Discovering a Semantic Basis of Neural Activity Using Simultaneous Sparse Approximation

Keywords: sparsity, simultaneous sparse approximation, multi-task feature learning, simultaneous variable selection, functional magnetic resonance imaging, fMRI

Mark Palatucci MPALATUC@CS.CMU.EDU
Tom Mitchell TOM@CS.CMU.EDU
Han Liu HANLIU@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

We consider the problem of predicting brain activation in response to arbitrary words in English. Whereas previous computational models have encoded words using predefined sets of features, we formulate a model that can automatically learn features directly from data. We show that our model reduces to a simultaneous sparse approximation problem and show two examples where learned features give insight about how the brain represents meanings of words.

1. Introduction

Over the last several years, researchers have designed algorithms to learn the complex patterns of brain (neural) activity from data generated by functional magnetic resonance imaging (fMRI). These algorithms are often called *cognitive state classifiers* and can be used to discriminate between different mental states. For example, one study has shown it is possible to distinguish between categories of objects a person is thinking about simply by observing an image of his/her neural activity (Mitchell et al., 2004). Others have shown it is possible to determine lies from truth (Davatzikos et al., 2005) and whether someone is a Democrat or Republican (Kaplan et al., 2007).

While a large literature has developed around cognitive state classification (which maps neural activity to cognitive states), little attention has been given to the inverse problem: is it possible to predict neural activity for a novel state? One recent study from Kay (2008) predicts the neural activity in the visual cortex in response to viewing a novel scene.

Another study from Mitchell (2008) predicts neural activity in response to thinking about an arbitrary word in English. In this work, the semantic meaning of a word is encoded by co-occurrence statistics with other words in a very large text corpus. Using a small number of training words, a generative model is learned that maps these co-occurrence statistics to images of neural activity recorded while thinking about those words. Their model can then predict images for new words that were not included in the training set. The model shows predicted images that are similar to observed images for those words.

In their initial model each word is encoded by a vector of co-occurrences with 25 sensory-action verbs (e.g. eat, ride, wear). For example, words related to foods such as "apples" and "oranges" would have frequent co-occurrences with the word "eat" but few co-occurrences with the word "wear". Conversely, words related to clothes such as "shirt" or "dress" would co-occur frequently with the word "wear" but not the word "eat". Thus "eat" and "wear" are example *basis words* used to encode relationships of a broad set of other words.

These 25 sensory-action verbs were chosen based on domain knowledge from the cognitive neuroscience literature and are considered a *semantic basis* of latent word meaning. A natural question is:

What is the optimal basis of words to represent semantic meaning across many concepts?

Rather than relying on models that require a predetermined set of words, our research tries to build models

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

that will perform automatic *variable selection* to learn a semantic basis of word meaning. We want to learn models that not only predict neural activity well, but also give insights into how the brain represents the meaning of different concepts.

1.1. Related Work

Regression models such as the $\mathcal{L}_1$ regularized Lasso (Tibshirani, 1996) have been used successfully to perform variable selection. The typical model usually involves a large number of explanatory variables (features) and a single response variable. The Lasso will yield the best prediction of the response using only a small number of variables.

Recently, some attempts have been made to perform *multiple response variable selection* which will select a small number of variables that can explain multiple responses well. In statistics this is known as the *simultaneous lasso* (Turlach et al., 2005). The problem has also been addressed in machine learning as *multi-task feature learning* (Argyriou et al., 2007) and in signal processing as *simultaneous sparse approximation* (Tropp, 2006).

Several methods have formulated and solved the problem using convex programming but the current approaches seem very limited in scale. To our knowledge, there is no formulation that can solve problems with thousands of responses and thousands of explanatory variables. An alternative approach that is more tractable and does not involve convex programming is the greedy pursuit method *simultaneous orthogonal matching pursuit* (SOMP) (Tropp et al., 2006). Neural networks also offer a convenient formulation for multiple outputs and could be used with regularization constraints.

2. Problem Formulation

We can formulate the *multiple response variable selection* problem as a convex program. Let N be the number of examples, T be the number of responses, d be the number of explanatory variables. Let $\mathbf{Y} \in \Re^{NxT}$ be the matrix of response variables and $\mathbf{X} \in \Re^{Nxd}$ be our design matrix of explanatory variables. Our objective then is to find a sparse matrix of coefficients $\mathbf{B} \in \Re^{dxT}$. Let λ be a regularization parameter that controls the row sparsity of the matrix. Formally,

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{XB}||_F^2 + \lambda \sum_i^d \max_{j \in [1...T]} |\mathbf{B}_{ij}| \quad (1)$$

Our problem of discovering a semantic basis of neural activation can be formulated using Equation (1).

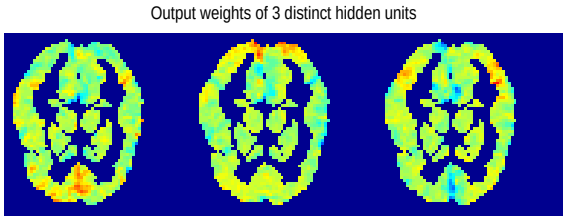

Figure 1. Weights on the output units for 3 hidden units in a 2-layer neural network. Regions with high weights are marked in red.

