

Bayesian Classifiers, Conditional Independence and Naïve Bayes

The slide features several light purple circles of varying sizes. One circle is positioned behind the word 'Bayesian' in the title. Another is behind 'Conditional'. A third is behind 'Independence'. A fourth is behind 'and'. A fifth is behind 'Naïve'. A sixth is behind 'Bayes'. Additionally, there are three more circles of the same color located in the lower half of the slide, arranged in a row.

Let's learn classifiers by learning $P(Y|X)$

- Suppose $Y = \text{Wealth}$, $X = \langle \text{Gender}, \text{HoursWorked} \rangle$

Gender	HrsWorked	$P(\text{rich} \mid G, \text{HW})$	$P(\text{poor} \mid G, \text{HW})$
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

How many parameters must we estimate?

- Suppose $X = \langle X_1, \dots, X_n \rangle$

Gender	<u>HrsWorked</u>	<u>P(rich G,HW)</u>	<u>P(poor G,HW)</u>
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

where X_i and Y are boolean RV's

To estimate $P(Y | X_1, X_2, \dots, X_n)$

- If we have 30 X_i 's instead of 2?

Can we reduce params by using Bayes Rule?

- Suppose $X = \langle X_1, \dots, X_n \rangle$
- where X_i and Y are boolean RV's

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Recall Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

Equivalently:

$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}$$

Naïve Bayes

- Naïve Bayes assumes

$$P(X_1 \dots X_n | Y) = \prod_i P(X_i | Y)$$

- i.e., that X_i and X_j are conditionally independent given Y , for all $i \neq j$

Conditional Independence

Definition: X is conditionally independent of Y given Z , if the probability distribution governing X is independent of the value of Y , given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

- Which we often write

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

E.g.,

$$P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y

- Given this assumption, then:

$$\begin{aligned}P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y)\end{aligned}$$

in general: $P(X_1 \dots X_n|Y) = \prod_i P(X_i|Y)$

- How many parameters to describe $P(X_1 \dots X_n|Y)$? $P(Y)$?
 - Without conditional independent assumption?
 - With conditional independent assumption?

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:

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

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

for each* value y_k

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

for each* value x_{ij} of each attribute X_i

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

- Classify (X^{new})

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

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

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

Estimating Parameters: Y, X_i discrete-valued

- Maximum likelihood estimates (MLE's):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in
dataset D for which $Y=y_k$

Naïve Bayes: Subtlety #1

- If unlucky, our MLE estimate for $P(X_i | Y)$ might be zero.
(e.g., X_{373} = Birthday_Is_January_30_1990)
- Why worry about just one parameter out of many?
- What can be done to avoid this?

Estimating Parameters

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

$$\hat{\theta} = \arg \max_{\theta} P(\mathcal{D} | \theta)$$

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

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\theta | \mathcal{D}) \\ &= \arg \max_{\theta} = \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}\end{aligned}$$

Estimating Parameters: Y, X_i discrete-valued

- Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

- MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + \alpha_k}{|D| + \sum_m \alpha_m}$$

