Naïve Bayes, Gaussian Distributions, Practical Applications

Required reading:

• Mitchell draft chapter, sections 1 and 2. (available on class website)

Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

Ground Hog's Day, 2009

Overview

Recently:

- learn P(Y|X) instead of deterministic f:X→Y
- Bayes rule
- MLE and MAP estimates for parameters of P
- Conditional independence
- classification with Naïve Bayes

Today:

- Text classification with Naïve bayes
- Gaussian distributions for continuous X
- Gaussian Naïve Bayes classifier
- Image classification with Naïve bayes

Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?

Article from rec.sport.hockey

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT
```

I can only comment on the Kings, but the most obvious candidate for pleasant surprise is Alex Zhitnik. He came highly touted as a defensive defenseman, but he's clearly much more than that. Great skater and hard shot (though wish he were more accurate). In fact, he pretty much allowed the Kings to trade away that huge defensive liability Paul Coffey. Kelly Hrudey is only the biggest disappointment if you thought he was any good to begin with. But, at best, he's only a mediocre goaltender. A better choice would be Tomas Sandstrom, though not through any fault of his own, but because some thugs in Toronto decided

Baseline: Bag of Words Approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)}$$

Sport ∈ {hockey, bball}

Assuming conditional independence among X_i's:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, classification rule for $X^{new} = \langle X_1, \ldots, X_n \rangle$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$X_i \in \{1, 0\} \rightarrow \theta_i = P(X_i = 1)$; $(1 - \theta_i) = P(X_i = 0)$

X_i = count of word$_i$ $\in \{0, 1, \ldots\}$

Learning to Classify Text

Target concept $Interesting?: Document \rightarrow \{+, -\}$

1. Represent each document by vector of words
   - one attribute per word position in document
