Advanced Data Science

Dr. Kira Radinsky

Slides Adapted from Tom M. Mitchell

Agenda

Topics Covered:

- Naive Bayes
- Gaussian Naive Bayes

Additional Reading:

- Bishop Ch. 1 thru 1.2.3
- Bishop Ch. 2 thru 2.2
- Andrew Moore's online tutorial (http://web.engr.oregonstate.edu/~xfern/classes/cs434/slides/prob-5-slides.pdf)
- Mitchell: "Naïve Bayes and Logistic Regression" (http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf)

Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Learning and the joint distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Suppose we want to learn the function f: <G, H> → W

Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) = 0.024 / (0.024+0.25)

Estimating Parameters

- **Maximum Likelihood Estimate (MLE)**: choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} \quad P(\mathcal{D} \mid \theta)$$

- **Maximum a Posteriori (MAP)** estimate: choose θ that is most probable given prior probability and the data

$$\widehat{\theta} = \arg\max_{\theta} \quad P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} \quad = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

Naïve Bayes in a Nutshell

Represent the joint probability $P(X,Y)$ and estimate its parameters via MLE or MAP

Naïve Bayes in a Nutshell

Bayes rule:

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

Assuming conditional independence among X_i's:

$$P(Y = y_k | X_1 \ldots X_n) =$$

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

$$Y^{new} \leftarrow \arg\max_{y_k}$$

Naïve Bayes in a Nutshell

Bayes rule:

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

Assuming conditional independence among X_i's:

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

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

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

Another way to view Naïve Bayes (Boolean Y): Decision rule: is this quantity greater or less than 1?

$$\frac{P(Y=1|X_1 \ldots X_n)}{P(Y=0|X_1 \ldots X_n)} = \frac{P(Y=1)\prod_i P(X_i|Y=1)}{P(Y=0)\prod_i P(X_i|Y=0)}$$

Another way to view Naïve Bayes (Boolean Y): Decision rule: is this quantity greater or less than 1?

$$1 \gtrless \frac{P(Y=1|X_1 \ldots X_n)}{P(Y=0|X_1 \ldots X_n)} = \frac{P(Y=1)\prod_i P(X_i|Y=1)}{P(Y=0)\prod_i P(X_i|Y=0)}$$

$$0 \gtrless \log \frac{P(Y=1|X_1 \cdots X_n)}{P(Y=0|X_1 \cdots X_n)} = \log \frac{P(Y=1)}{P(Y=0)} + \sum_i \log \left[\frac{P(X_i|Y=1)}{P(X_i|Y=0)}\right]$$

$$\hat{\theta}_{ik} = \hat{P}(X_i=1|Y=k) \qquad 0 \gtrless \log \frac{P(Y=1)}{P(Y=0)} + \sum_i \left[X_i \log \frac{\theta_{i1}}{\theta_{i0}} + (1-X_i) \log \frac{(1-\theta_{i1})}{(1-\theta_{i0})}\right]$$

$$1-\hat{\theta}_{ik} = \boxed{\hat{P}(X_i=0|Y=k)}$$

Naïve Bayes: classifying text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?

> I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.
>
> I would like to thank Frank Pfenning, who has served ably in this role for the past two years.
>
> *********************************
>
> Randal E. Bryant
> Dean and University Professor

How shall we represent text documents for Naïve Bayes?

Learning to classify documents: P(Y|X)

Y discrete valued.

- e.g., Spam or not

$X = \langle X_1, X_2, \ldots X_n \rangle$ = document

X_i is a random variable describing…

I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.

I would like to thank Frank Pfenning, who has served ably in this role for the past two years.

****

Randal E. Bryant
Dean and University Professor

Learning to classify documents: P(Y|X)

Y discrete valued.

- e.g., Spam or not

$X = \langle X_1, X_2, \ldots X_n \rangle$ = document

X_i is a random variable describing…

Answer 1:

X_i is boolean, 1 if word i is in document, else 0 e.g., $X_{pleased} = 1$

Issues?

> I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.
>
> I would like to thank Frank Pfenning, who has served ably in this role for the past two years.
>
> ********************************
>
> Randal E. Bryant
> Dean and University Professor

Learning to classify documents: P(Y|X)

Y discrete valued.

- e.g., Spam or not

$X = <X_1, X_2, \ldots X_n>$ = document

X_i is a random variable describing…

I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.

I would like to thank Frank Pfenning, who has served ably in this role for the past two years.

