 800, 1600\}$ for Verb

Table II. Picture vs Sentence Study - errors by subject and feature selection method. The first column indicates the feature selection method, with each row reporting errors achieved using GNB with this method (feature selection method "All" uses all available features). The second column indicates the average error over all 13 subjects. Remaining columns indicate errors for individual subjects A through M. Errors in columns A through M are the minimum errors achieved by varying the number of features selected.

Feature Sel.	Average Error	A	B	C	D	E	F	G
All	0.29	0.48	0.10	0.31	0.03	0.31	0.40	0.45
Discrim	0.26	0.26	0.09	0.34	0.06	0.24	0.31	0.33
Active	0.16	0.21	0.08	0.19	0.01	0.11	0.31	0.33
roiActive	0.18	0.30	0.09	0.24	0.03	0.13	0.29	0.31
roiActiveAvg	0.21	0.45	0.16	0.31	0.01	0.25	0.28	0.26

		H	I	J	K	L	M
		0.06	0.16	0.11	0.13	0.31	0.09
		0.19	0.31	0.29	0.36	0.45	0.15
		0.05	0.13	0.19	0.15	0.33	0.10
		0.11	0.19	0.19	0.19	0.24	0.11
		0.21	0.28	0.28	0.23	0.46	0.20

Tables II through V present summarized results for each of the four fMRI studies. Each table shows the best errors obtained for each feature selection method and for each subject considered in the study, when using a GNB classifier. Here the best error refers to the lowest error achieved by varying the number of selected voxels. The optimal number of voxels selected varied by study, subject, and feature selection method, typically ranging from 5 to 20% of the total number available in the selected ROIs.

6.3.3. *Analysis*

These results indicate that using feature selection leads to improved classifier error in all studies considered. More specifically, the best feature selection method outperforms no feature selection for every human subject in every case study except the "Noun versus Verb" study, where it outperforms no feature selection in three of the four subjects.

A second strong trend in the results is the dominance of feature selection methods based on activity (Active, roiActive, roiActiveAvg)

vs Noun and $\{100, 200, 400, 800, 1200, 1600, 2000, 2400, 2800, 3200, 3600, 4400\}$ for Semantic Categories.

Table III. Syntactic Ambiguity Study - GNB errors by subject and feature selection method. Results are presented using the same format as Table II.

Feature Selection	Average Error	A	B	C	D	E
All	0.43	0.30	0.55	0.55	0.35	0.40
Discrim	0.34	0.25	0.50	0.30	0.30	0.35
Active	0.25	0.20	0.35	0.25	0.25	0.20
roiActive	0.27	0.15	0.35	0.35	0.30	0.20
roiActiveAvg	0.27	0.25	0.30	0.30	0.30	0.20

Table IV. Noun vs Verb Study - GNB errors by subject and feature selection method. Results are presented using the same format as Table II.

Feature Selection	Average Error	A	B	C	D
All	0.36	0.38	0.19	0.38	0.50
Discrim	0.36	0.38	0.19	0.38	0.50
Active	0.34	0.31	0.38	0.31	0.38
roiActive	0.31	0.25	0.31	0.31	0.38
roiActiveAvg	0.23	0.19	0.25	0.25	0.25

over those based on discriminability (Discrim). As can be seen in the tables, the average error for the best activity-based method outperforms the average error for the discriminability method in all four case studies. The best activity-based method dominates the discriminability-base method for 11/13 subjects in the "Picture versus Sentence" study, for 8/10 subjects in the "Semantic Categories" study, 5/5 subjects in the "Syntactic Ambiguity" study and 3/4 subjects in the "Noun versus Verb" study. Under the null hypothesis that activity-based and discrimination-based methods perform better equally often, the prob-

Table V. Semantic Categories - GNB errors by subject and feature selection method. Results are presented using the same format as Table II.