Only difference:
"imaginary" examples

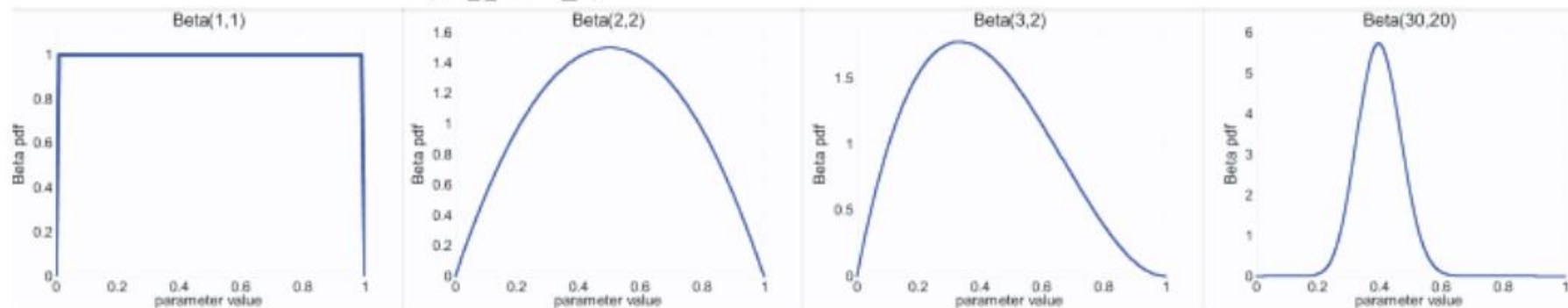
$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + \alpha'_k}{\#D\{Y = y_k\} + \sum_m \alpha'_m}$$

Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

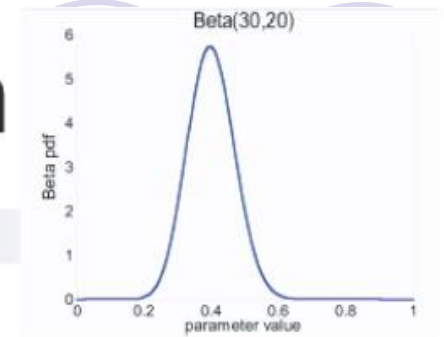
Mean:

Mode:



- Likelihood function: $P(\mathcal{D} | \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T}$
- Posterior: $P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$

MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- MAP: use most likely parameter:

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

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is “forgotten”
- **But, for small sample size, prior is important!**

Dirichlet distribution

- number of heads in N flips of a two-sided coin
 - follows a binomial distribution
 - Beta is a good prior (conjugate prior for binomial)
- what it's not two-sided, but k-sided?
 - follows a multinomial distribution
 - Dirichlet distribution is the conjugate prior

$$P(\theta_1, \theta_2, \dots, \theta_K) = \frac{1}{B(\alpha)} \prod_i^K \theta_i^{(\alpha_i - 1)}$$

Lejeune Dirichlet



Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	 Germany
Nationality	 German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

Naïve Bayes: Subtlety #2

- Often the X_i are not really conditionally independent
- We use Naïve Bayes in many cases anyway, and it often works pretty well
 - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])
- What is effect on estimated $P(Y|X)$?
 - Special case: what if we add two copies: $X_i = X_k$



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?

Baseline: Bag of Words Approach

the world of

TOTAL



all about the company

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

► All About The Company

- Global Activities
- Corporate Structure
- TOTAL's Story
- Upstream Strategy
- Downstream Strategy
- Chemicals Strategy
- TOTAL Foundation
- Homepage

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

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

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

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

one more assumption:

$$P(a_i = w_k | v_j) = P(a_m = w_k | v_j), \forall i, m$$

Twenty NewsGroups

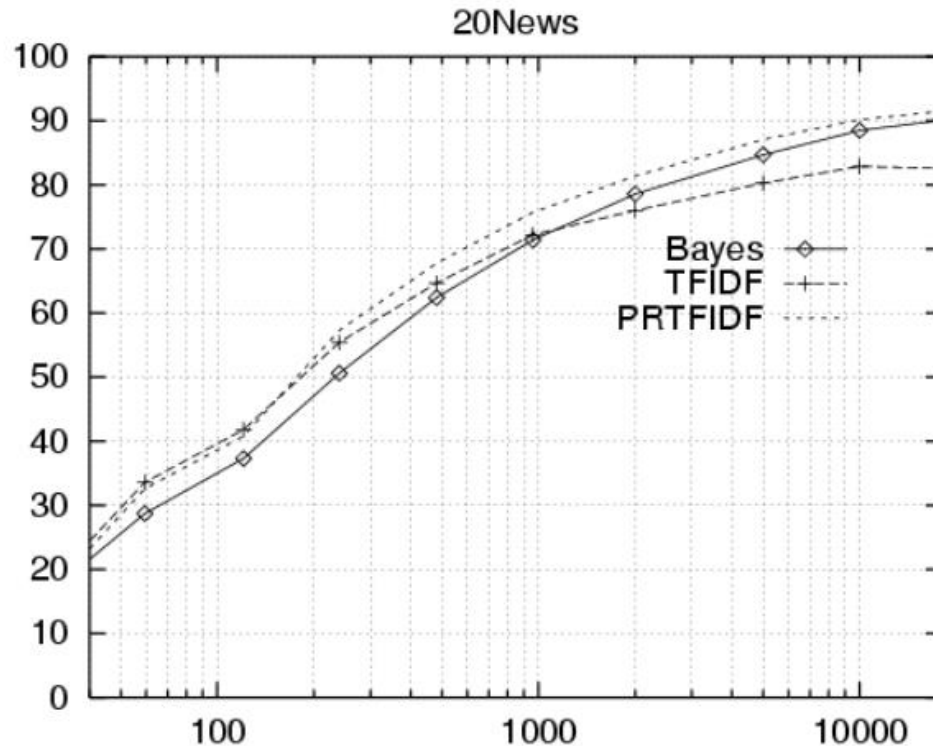


Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

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

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 i^{th} pixel



What if we have continuous X_i ?

- Eg., image classification: X_i is i^{th} pixel



Still have:

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

Just need to decide how to represent $P(X_i | Y)$

What if we have continuous X_i ?

- Eg., image classification: X_i is i^{th} pixel
- Gaussian Naïve Bayes (GNB): assume

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

- 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

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

- Classify (X^{new})

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

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

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

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

Diagram annotations for the equation above:

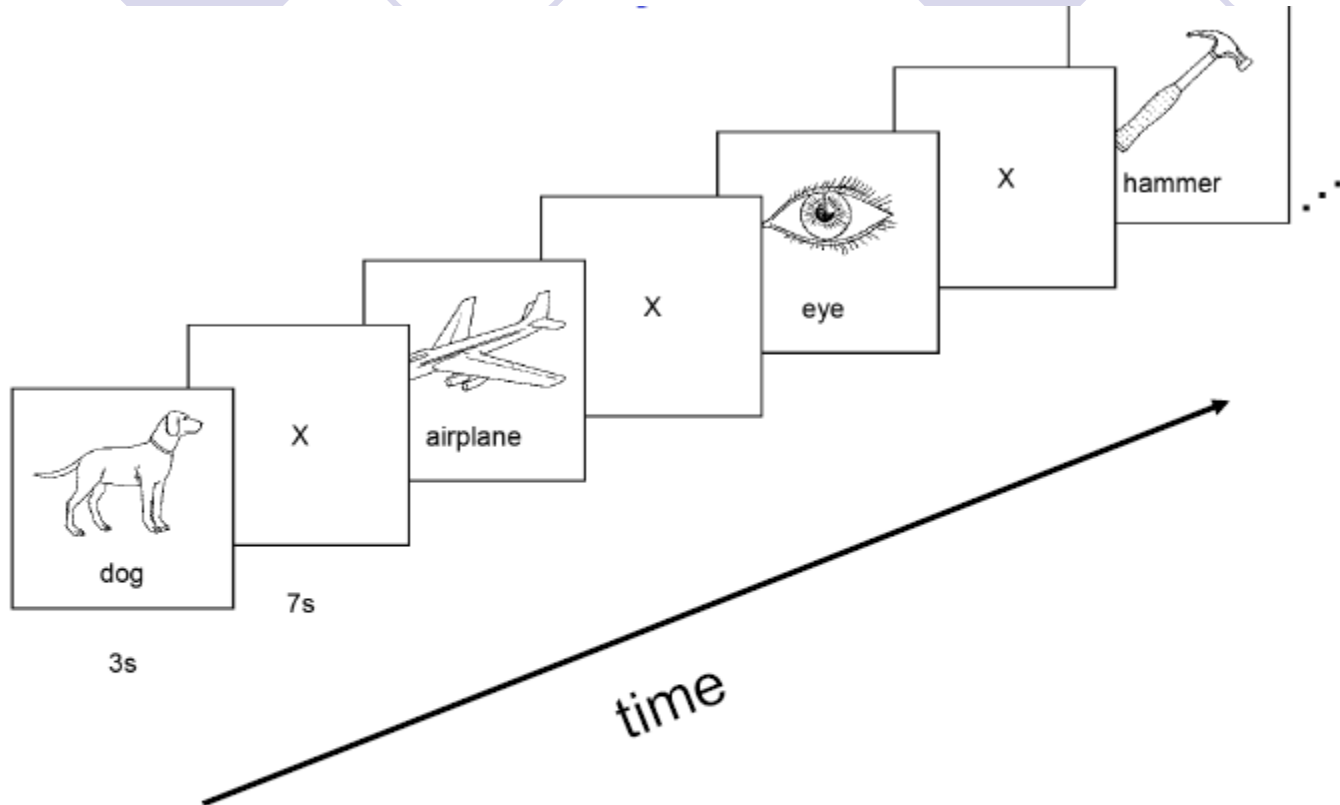
- $\hat{\mu}_{ik}$: ith feature, kth class
- X_i^j : jth training example
- $\delta(z) = 1$ if z true, else 0