2. Learning: Use training examples to estimate
   - $P(+)$
   - $P(-)$
   - $P(doc|+)$
   - $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|v_j)$$

i th location

where $P(a_i = w_k|v_j)$ is probability that word in position i is w_k, given v_j

θ_{kj}

j th class

k th word

one more assumption:
$P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

Twenty NewsGroups

Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy

LEARN_NAIVE_BAYES_TEXT($Examples, V$)

1. *collect all words and other tokens that occur in Examples*

~50k

- $Vocabulary \leftarrow$ all distinct words and other tokens in $Examples$

50K − 1 params

2. *calculate the required* $P(v_j)$ *and* $P(w_k|v_j)$ *probability terms*

- For each target value v_j in V do
  - $docs_j \leftarrow$ subset of $Examples$ for which the target value is v_j
  - $P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$
  - $Text_j \leftarrow$ a single document created by concatenating all members of $docs_j$
  - $n \leftarrow$ total number of words in $Text_j$ (counting duplicate words multiple times)
  - for each word w_k in $Vocabulary$
    * $n_k \leftarrow$ number of times word w_k occurs in $Text_j$
    * $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$

MLE

For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"

CLASSIFY_NAIVE_BAYES_TEXT(Doc)

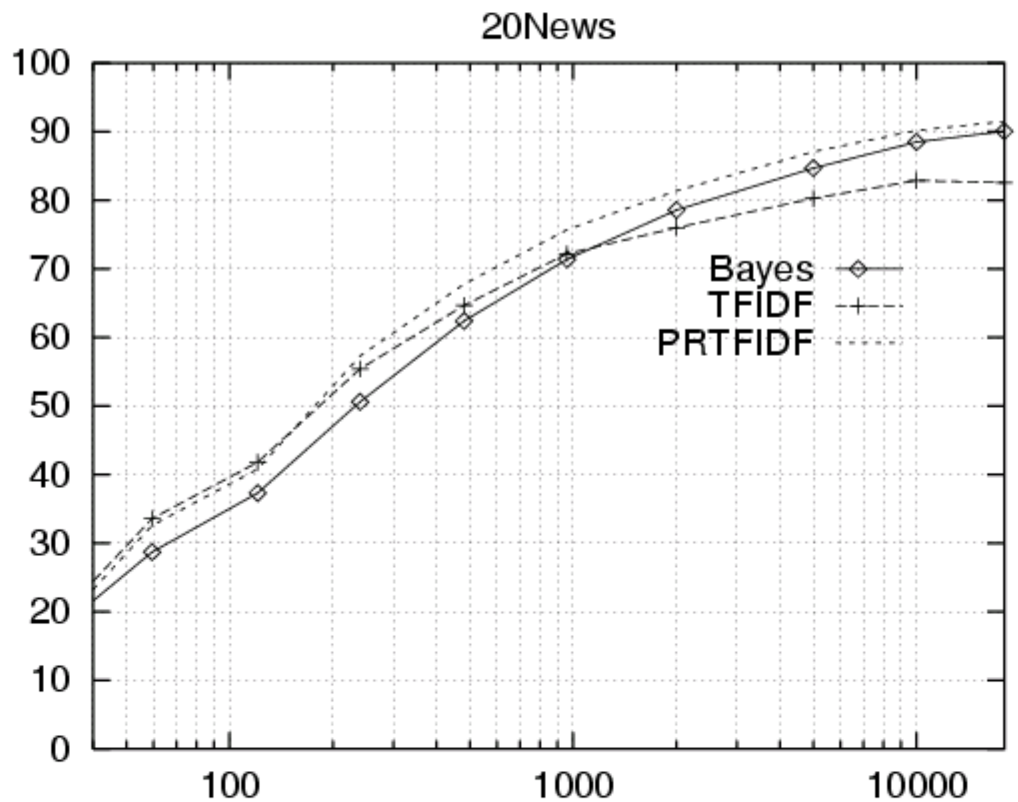
- $positions \leftarrow$ all word positions in Doc that contain tokens found in $Vocabulary$
- Return v_{NB}, where

$$v_{NB} = \operatorname*{argmax}_{v_j \in V} \underbrace{P(v_j) \prod_{i \in positions} P(a_i|v_j)}$$

$$v_{NB} = \operatorname*{argmax}_{v_j} \log(\uparrow) \qquad \log P(v_j) + \sum_i \log P(a_i|v_j)$$

$$P(v = v_j | Doc) = \frac{P(v_j) \prod P(a_i | v_j)}{P(Doc)}$$

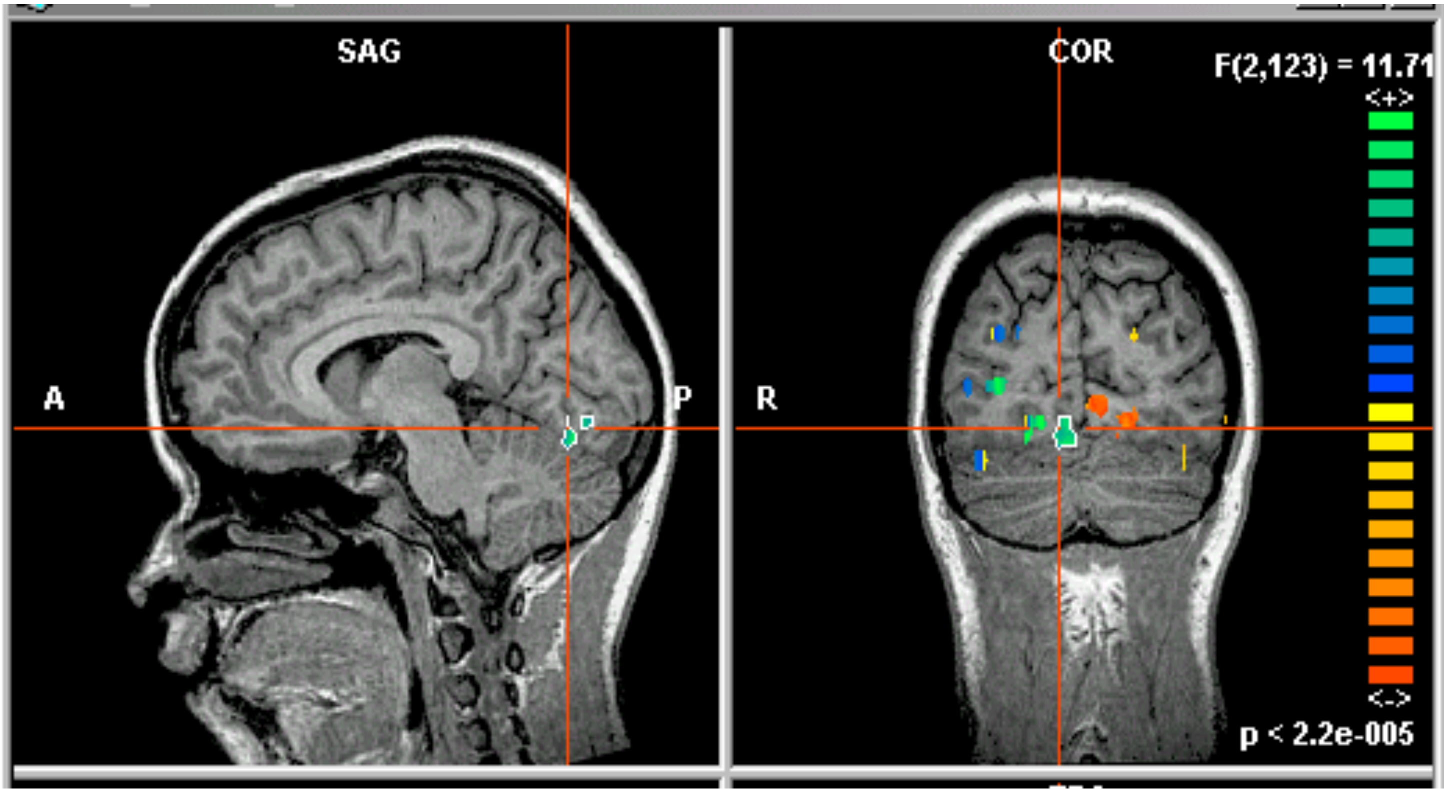
Learning Curve for 20 Newsgroups



Accuracy vs. Training set size (1/3 withheld for test)

What if we have continuous X_i ?

Eg., image classification: X_i is real-valued ith pixel



What if we have continuous X_i ?