Feature Sel.	Avg Error	A	B	C	D	E	F	G	H	I	J
All	0.10	0.13	0.17	0.04	0.12	0.06	0.07	0.20	0.04	0.14	0.05
Discrim	0.10	0.10	0.17	0.04	0.12	0.06	0.07	0.19	0.04	0.13	0.05
Active	0.08	0.11	0.12	0.04	0.10	0.06	0.06	0.11	0.04	0.13	0.04
roiActive	0.09	0.12	0.13	0.05	0.11	0.06	0.07	0.12	0.04	0.14	0.04

ability of the activity-based methods dominating at least this often is, respectively, 0.05, 0.01, 0.03 and 0.31 (using a binomial distribution with a parameter of 0.5, where the number of trials corresponds to the number of subjects in the study).

It is at first surprising to observe that selecting features based on their activity level works dramatically better than selecting them based on their ability to discriminate the target classes. Given that the end goal is to discriminate the target classes, and that selecting features based on discriminability is the norm in machine learning applications, one might well expect discriminability to have been the dominant method. Below we look deeper into why we observe the opposite result in all four fMRI studies.

One situation in which we might expect activity-based feature selection to outperform discrimination-based methods is when data dimensionality is very high, noise levels are high, training data are sparse, and very few voxels contain a signal related to the target classes. In such cases, we should expect to find that some voxels that are truly irrelevant appear nonetheless to be good discriminators over the sparse sample of training data - even when using cross validation to test their discrimination power. The larger the set of such irrelevant voxels, the more likely that a feature selection strategy focused on discrimination would select such overfitting voxels, and be unable to distinguish these from truly informative discriminating voxels. However, in this same case we might expect that choosing voxels with high signal-to-noise ratios would be a useful strategy, as it would remove from consideration the large number of irrelevant voxels (i.e., those with no signal, but only noise). In fact, our activity-based feature selection strategies select exactly this kind of high signal-to-noise ratio voxels. The bottom line is that each feature selection strategy runs its own risk: discrimination-based methods run the risk of selecting voxels that only coincidentally fit the noisy training sample, whereas activity-based methods run the risk of choosing high signal-to-noise voxels that cannot discriminate the target classes. Which risk is greater depends on the exact problem, but the relative risk for the discrimination-based method grows more quickly with increasing data dimension, increasing noise level, decreasing training set size, and an increasing fraction of irrelevant features.

To explore this conjecture, let us examine the actual characteristics of the voxels selected by these two methods in our data. In particular, let us focus on a single subject in the Semantic Categories study: subject G, whose best average normalized rank error (0.108) is obtained by a GNB classifier using 800 voxels chosen using the Active feature selection strategy. Using the same number of voxels selected instead by

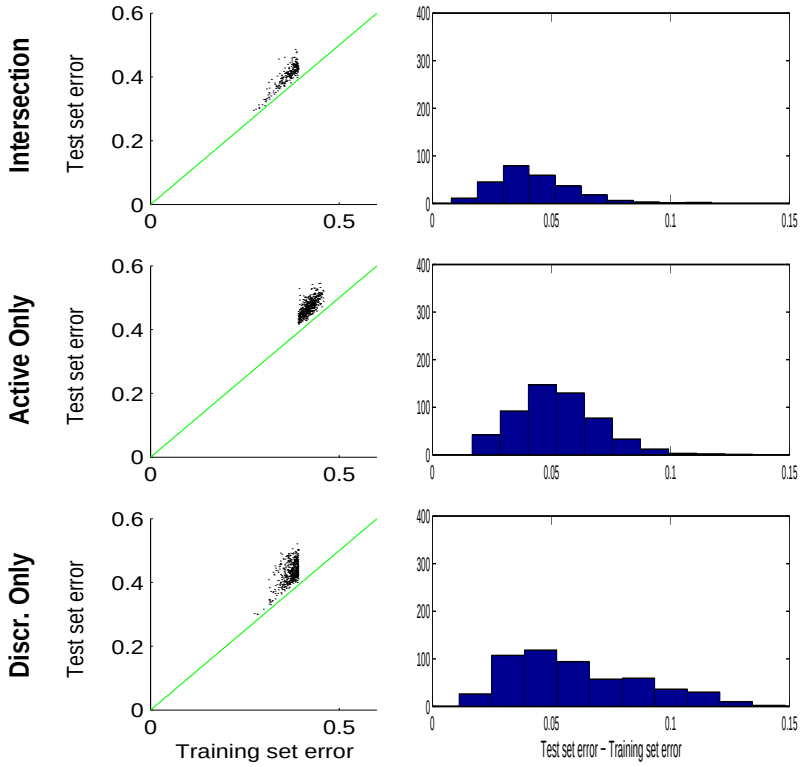


Figure 4. (a) Scatter plots of training set error (horizontal axis) against test set error (vertical axis) for the voxels in each subset (row). (b) Histograms depicting for each voxel subset the number of voxels that overfit to various degrees. The horizontal axis in this case is (test error minus training error), measuring the degree of overfitting.

the Discrim method leads to a substantially higher error (0.190). Are there in fact differences in the degree of overfitting between these two sets of selected voxels?