****

Randal E. Bryant
Dean and University Professor

<u>Answer 2</u>:

X_i represents the i^{th} *word position* in document

- X_1 = "I", X_2 = "am", X_3 = "pleased"
- and, let's assume the X_i are iid (indep, identically distributed) $P(X_i|Y) = P(X_j|Y) \quad (\forall i, j)$

Learning to classify document: P(Y|X) the "Bag of Words" model

- Y discrete valued. e.g., Spam or not
- $X = <X_1, X_2, \ldots X_n>$ = document

- X_i are iid random variables. Each represents the word at its position i in the document
- Generating a document according to this distribution = rolling a 50,000 sided die, once for each word position in the document

- The observed counts for each word follow a ??? distribution

Multinomial Distribution

- P(θ) and P(θ|D) have the same form

50000

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

Count for side K

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

Multinomial Bag of Words

ebay Shop by category | Search... | All Categories | Search | Advanced

Sharp shooting

Shot detailed, realistic images thanks to the 20-megapixel sensor which is powered by DIGIC 4+ processor.

Images are sharp and video is smooth thanks to Canon's Image Stabilisation which combats camera shake – ideal when using the 8 x optical zoom.

User-friendly

Simply point and shoot – there's no need to adjust settings every time you want to take a picture thanks to Smart Auto with Auto Button which optimises the settings for each scene.

Framing the perfect portrait is easy thanks to the 2.7" screen with adjustable brightness and Face Detection technology.

Shooting HD video with the Canon **IXUS 175 Compact Camera** is also made simple thanks to the dedicated movie button.

Technical specifications for CANON IXUS 175 Compact Camera - Silver

OVERVIEW	
Camera type	Compact camera
Processor	DIGIC 4+ with iSAPS technology
SENSOR	
Resolution	20 megapixels
Type	CCD
Size	1/2.3" / 6.17 x 4.55 mm
ISO sensitivity	AUTO, 100 - 1600
Image stabiliser	Image Stabilisation (IS)
LENS	
Focal length	5 - 40 mm
35 mm equivalent	28 - 224 mm
Maximum aperture	f/3.2 - f/6.9
Minimum distance	1 cm
Normal distance	1 cm to infinity
Focusing	Autofocus

→

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

50000

MAP estimates for bag of words

Map estimate for multinomial

$$\theta_i = \frac{\alpha_i + \beta_i - 1}{\sum_{m=1}^{k} \alpha_m + \sum_{m=1}^{k} (\beta_m - 1)}$$

MLE

$$\theta_{aardvark} = P(X_i = \text{aardvark}) = \frac{\#\ \text{observed 'aardvark'} + \#\ \text{hallucinated 'aardvark'} - 1}{\#\ \text{observed words} + \#\ \text{hallucinated words} - k}$$

What β's should we choose?

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each value y_k

  P(Category = 'Phones')

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

  for each value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

  prob that word x_{ij} appears in position i, given $Y=y_k$

- Classify (X^{new})

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

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* Additional assumption: word probabilities are position independent $\theta_{ijk} = \theta_{mjk}$ for $i \neq m$

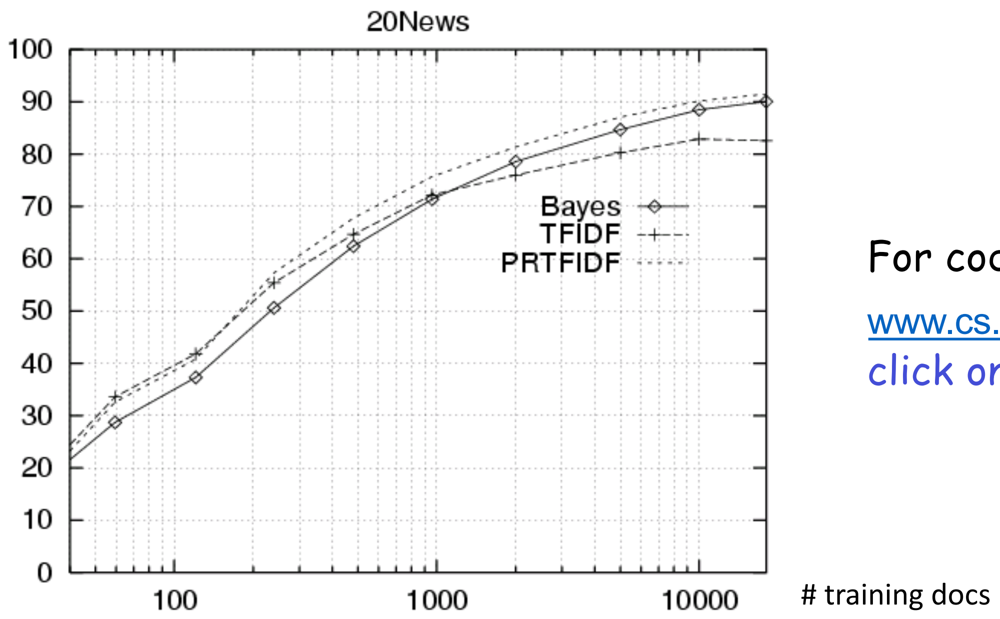
Twenty NewsGroups

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

Learning Curve for 20 Newsgroups



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

For code and data, see

www.cs.cmu.edu/~tom/mlbook.html

click on "Software and Data"