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

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

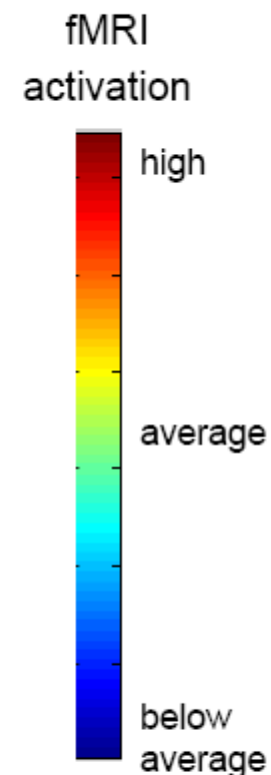
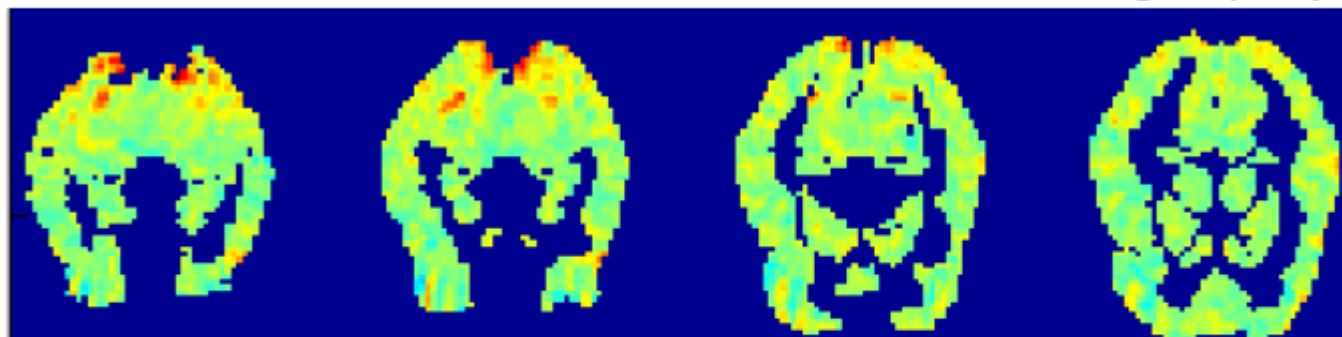
- are they reading a sentence or viewing a picture?
- reading the word “Hammer” or “Apartment”
- viewing a vertical or horizontal line?
- answering the question, or getting confused?