To explore this question, let us consider three sets of voxels: voxels chosen by Active feature selection but not by Discrim ("ActiveOnly"), voxels chosen by Discrim by not by Active ("DiscriminatingOnly"), and voxels chosen independently by both methods ("Intersection"). For this particular subject, there are 251 voxels in the Intersection set, and 549 in each of ActiveOnly and DiscriminatingOnly. Training a GNB classifier using Intersection yields an error of (0.106), slightly but not significantly better than the error from the Active voxels.

Figure 4 shows the degree of overfitting for each of the three sets of voxels. On the left, panel (a) provides a scatterplot of training set error (horizontal axis) versus test set error (vertical axis). The straight line

indicates where training error equals test error. Notice all three sets of voxels overfit to some degree (i.e., test error is generally greater than or equal to training error), but that the cluster of voxels ranges furthest from the straight line for the DiscriminatingOnly voxels. On the right, panel (b) provides a histogram showing the number of voxels in each set that overfit to varying degrees. Note the DiscriminatingOnly voxel set contains many more voxels that overfit to a large degree. Based on the data summarized in this figure, it is clear that the degree of overfitting is indeed greater in this case for voxels selected by Discrim than those selected by Active. It is also clear that the voxels in Intersection suffer the least overfitting.

A different view into the character of these three voxel sets is provided by Figure 5. Panel (a) plots the 10 voxels with the best training set error from each of the three sets. Each voxel plot shows the learned Gaussian model for each of the twelve target classes. Notice the greater spread of these models for the voxels chosen by the Discrim method (DiscriminatingOnly and Intersection) than for the ActiveOnly set. Panel (b) provides a scatter plot of standard deviation (horizontal axis) versus test error (vertical axis) for the three voxel sets. Notice the significantly lower standard deviation for the ActiveOnly set.

Above we suggested that the Discrim method for feature selection carries a risk of selecting voxels that overfit the data. The above data, especially from Figure 4 indicates that in fact overfitting is greater for Discrim than for Active in our data. We also suggested the Active method carries the counterbalancing risk of selecting irrelevant voxels. Is this in fact occurring in our case? The plots in Figure 5 show that the ActiveOnly set of voxels does appear to contain voxels that are poor discriminators among the twelve target classes.

To understand the impact of poor discriminators (irrelevant voxels) selected by the Active method, consider the relative weight of the relevant versus irrelevant voxels used by a GNB. Given an instance X to be classified, the log odds assigned by the GNB for two classes c_i and c_j is

$$\log \frac{\hat{P}(c_i|X)}{\hat{P}(c_j|X)} = \log \frac{\hat{P}(c_i)}{\hat{P}(c_j)} + \sum_k \log \frac{\hat{P}(x_k|c_i)}{\hat{P}(x_k|c_j)}$$

where x_k is the observed value for the kth feature (i.e., the kth voxel) of X, and where $\hat{P}$ denotes distributions estimated by GNB based on the training data. Note the GNB classifier will predict class c_i if the above log odds ratio is positive, and c_j if it is negative. Thus, the decision of the GNB is determined by a linear sum, where each voxel contributes one term to the sum.