Eg., image classification: X_i is real-valued i[th] pixel

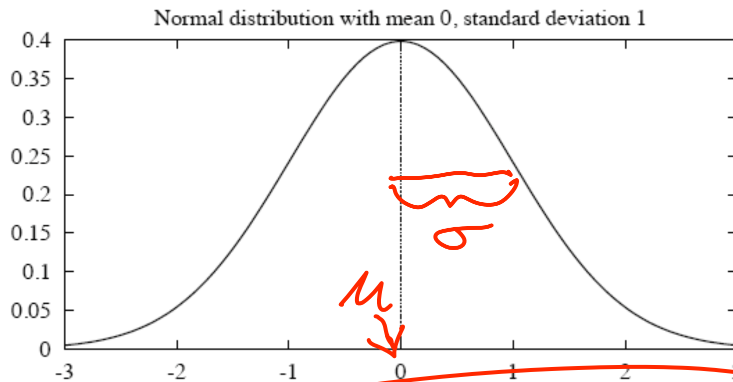
Naïve Bayes requires $P(X_i \mid Y_k)$, but X_i is real (continuous)

$$P(Y/X) = \frac{P(Y)\,P(X\mid Y)}{P(Y)} \xrightarrow{C.I.} \frac{P(Y)\prod_i P(X_i\mid Y)}{P(Y)}$$

Common approach: assume $P(X_i \mid Y_k)$ follows a normal (Gaussian) distribution

Gaussian Distribution

(also known as “Normal” distribution)



μ = mean

σ = std. dev

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that X will fall into the interval (a, b) is given by

$$P(a<X<b) = \int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- Variance of X is

$$Var(X) = \sigma^2$$

- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

p(x) is a *probability density function*, whose integral (not sum) is 1

What if we have continuous X_i ?

1 par