In our problem, we have a design matrix $\mathbf{X}$ which is the co-occurrences of $N = 60$ training words with $d = 50,000$ other English words. Each column of the response matrix $\mathbf{Y}$ contains the neural activations for a single voxel (volume-element) across the N fMRI images. There are $T = 20,000$ voxels. Our goal is to find a small number of words in $\mathbf{X}$ that can accurately predict the neural activity across multiple voxels in $\mathbf{Y}$.

3. Open Questions

What scalable techniques solve Equation (1) when there are thousands of response and explanatory variables?

Solving Equation (1) for our neural semantic problem is non-trivial. One approach may be to modify the large-scale interior-point method of Kim (2008) for multiple responses. Another option may be descent techniques with gradient projections.

What matrix norms should be imposed on the coefficient matrix $\boldsymbol{B}$?

Equation (1) imposes a $\mathcal{L}_1\mathcal{L}_\infty$ norm on the coefficient matrix $\mathbf{B}$. Other norms such as $\mathcal{L}_1\mathcal{L}_2$ could be imposed instead. Different norms will lead to different solutions and will affect the complexity of the optimization. Also, computing the full regularization path for λ may be easier for certain norms due to piecewise linearity of the regulation path.

How can spacial smoothness be imposed on the coefficient matrix $\boldsymbol{B}$?

In the case of fMRI images the coefficients in $\mathbf{B}$ have a geometric relationship. Since we are interested in how the brain performs at a regional rather individual voxel area, it may be preferable to have solutions that are spatially smooth, meaning that if a particular coefficient $\mathbf{B}_{ij}$ has high weight, then a local neighbor $\mathbf{B}_{ik}$ is biased to have high weight as well.

Table 1. Top 5 words ranked by weight in the three hidden units in Figure (1)

HIDDEN UNIT 1	HIDDEN UNIT 2	HIDDEN UNIT 3
WHITE	CHOPPED	FETISH
POTTERY	CUP	FISHING
CATHOLIC	CARROTS	SQUARE
BAPTIST	HOOD	POTTERY
CHRIST	PACKARD	FOOT

4. Convex Programming Alternatives

We have implemented two alternative methods that scale to very large problems and do not require convex programming. The first is a regularized 2-layer neural network with 50,000 inputs (a word is represented by its co-occurrences with 50,000 others words) and 20,000 output units (each fMRI image has neural activity for 20,000 voxels).

Figure (1) shows the output weights of the three most distinct hidden units. Interestingly, the network learns high weights in local neighborhoods of the brain. Table (1) shows the top five words for each hidden unit according to learned input weight. While the model is not truly sparse since most input weights are non-zero, the large non-zero weights can still give us insight into semantic meaning. For example, the first hidden unit contains several terms related to religion while the second hidden unit contains terms related to eating.

A second method we implemented was simultaneous orthogonal matching pursuit. Using this technique we greedily selected 25 words from a candidate set of 486 verbs and compared them against the 25 sensory-action verbs selected in the Mitchell (2008) model. The learned basis of 25 words was: *surround stitch nail bruise dried cut ran lick open press unpack chop employ dig cheat glow belong flew force train bang stain flood tip saw*

The original 25 sensory-action verbs were: *see hear listen taste smell eat touch rub lift manipulate run push fill move ride say fear open approach near enter drive wear break clean*

It is interesting to note the semantic similarities between words in both groups:

LICK →TASTE	STAIN → CLEAN	FLEW → DRIVE
RAN →RUN	SAW → SEE	PRESS → TOUCH
STITCH →WEAR	FORCE → MOVE	TIP → PUSH

5. Conclusion

Simultaneous sparse approximation is useful for learning a semantic basis of neural activation. This statistical method can help us discover useful ways to encode word meaning and build computational models of neural activation. Many practical challenges remain for building scalable methods for simultaneous variable selection. There are also a number of interesting theoretical questions regarding solution performance and the effect of different constraint norms.

References

Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Convex multi-task feature learning. *INSEAD Business School Research Paper No. 2007/13/TOM/DS*.

Davatzikos, C., et al. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, *28*, 663–668.

Kaplan, J. T., Freedman, J., & Iacoboni, M. (2007). Us versus them: Political attitudes and party affiliation influence neural response to faces of presidential candidates. *Neuropsychologia*, *45*, 55–64.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*, 352–356.

Kim, S.-J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2008). An interior-point method for large-scale l1 regularized least squares. *To appear in IEEE Journal on Selected Topics in Signal Processing*.

Liu, H., & Zhang, J. (2008). On the l1-lq regularized regression. *Carnegie Mellon Statistics Technical Report 860*.

Mitchell, T., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *To appear in the journal Science*.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, *57*, 145–175.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*, 267–288.

Tropp, J. A. (2006). Algorithms for simultaneous sparse approximation: Part ii: Convex relaxation. *Signal Processing*, *86*, 589–602.

Tropp, J. A., Gilbert, A. C., & Strauss, M. J. (2006). Algorithms for simultaneous sparse approximation: part i: Greedy pursuit. *Signal Processing*, *86*, 572–588.

Turlach, B., Venables, W., & Wright, S. (2005). Simultaneous variable selection. *Technometrics*, *47*, 349–364.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Royal Statistical Society*, *68*, 49–67.

Zhao, P., Rocha, G., & Yu, B. (2007). Grouped and hierarchical model selection through composite absolute penalties. *University of California, Berkeley, Statistics Technical Report 703*.