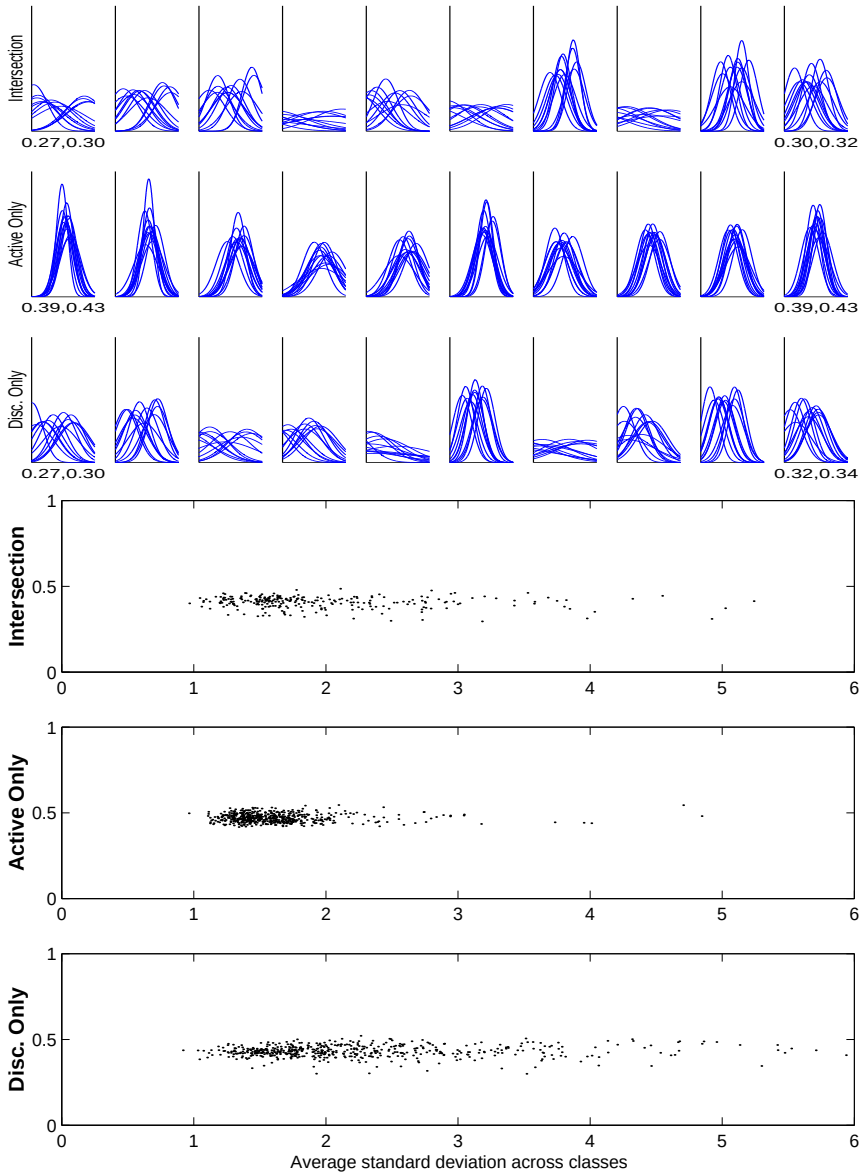


Figure 5. In the top half of the figure, each plot shows the 12 learned class probability densities for a given voxel, with the x axis ranging from -5 to 5. Each row contains the 10 voxels with the lowest training set errors from each voxel subset, sorted by increasing error. For reference, the leftmost and rightmost plots in each row have their (training set,test set) error values below them. The bottom half of the figure provides scatterplots depicting the average class standard deviation (horizontal axis) against the test error (vertical axis) for each voxel subset (row). Note the higher variance for the Discriminating voxels (Intersection and DiscriminatingOnly).

Now let us consider a voxel x_k which is truly irrelevant to the classification (i.e., where the true distributions $P(x_k|c_i)$ and $P(x_k|c_j)$ are identical). First, consider the situation in which the learned estimates $\hat{P}(x_k|c_i)$ and $\hat{P}(x_k|c_j)$ are also identical. In this case the fraction involving x_k will be equal to 1, its log will be equal to 0 regardless of the observed value of x_k, and voxel x_k will therefore have no influence on the final GNB decision. Now consider the situation in which $\hat{P}(x_k|c_i)$ and $\hat{P}(x_k|c_j)$ differ (e.g., due to overfitting) despite the fact that $P(x_k|c_i) = P(x_k|c_j)$. In this case, the x_k term will in fact be nonzero, and will have a detrimental, randomizing influence on the final GNB classification. Is this in fact occurring in our data? The plots in panels (a) and (b) of Figure 5 suggest that the Active voxels that are irrelevant (i.e., those in ActiveOnly) do indeed have strongly overlapping $\hat{P}(x_k|c_i)$ distributions, limiting the magnitude of their contribution to the final GNB classification.