$P(X|Y)$, $P(Y)$

2n×2

Eg., image classification: X_i is i[th] pixel

Gaussian Naïve Bayes (GNB): assume

Normaliz.

$$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2}$$

for single val of Y, 2n params for P(X|Y)
for both vals , 2n × 2

How many params?
for Y boolean, $X = \langle X_1 \ldots X_N \rangle$

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

$1 = \int_{x,y} P(x, y)$

Gaussian Naïve Bayes Algorithm – continuous X_i (but still discrete Y)

- Train Naïve Bayes (examples)

  for each value y_k

  estimate* $\pi_k \equiv P(Y = y_k)$

  for each attribute X_i estimate

  class conditional mean μ_{ik}, variance σ_{ik}

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

jth training example

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

ith feature

kth class

δ(z)=1 if z true, else 0

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, $X=\langle X_1, \ldots X_n \rangle$?

What is form of decision surface for Gaussian Naïve Bayes classifier?

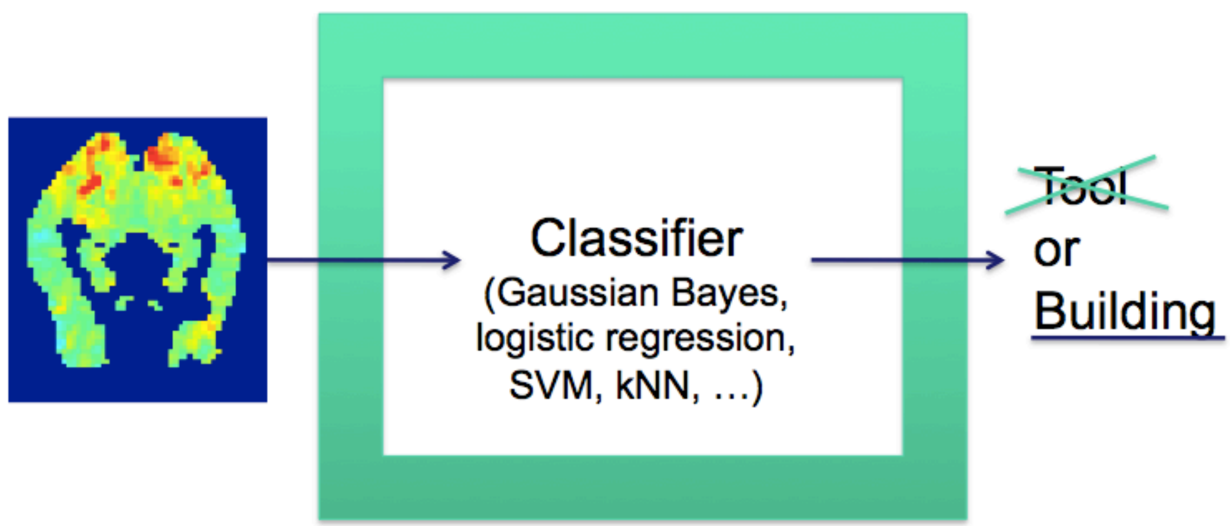
eg., if distributions are spherical, attributes have same variance

($\sigma_{ik} = \sigma$)

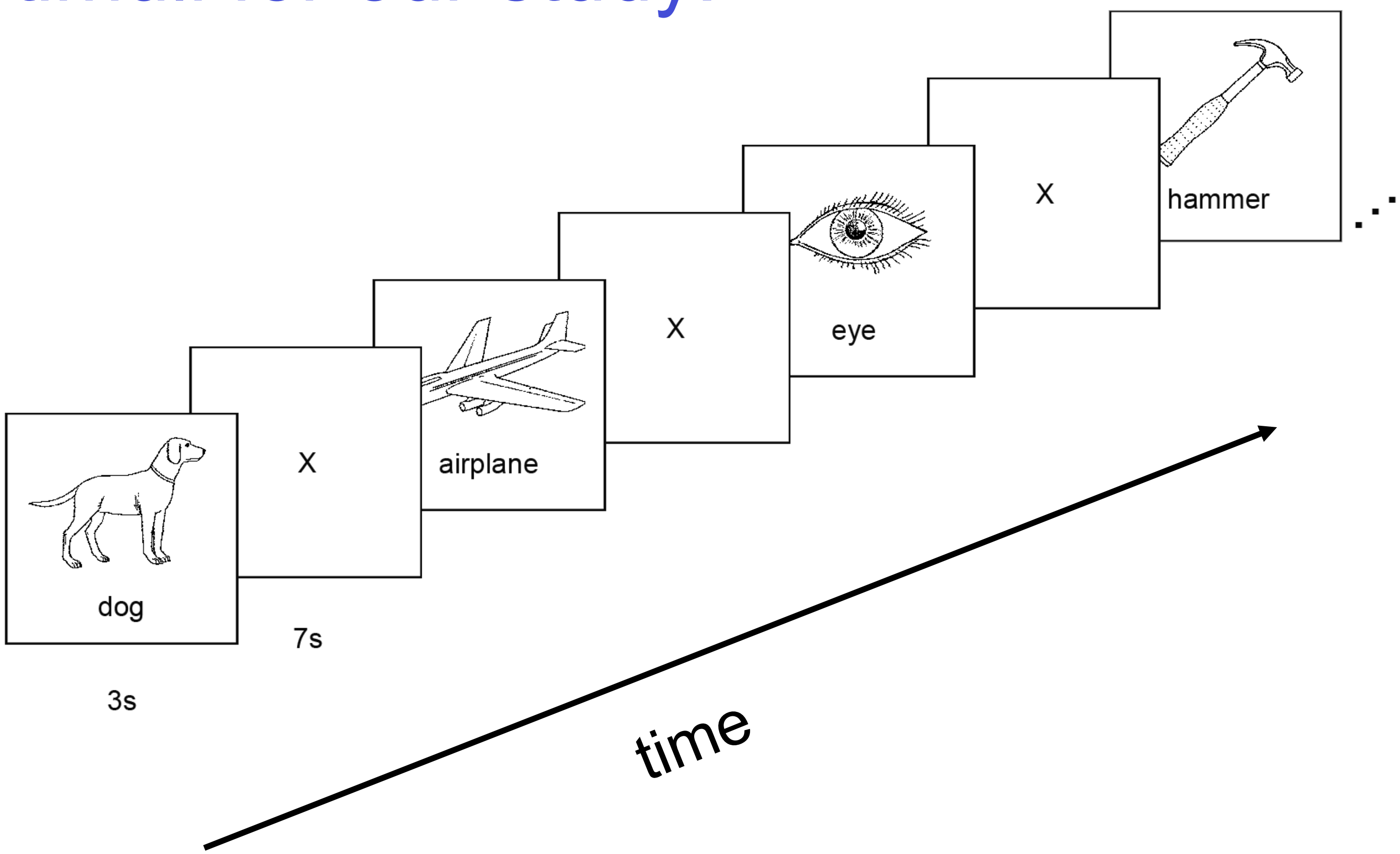
What is form of decision surface for Naïve Bayes classifier?

GNB Example: Classify a person's cognitive activity, based on brain image

- reading a sentence or viewing a picture?
- reading the word "Hammer" or "Apartment"
- viewing a vertical or horizontal line?
- answering the question, or getting confused?



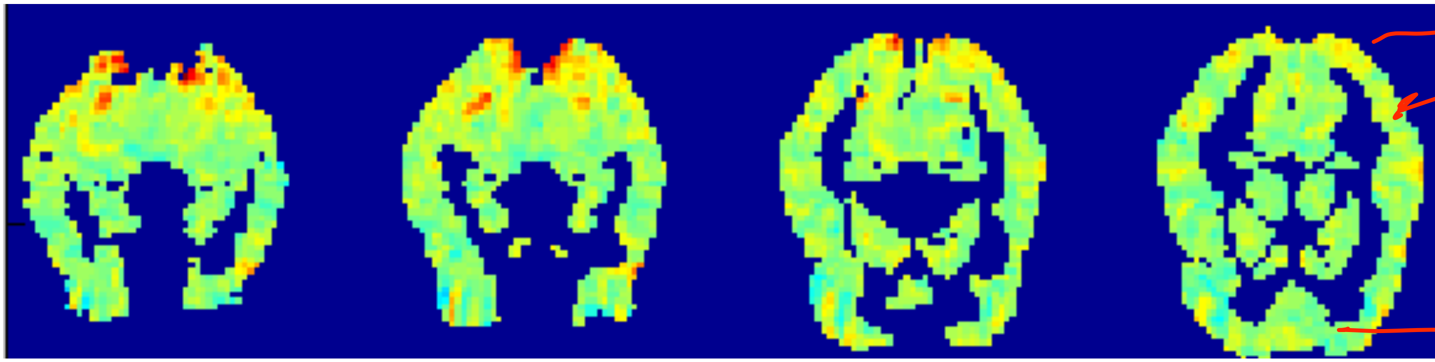
Stimuli for our study:

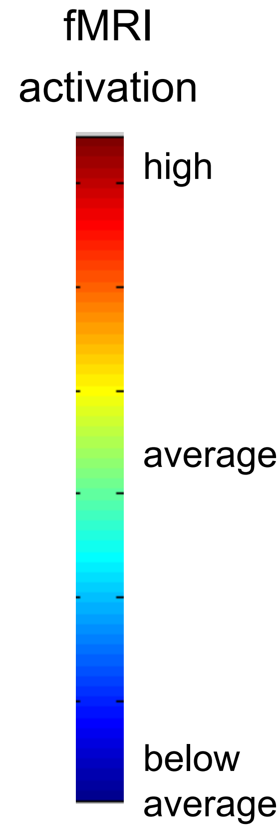


60 distinct exemplars, presented 6 times each

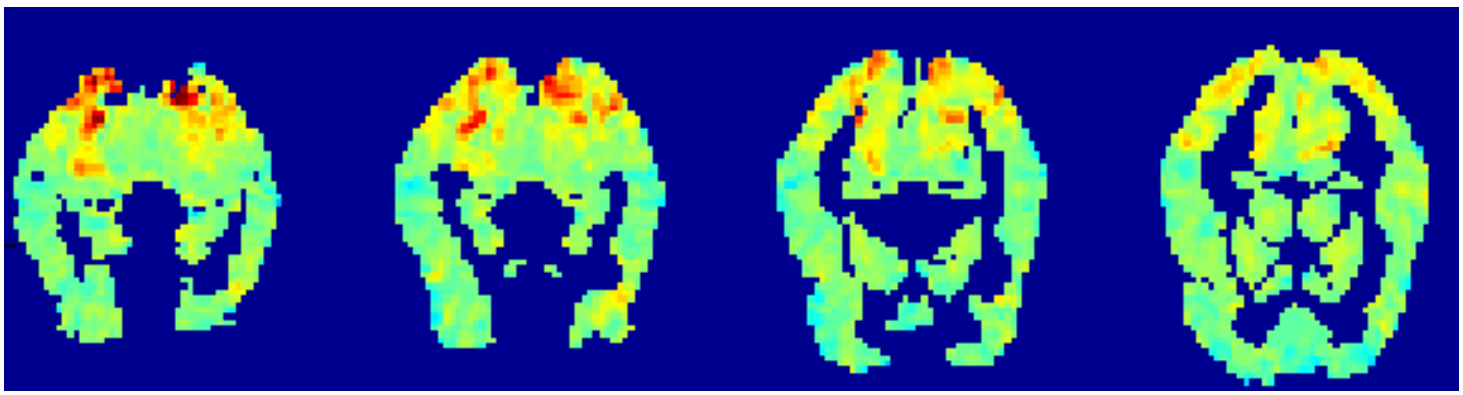