Stimuli for our study:

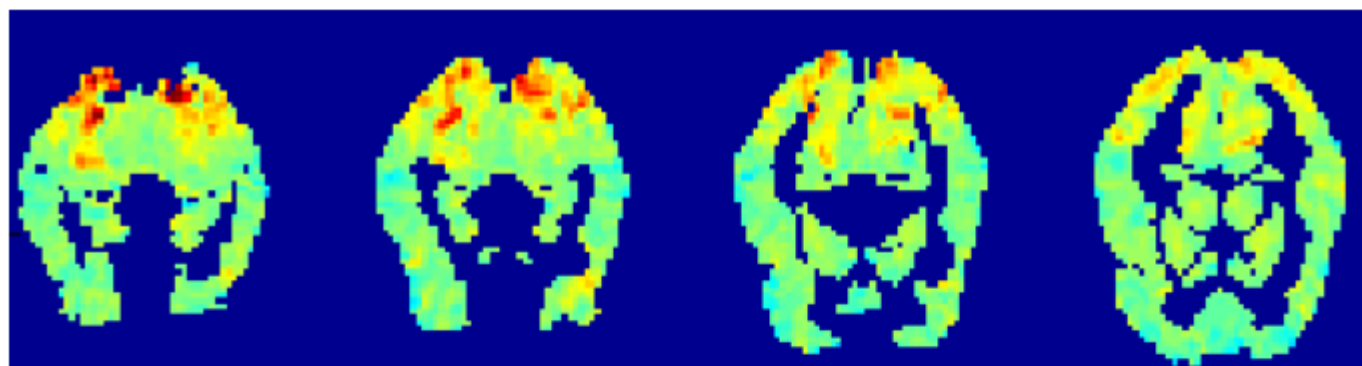


60 distinct exemplars, presented 6 times each

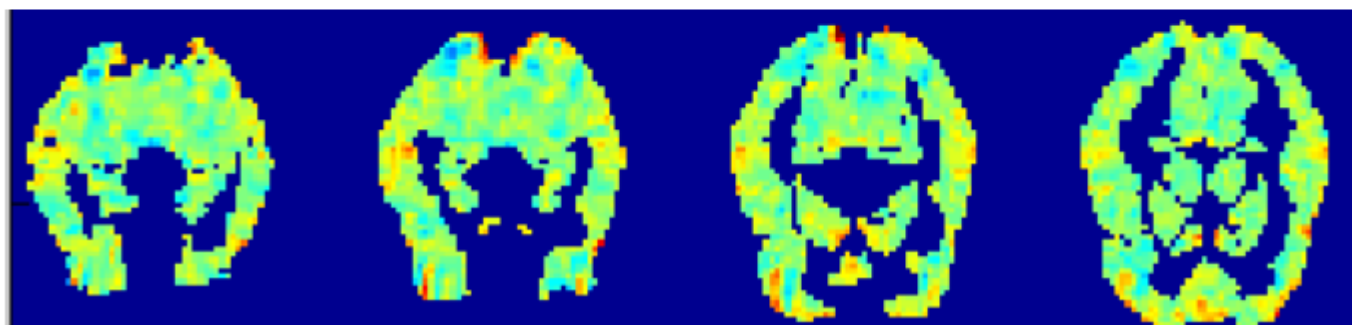
fMRI voxel means for “bottle”: means defining $P(X_i | Y=\text{“bottle”})$