To summarize, we find clear empirical evidence that feature selection consistently improves classification error, and that activity-based feature selection outperforms discrimination-based feature selection in all four studies. Our classification problem setting involves an interesting third category of "fixation" data, as illustrated in Figure 3. In fact, we believe that a variety of sensor-based classification problems have a similar property, providing a zero-signal class of data. We conjecture, and support with a variety of empirical observations, that activity-based feature selection may outperform discrimination-based feature selection in zero-signal classification problems, especially with increasing data dimension, noise, and sparsity, and as the proportion of truly relevant features decreases.

6.4. Can One Train Classifiers Across Multiple Subjects?

All results discussed so far in this paper have focused on the problem of training subject-specific classifiers. This section considers the question of whether it is possible to train classifiers that apply across multiple human subjects, including subjects beyond the training set.

The biggest obstacle to inter-subject analysis of fMRI data is anatomical variability among subjects. Different brains have different shapes and sizes, making it problematic to register the many thousands of voxels in one brain to their precise corresponding locations in a second brain. One common approach to this problem is to transform (geometrically morph) fMRI data from different subjects into some standard anatomical space, such as Talairach coordinates (Talairach and Tournoux, 1988). The drawback of this method is that the transformation always introduces some degree of error into the spatial map.

However, some feature selection and abstraction methods used in our studies are immune to anatomical variability. For example, by averaging the voxels in a particular ROI into a supervoxel (and treating it as a single voxel ROI afterwards), we can easily map one brain to another in terms of these anatomically defined ROI supervoxels. This approach provides a successful way to train classifiers across subjects.

A second difficulty that arises when training multiple-subject classifiers is that the intensity of fMRI response to a particular stimulus is usually different across subjects. We employ a normalization method that linearly rescales the data from different subjects into the same range to partially address this issue. While there are many inter-subject differences that cannot be addressed by this simple linear transformation, we have found this normalization to be useful. We have also found a similar normalization method can sometimes reduce classification error for single-subject classifiers, when used to normalize data across different trials for that subject.

We performed experiments to train multiple subject classifiers using two data sets: the Picture versus Sentence data, and the Syntactic Ambiguity data. The following two subsections describe these experiments in turn.

6.4.1. *Sentence Versus Picture Study*

We trained multiple-subject classifiers for the Sentence versus Picture study, to discriminate whether the subject was viewing a picture or a sentence. Multiple-subject classifiers were trained using data from 12 of the 13 subjects, abstracting the data from each subject into ROI supervoxels as described above. To evaluate the error of these trained classifiers, we used leave-one-subject-out cross validation. In particular, for each subject we trained on the remaining 12 subjects, measured the error on this held out subject, then calculated the mean error over all held out subjects.

The results, summarized in Table VI, show that the linear SVM learns a cross-subject classifier that achieves error of 0.25 ± 0.026 over the left out subject. This is highly statistically significant compared to the 0.50 error expected of random guessing, indicating that it is indeed possible to train a classifier to capture significant subject-independent regularities in brain activity that are sufficiently strong to detect single-interval cognitive states in human subjects who are not part of the training set. As in earlier experiments, we note that SVM and GNB again outperform kNN.

In a second set of experiments, we partitioned the Sentence versus Picture data into two disjoint subsets: trials in which the sentence was presented before the picture (which we will refer to as S-then-P), and

Table VI. Errors for multiple subject classifier, Sentence versus Picture study. The last column shows the error of a multi-subject classifier when applied to a subject withheld from the training set. Results are obtained using normalization and 7 ROIs. All classifiers are trained by averaging all voxels in an ROI into a supervoxel. 95% confidence intervals are computed under the assumption that test examples are i.i.d. Bernoulli distributed. The error of a random classifier is 0.50.