fMRI voxel means for “bottle”: means defining P(Xi | Y=“bottle)

$\mu_{i,bottle}$

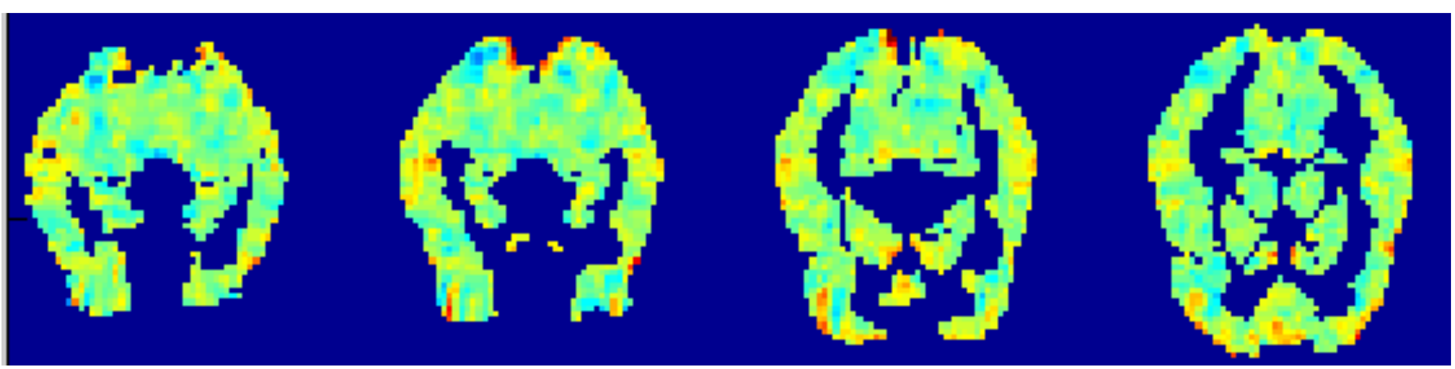




Mean fMRI activation over all stimuli:



“bottle” minus mean activation:



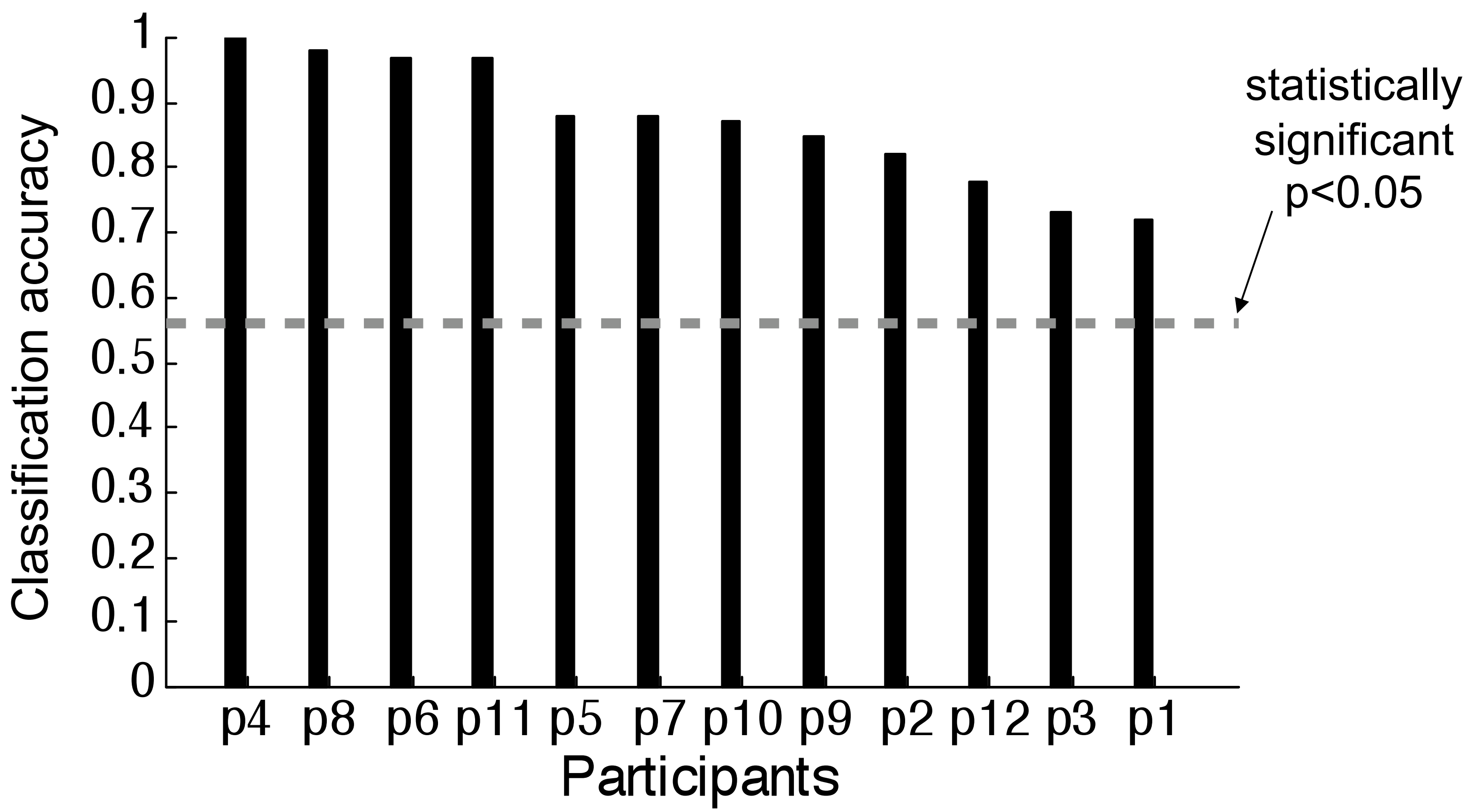
Training Classifiers over fMRI sequences

- Learn the classifier function
  Mean(fMRI(t+4), ...,fMRI(t+7)) → WordCategory
  - Leave one out cross validation

X → Y

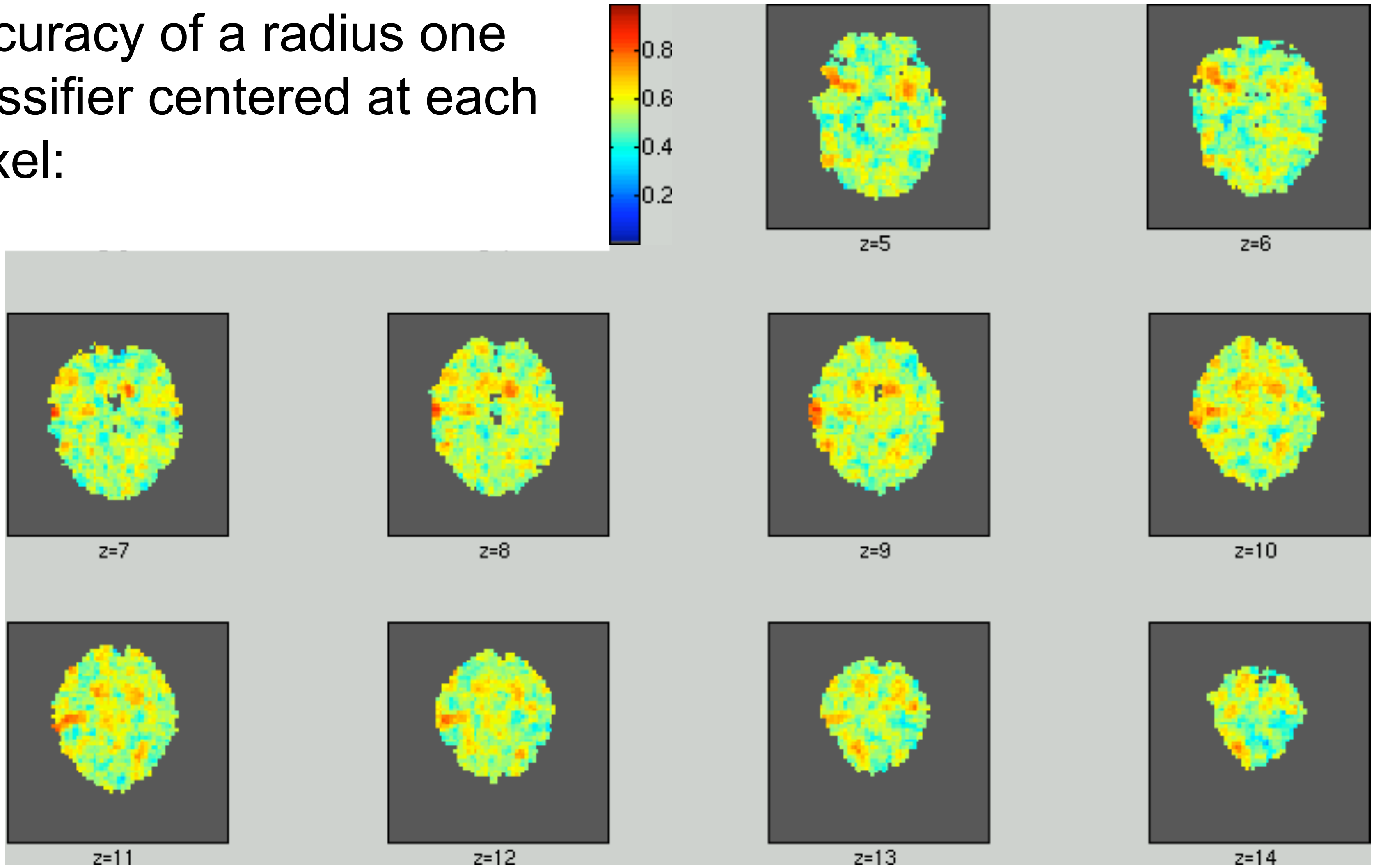
- Preprocessing:
  - Adjust for head motion
  - Convert each image x to standard normal image $x(i) \leftarrow \frac{x(i) - \mu_x}{\sigma_x}$
- Learning algorithms tried:
  - kNN (spatial correlation)
  - SVM
  - SVDM
  - Gaussian Naïve Bayes
  - Regularized Logistic regression
- Feature selection methods tried:
  - Logistic regression weights, voxel stability, activity relative to fixation,...