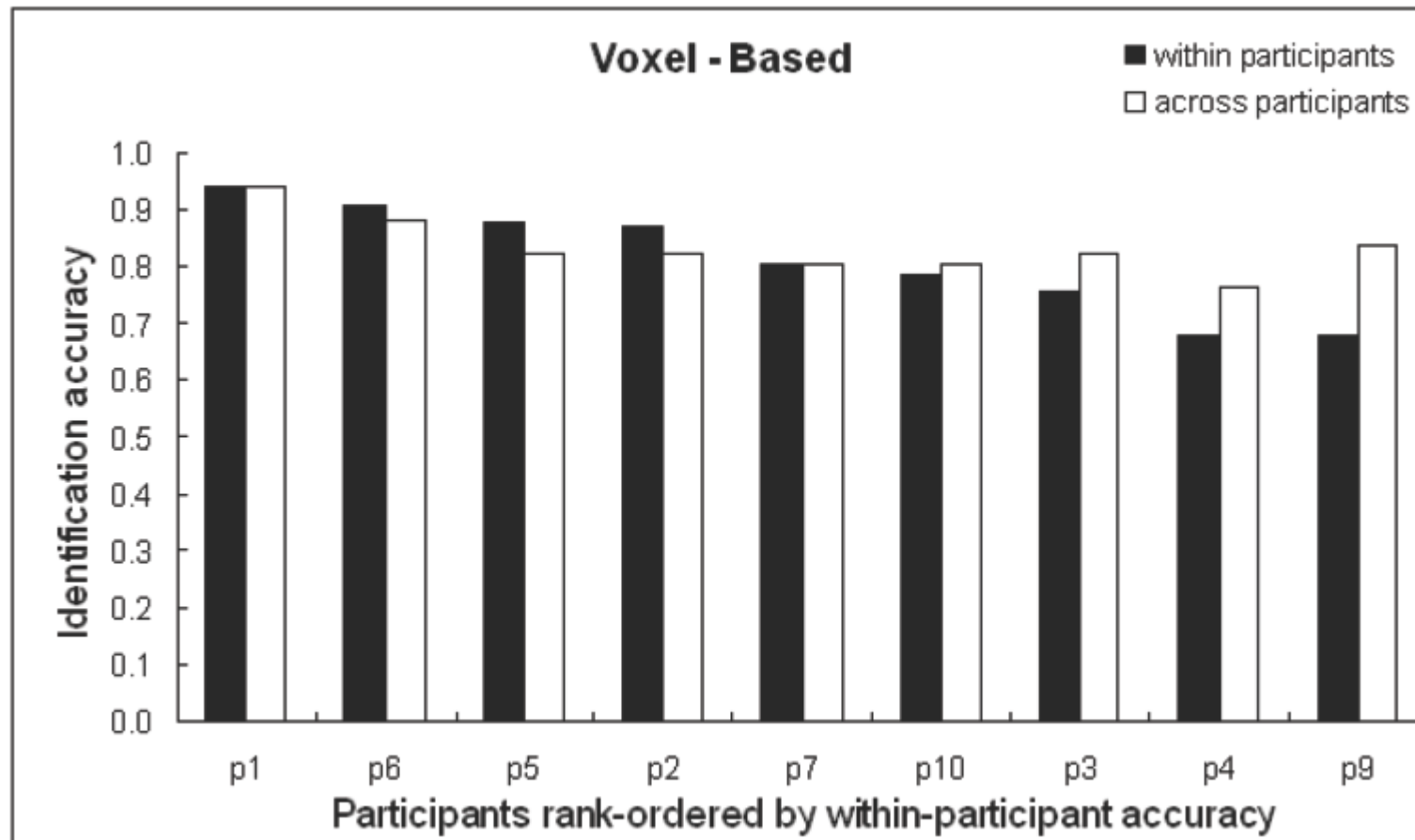
Mean fMRI activation over all stimuli:



“bottle” minus mean activation:



Rank Accuracy Distinguishing among 60 words



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 and continuous (Gaussian)

Questions:

- What is the error will classifier achieve if Naïve Bayes assumption is satisfied and we have infinite training data?
- 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?