Summary

- **Maximum Likelihood Estimate (MLE)**: choose θ that maximizes probability of observed data $\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$

- **Maximum a Posteriori (MAP) Estimate**: choose θ that is most probable given prior probability and the data $\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$

- **Naive Bayes:** Represent the joint probability P(X,Y) and estimate its params via MLE or MAP
  - Representation of P(X,Y) done assuming bayes rule: $P(X,Y) = P(Y)P(X|Y)$
  - Training done by estimating the following parameters (via MLE or MAP):

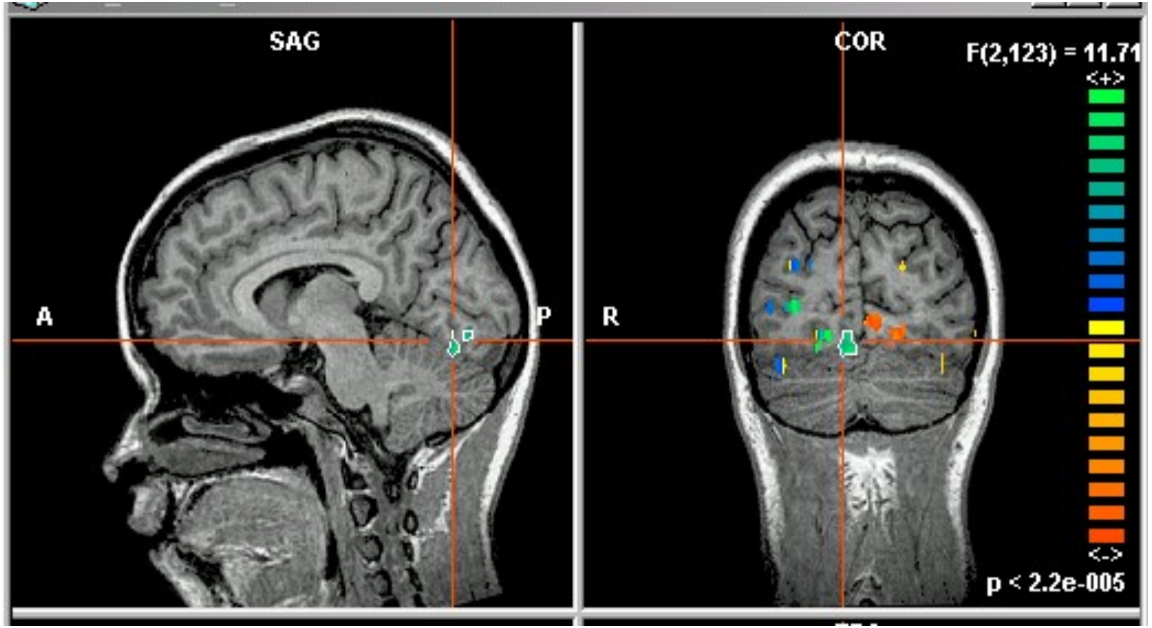
$$P(Y = y_k) \qquad \theta_{ijk} \equiv P(X_i = x_{ij}|Y = y_k)$$

  - Prediction:

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

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 $P(X_i \mid Y=y_k)$, but X_i is real (continuous)

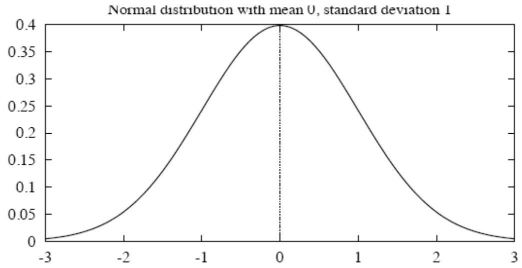
$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Common approach: assume $P(X_i \mid Y=y_k)$ follows a Normal (Gaussian) distribution

Gaussian Distribution

(also called "Normal")

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



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

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

$$\int_a^b p(x)dx$$

- 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

$$\sigma_X = \sigma$$

What if we have continuous X_i?