Classification task: is person viewing a "tool" or "building"?



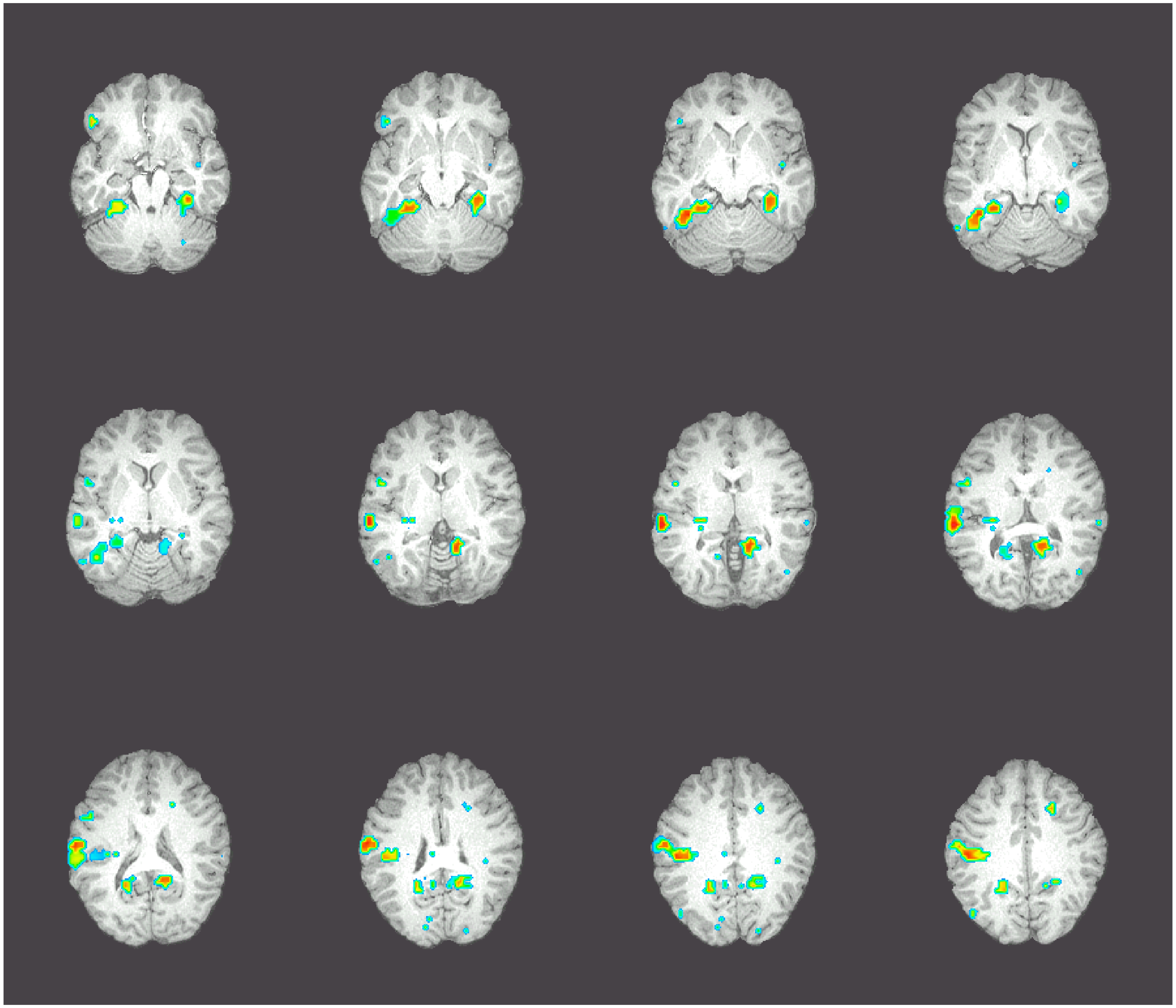
Where in the brain is activity that distinguishes tools vs. buildings?