Classifier	*Leave-1-subject-out error*
GNB	0.30±0.028
SVM	0.25±0.026
1NN	0.36±0.029
3NN	0.33±0.029
5NN	0.32±0.028

Table VII. Errors for single-subject and multiple-subject classifiers, when trained on P-then-S, and S-then-P data. The third column shows the average error of classifiers trained for single subjects. The fourth column shows the error of multi-subject classifiers applied to subjects withheld from the training set. Results are obtained using normalization. All classifiers are trained based upon averaging all voxels in an ROI into a supervoxel. 95% confidence intervals are computed under the assumption that test examples are i.i.d. Bernoulli distributed. The error of a random classifier is 0.50.

Data Set	*Classifier*	*Avg Single Subject Error*	*Leave-1-subject-out*
S-then-P	GNB	0.10±0.024	0.14±0.030
S-then-P	SVM	0.11±0.025	0.13±0.029
S-then-P	1NN	0.13±0.028	0.15±0.031
S-then-P	3NN	0.12±0.027	0.13±0.029
S-then-P	5NN	0.10±0.025	0.11±0.027
P-then-S	GNB	0.20±0.033	0.20±0.034
P-then-S	SVM	0.17±0.031	0.22±0.036
P-then-S	1NN	0.38±0.041	0.26±0.038
P-then-S	3NN	0.31±0.039	0.24±0.037
P-then-S	5NN	0.26±0.037	0.21±0.035

trials in which the picture was presented before the sentence (which we will call P-then-S). For each of these sets we repeated the training of multi-subject classifiers, training and testing using only the data from the subset. We performed this experiment to understand the impact of these two differing contexts on the difficulty of the classification problem. The results are summarized in Table VII. Note for comparison

we present both the leave-1-subject-out error of the multiple subject classifiers, and the average leave-one-example-per-class-out error of the corresponding single subject classifiers.

As in the first table, the multiple subject classifiers achieve accuracies significantly greater than the 0.5 expected from random guessing. In fact, these classifiers are significantly more accurate than those in the first experiment, despite the fact that they are trained using only half of the data. Presumably the explanation for this greater accuracy is that the classification task is easier here than when using the full data – in the full data examples come from a greater diversity of temporal contexts, and the effects of these different contexts can remain apparent for several seconds due to the temporally delayed BOLD response.

A second interesting trend apparent in Table VII is that the error on the left out subject for the multiple subject classifiers is often very close to the average error of the single subject classifiers, and in several cases it is statistically significantly better than the corresponding single subject classifiers. Presumably this better performance by the multiple subject classifier can be explained by the fact that it is trained using an order of magnitude more training examples, from twelve subjects rather than one.

6.4.2. *Syntactic Ambiguity Study*
We also attempted to train multiple subject classifiers for the Syntactic Ambiguity study, to discriminate whether the subject was reading an ambiguous or non-ambiguous sentence. In this case, the best error of a multi-subject classifier obtained was 0.36 ± 0.094 under leave-one-subject-out cross validation, and correspondingly the average error of single subject classifiers is 0.35 ± 0.092. The setting which produced this result was using GNB, Normalization, and feature selection method roiActiveAvg, averaging the 20 most active voxels from each ROI into a supervoxel. These errors are significantly better than expected from a random classifier, 0.50. Unlike the Sentence versus Picture study, however, these results are quite sensitive to the particular selection of learning method and feature selection. Although we cannot draw strong conclusions from this result, it provides modest additional support for the feasibility of training multiple subject classifiers.

7. Summary and Conclusions

We have presented results from four different fMRI studies demonstrating the feasibility of training classifiers to distinguish a variety of cognitive states, based on single-interval fMRI observations. This

problem is interesting both because of its relevance to studying human cognition, and as a case study of machine learning in high dimensional, noisy, sparse data settings.

Our comparison of classifiers indicates that Gaussian Naive Bayes (GNB) and linear Support Vector Machine (SVM) classifiers outperform k Nearest Neighbor across all four studies, and that feature selection methods consistently improve classification error in all four studies. In comparing GNB to SVM, we found trends consistent with the observations in (Ng and Jordan, 2002), that the relative performance of generative versus discriminative classifiers depends in a predictable fashion on the number of training examples and data dimension. In particular, our experiments are consistent with the hypothesis that the accuracy of SVM's increases relatively more quickly than the accuracy of GNB as the data dimension is reduced via feature selection, and as the number of training examples increases.