Gaussian Naïve Bayes (GNB): assume

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

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)

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 $P(X_i|Y = y_k)$

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

- Classify (X^{new})

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

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

* 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

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

ith feature

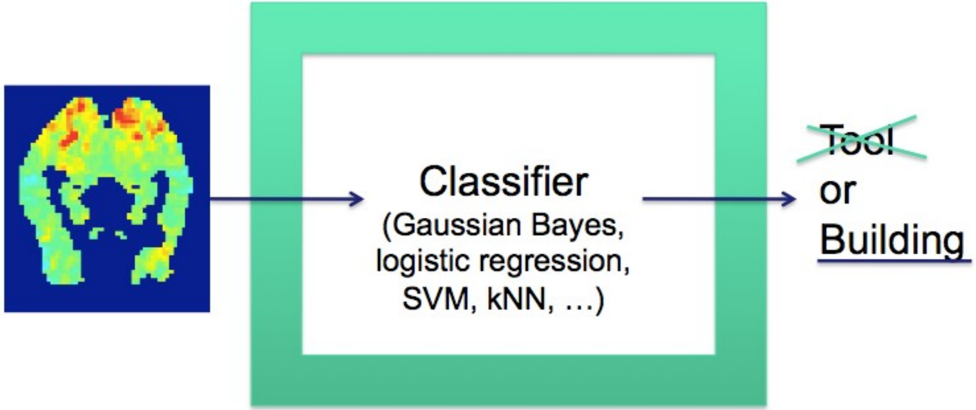
kth class

δ()=1 if ($Y^j=y_k$) else 0

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

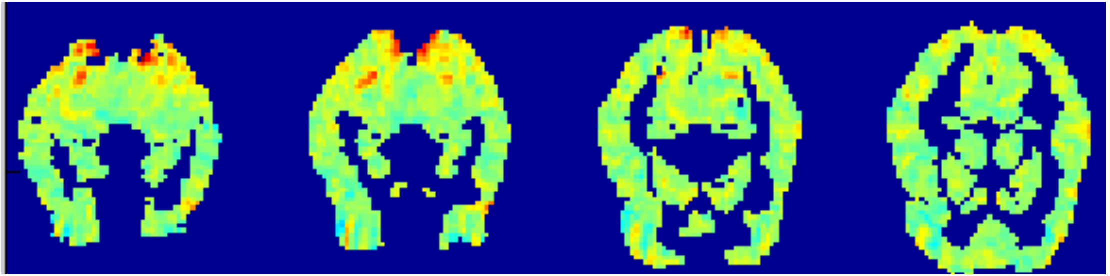
GNB Example: Classify a person's cognitive state, based on brain image

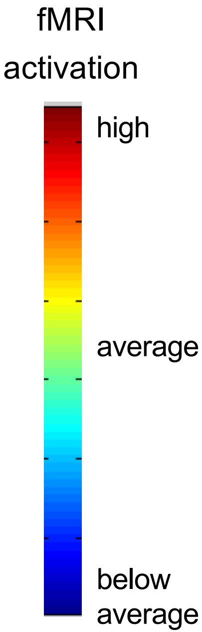
- reading a sentence or viewing a picture?
- reading the word describing a "Tool" or "Building"?
- answering the question, or getting confused?



Mean activations over all training examples for Y="bottle"