Accuracy of a radius one classifier centered at each voxel:



voxel clusters: searchlights

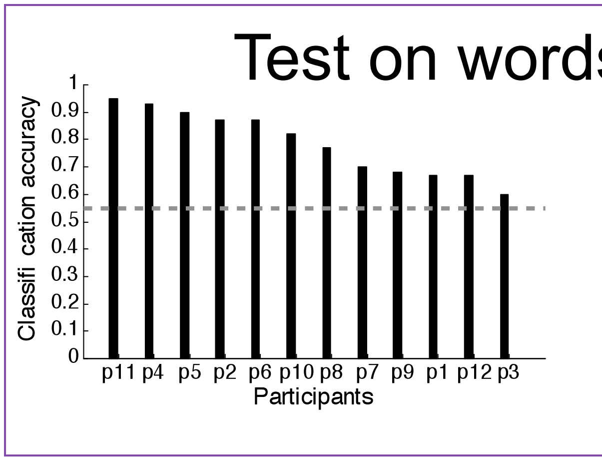
Accuracies of cubical 27-voxel classifiers centered at each significant voxel [0.7-0.8]

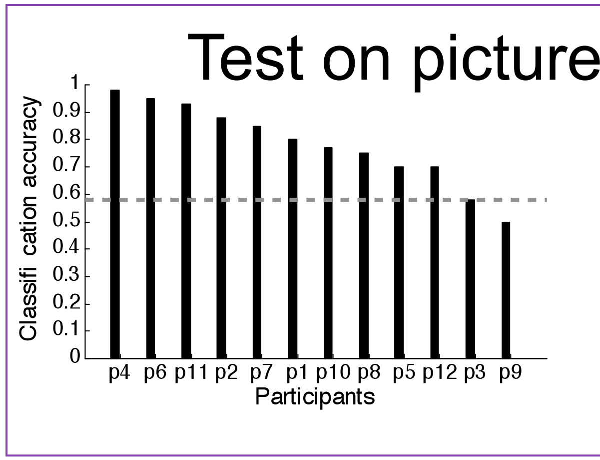


Are classifiers detecting neural representations of <u>meaning</u> or <u>perceptual features</u>?

ML: Can we train on word stimuli, then decode picture stimuli?

<u>YES</u>: We can train classifiers when presenting English words, then decode category of picture stimuli, or Portuguese words



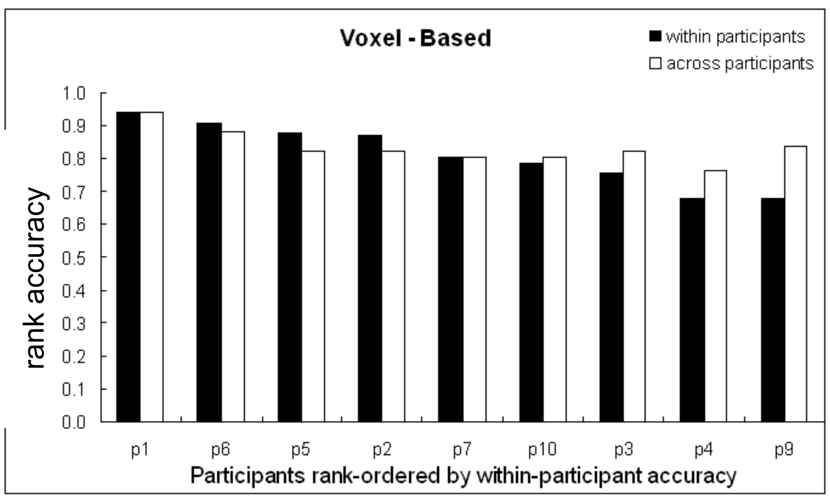


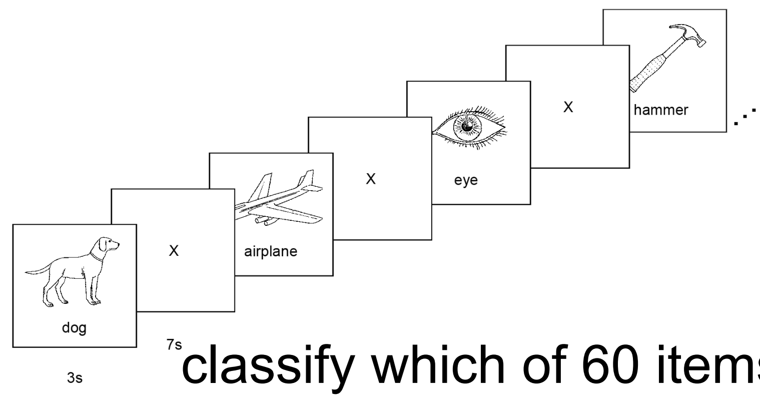
Therefore, the learned neural activation patterns must capture how the brain represents the <u>meaning</u> of input stimulus

Are representations similar across different people?

ML: Can we train classifier on data from a collection of people, then decode stimuli for a new person?

YES: We can train on one group of people, and classify fMRI images of new person





classify which of 60 items

Therefore, seek a theory of neural representations common to all of us (and of how we vary)

What you should know:

- Training and using classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it’s important
- Naïve Bayes
  - What it is
  - Why we use it so much
  - Training using MLE, MAP estimates
  - Discrete variables (Bernoulli) and continuous (Gaussian)

Questions to think about:

- Can you use Naïve Bayes for a combination of discrete and real-valued X_i?
- How can we easily model just 2 of n attributes as dependent?
- What does the decision surface of a Naïve Bayes classifier look like?
- How would you select a subset of X_i's?