Feature selection is an important aspect in the design of classifiers for high dimensional, sparse, noisy data. We defined a new classifier setting (the zero signal setting) that captures an important aspect of our fMRI classification problem, as well as a variety of other classification problems involving sensor data. In this setting, the available data includes not only examples of the classes to be discriminated (e.g., data when the subject is reading a noun, or a verb), but also a class of "zero signal" data (e.g., when the subject is reading neither a noun nor a verb, but is simply fixating on the screen). Our experiments show that feature selection methods taking advantage of this zero signal data consistently outperform traditional feature selection methods that use only data from the target classes. We plan further research to develop a more precise formal model of this zero signal setting, and to develop and experiment with feature selection strategies tuned to take maximal advantage of this setting.

In addition to training classifiers to detect cognitive states in single subjects, we also explored the feasibility of training cross-subject classifiers to make predictions across multiple human subjects. In this case, we found it useful to abstract the fMRI data by using the mean fMRI activity in each of several anatomically defined brain regions. Using this approach, it was possible to train classifiers to distinguish, e.g., whether the subject was viewing a picture or a sentence describing a picture, and to apply these successfully to subjects outside the training set. In some cases, the classification accuracy for subjects outside the training set equalled the accuracy achieved by training on data from just this single subject. Given this success in training cross-subject classifiers, we plan additional research to explore a number of alternative approaches to cross-subject classification (e.g., instead of abstracting the data for each

subject, map the different brain structures to a standard coordinate system such as Talairach coordinates).

There are many additional opportunities for machine learning research in the context of fMRI data analysis. For example, it would be useful to learn models of temporal behavior, in contrast to the work reported here which considers only data at a single time or time interval. Machine learning methods such as Hidden Markov Models and Dynamic Bayesian Networks appear relevant. A second research direction is to develop learning methods that take advantage of data from multiple studies, in contrast to the single study efforts described here. In our own lab, for example, we have accumulated fMRI data from over 800 human subjects. A third research topic is to develop machine learning methods that could take as a starting point computational cognitive models of human processing, such as ACT-R (Anderson et al., in press) and 4CAPS (Just et al., in press), using these as prior knowledge for guiding the analysis of fMRI, and automatically refining these models to better fit the observed fMRI data.

Acknowledgements

We are grateful to Luis J. Barrios for helpful discussions and detailed comments on various drafts of this paper. Thanks to Vladimir Cherkassky and Joel Welling for useful observations and suggestions during the course of this work. We thank Paul Bennett for many helpful discussions and for writing part of the code used for the Word Categories study.

Radu Stefan Niculescu was supported by a Graduate Fellowship from the Merck Computational Biology and Chemistry Program at Carnegie Mellon University established by the Merck Company Foundation and by National Science Foundation (NSF) grant no. CCR-0122581. Francisco Pereira was funded by a PRAXIS XXI scholarship from Fundação para a Ciência e Tecnologia, Portugal (III Quadro Comunitário de Apoio, comparticipado pelo Fundo Social Europeu) and a PhD scholarship from Fundação Calouste Gulbenkian, Portugal. Rebecca Hutchinson was supported by an NSF Graduate Fellowship.

References

Aguirre, G. K., E. Zarahn, and M. D'Esposito. An Area within Human Ventral Cortex Sensitive to Building Stimuli: Evidence and Implications. *Neuron*, 21, 373-383, 1998.

Anderson, J. R., et al. An information-processing model of the BOLD response in symbol manipulation tasks. in press.

Battig, W. F., and W. E. Montague Category Norms for Verbal Items in 56 Categories: A replication and extension of the Connecticut Norms. *Journal of Experimental Psychology Monograph*, 80, (3), 1-46, 1968.

Blankertz, B., G.Curio, and K. R. Mller Classifying single trial EEG: Towards brain computer interfacing. *Advances in Neural Inf. Proc. Systems (NIPS 2001)*, 14, 157-164, 2002.

Bly, B. M. When you have a General Linear Hammer, every fMRI time-series looks like independent identically distributed nails. *Concepts and Methods in NeuroImaging workshop*, NIPS 2001.