$\mu \mid Y = bottle$

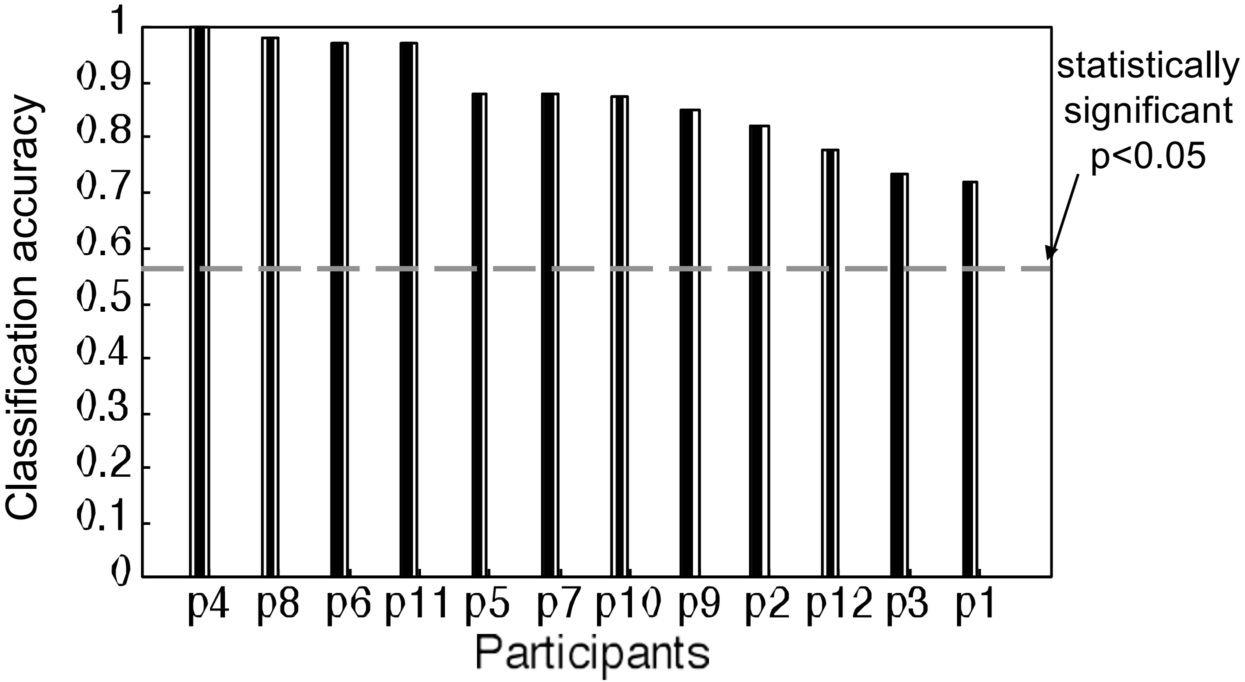




Y is the mental state (reading "house" or "bottle")
X_i are the voxel activities,

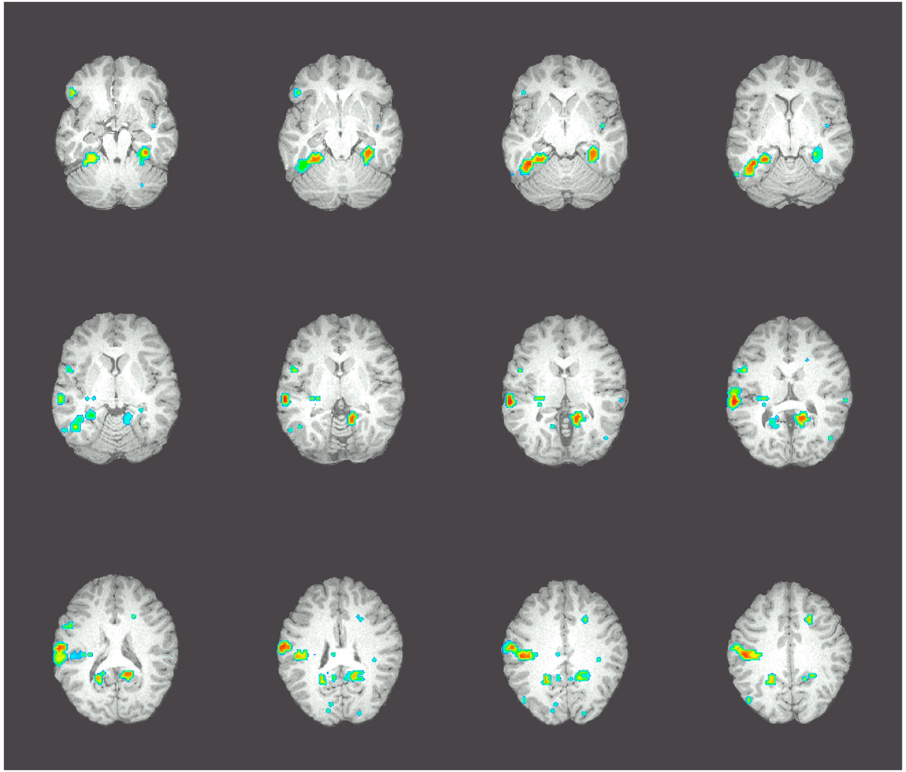
this is a plot of the μ's defining $P(X_i \mid Y=\text{"bottle"})$

Classification task: is person viewing a “tool” or “building”?



Where is information encoded in the brain?

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



Naïve Bayes: What you should know

- Designing classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it's important
- Naïve Bayes assumption and its consequences
  - Which (and how many) parameters must be estimated under different generative models (different forms for $P(X|Y)$)
    - and why this matters
- How to train Naïve Bayes classifiers
  - MLE and MAP estimates
  - with discrete and/or continuous inputs X_i

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?