Burges, C. A Tutorial on Support Vector Machines for Pattern Recognition. *Journal of data Mining and Knowledge Discovery*, 2(2),121-167, 1998.

Caviness, V. S., D. N. Kennedy, J. Bates, and N. Makris. MRI-based parcellation of human neocortex: an anatomically specified method with estimate of reliability. *Journal of Cognitive Neuroscience*, 8, 566-588, 1996.

Chao, L., J. V. Haxby, and A. Martin. Attribute-based Neural Substrates in Temporal Cortex for Perceiving and Knowing about Objects. *Nature Neuroscience*, 2, 913-919, 1999.

Chao, L., J. Weisberg, and A. Martin. Experience-dependent Modulation of Category-related Cortical Activity. *Cerebral Cortex*, 12, 545-551, 2002.

Cover, T. and Thomas, J. Elements of Information Theory. *Wiley and Sons*, 1991.

Eddy, W., et al. The Challenge of Functional Magnetic Resonance Imaging. *Journal of Computational and Graphical Statistics*, Volume 8, Number 3, Page 545-558, 1998.

Friston, K. J., et al. Statistical Parametric Maps in Functional Imaging: A General Linear Approach. *Human Brain Mapping*, 2, 189-210, 1995.

Genovese, C. Statistical Inference in Functional Magnetic Resonance Imaging. CMU Statistics Tech Report 674, 1999.

Goutte, C., P. Toft, E. Rostrup, F. A. Nielsen, and L. K. Hansen. On clustering fMRi time series. Technical Report IMM-REP-1998-11, 1998.

Haxby, J., et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2425-2430, 2001.

Hojen-Sorensen, P., L. K. Hansen, and C. E. Rasmussen. Bayesian Modeling of fMRI Time Series. NIPS*99. Denver, November 29 - December 4, 1999.

Ishai, A., L. G. Ungerleider, A. Martin, J. L. Schouten, and J. V. Haxby. Distributed Representation of Objects in the Human Ventral Visual Pathway. *Proc. Nat. Acad. Sci. USA*, 96, 9379-9384, 1999.

Joachims, T. A Statistical Learning Model of Text Classification with Support Vector Machines. *Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR)*, ACM, 2001.

Just, M. A., P.A. Carpenter, and Varma, S. Computational modeling of high-level cognition and brain function. *Human Brain Mapping*, in press.

Keller, T. A., Just, M. A., and Stenger, V. A. Reading span and the time-course of cortical activation in sentence-picture verification *Annual Convention of the Psychonomic Society*, Orlando, FL, November 2001.

Mason, R., M. Just, T. Keller, and P. Carpenter. Ambiguity in the Brain: What brain imaging reveals about the processing of syntactically ambiguous sentences, in press, 2003.

McKeown, M. J., et al. Analysis of fMRI data by blind separation into independent spatial components, *Human Brain Mapping*, Vol. 6, No. 3, pp. 160-188, 1998.

Menon, R. S., Luknowsky, D. L., Gati, J. S. Mental chronometry using latency-resolved functional magnetic resonance imaging, *Proc. Natl. Acad. Sci (U.S.A.)*, 95:10902-10907, 1998

Mitchell, T. M. *Machine Learning*, McGraw-Hill, 1997.

Nigam, K., A. McCallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, No. 39, pp. 103-134, 2000.

Ng, A.Y., and M. Jordan. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. *Neural Information Processing Systems*, Vol. 14, 2002.

Penny, W. Mixture Models with Adaptive Spatial Priors. *Concepts and Methods in NeuroImaging* workshop at NIPS*01, Vancouver, British Columbia, Canada, December 3 - 8, 2001.

Rademacher, J., A. M. Galaburda, D. N. Kennedy, P. A. Filipek, and V. S. Caviness. Human cerebral cortex: Localization, parcellation, and morphometry with magnetic resonance imaging. *Journal of Cognitive Neuroscience*, 4, 352-374, 1992.

Talairach, J., Tournoux, P. Co-planar Stereotaxic Atlas of the Human Brain. *Thieme, New York*, 1988.

Wagner, A. D., et al. Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, 281, 1188-1191, 1998.

Yang, Y. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1, (1,2), 67-88, 1999.