# Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 20, 2011

Today:
- Bayes Classifiers
- Naïve Bayes
- Gaussian Naïve Bayes

Readings:
Mitchell:
"Naïve Bayes and Logistic Regression"
(available on class website)

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

Consider Y=Wealth, X=<Gender, HoursWorked>

| 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 |

| Gender | HrsWorked | P(rich \| G,HW) | P(poor \| G,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 $X = \langle X_1, \ldots X_n \rangle$
where $X_i$ and $Y$ are boolean RV's

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

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

$2^n$

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

$2^{30} \sim 1$ Billion

## 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:

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

## Can we reduce params using Bayes Rule?

Suppose X =<$X_1$,… $X_n$>
where $X_i$ and Y are boolean RV's

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

how many params for $P(X_1 \cdots X_n|Y)$ $(2^n-1)\cdot 2$

how many for $P(Y) = 1$

## Naïve Bayes

Naïve Bayes assumes

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

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

# Conditional Independence

Definition: X is <u>conditionally independent</u> 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.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

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

Given this assumption, then:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$ Chain rule

$$= P(X_1|Y)P(X_2|Y)$$ Cond. Indep.

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

How many parameters to describe $P(X_1...X_n|Y)$? $P(Y)$?

- Without conditional indep assumption? $2(2^n - 1) + 1$
- With conditional indep assumption? $2n + 1$

# 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:

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

So, classification rule for $X^{new} = < X_1, \ldots, X_n >$ 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 of these...

## Estimating Parameters: $Y$, $X_i$ discrete-valued

Maximum likelihood estimates (MLE's):

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

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

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

## Example: Live in Sq Hill? P(S|G,D,M)

80, 8

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- M=1 iff Rachel Maddow fan

What probability parameters must we estimate?

## Example: Live in Sq Hill? P(S|G,D,M)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- M=1 iff Rachel Maddow fan

| | | | |
|---|---|---|---|
| P(S=1) : | .4   32/80 = .4 | P(S=0) : | .6 |
| P(D=1 \| S=1) : | 6/32 = .188 | P(D=0 \| S=1) : | |
| P(D=1 \| S=0) : | 18/48 = .38 | P(D=0 \| S=0) : | |
| P(G=1 \| S=1) : | 29/32 = .91 | P(G=0 \| S=1) : | |
| P(G=1 \| S=0) : | 17/48 = .35 | P(G=0 \| S=0) : | |
| P(M=1 \| S=1) : | 2/32 = .06 | P(M=0 \| S=1) : | |
| P(M=1 \| S=0) : | 3/48 = .06 | P(M=0 \| S=0) : | .94 |

$P(S=1 \mid G\,M\,D) \Rightarrow P(S=1)\cdot P(G=g \mid S=1)\,P(D=d \mid S=1)\,P(M=m \mid S=1)$

Test: $\langle S=1\;\; G=1\;\; D=1\;\; M=0 \rangle$

$P(S=1 \mid X^{Test}) \propto .4 \cdot .91 \cdot .19 \cdot .94$

$P(S=0 \mid X^{Test}) \propto .6 \cdot .35 \cdot .38 \cdot .94$

$P(S=0 \mid X^{Test}) = .54$

## Naïve Bayes: Subtlety #1

If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero. (e.g., $X_i$ = Birthday_Is_January_30_1990)

- Why worry about just one parameter out of many?

$$P(Y \mid X) \propto P(Y) \prod_i \underbrace{P(X_i = x_i^{New} \mid Y)}_{0}$$

- What can be done to avoid this?

## Estimating Parameters

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

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

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

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

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

## Estimating Parameters: $Y$, $X_i$ discrete-valued

Maximum likelihood estimates:

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

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

MAP estimates (Beta, Dirichlet priors):

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

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

Only difference: "imaginary" examples

## Estimating Parameters

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

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

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

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

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

## Conjugate priors

- $P(\theta)$ and $P(\theta|D)$ have the same form

Eg. 1 Coin flip problem

Likelihood is ~ Binomial

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If prior is Beta distribution,

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

Then posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

**For Binomial, conjugate prior is Beta distribution.**

[A. Singh]

## Conjugate priors

- $P(\theta)$ and $P(\theta|D)$ have the same form

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

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

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

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.**

[A. Singh]

## 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: $X_i = X_k$

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


→

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

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

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

- $X_i$ is a random variable describing the word at position i in the document
- possible values for $X_i$ : any word $w_k$ in English

- Document = bag of words: the vector of counts for all $w_k$'s
- This vector of counts follows a ?? distribution

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

  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

$$\theta_{ijk} = \theta_{mjk} \text{ for } i \neq m$$

## MAP estimates for bag of words

Map estimate for multinomial

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

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

What $\beta$'s should we choose?

## Twenty NewsGroups

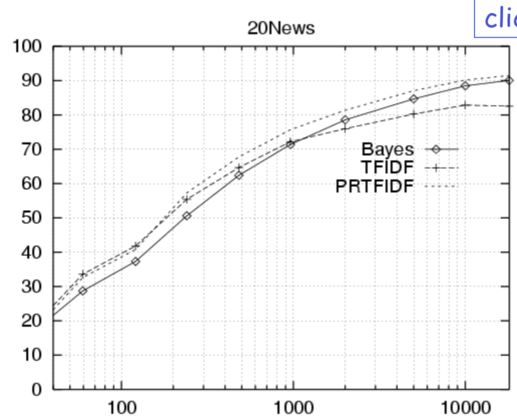
Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

| | |
|---|---|
| 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

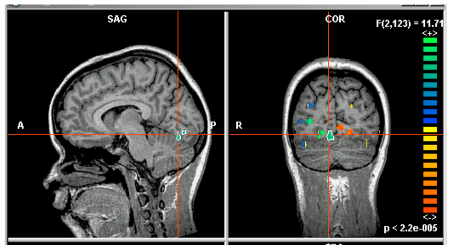
For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"




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

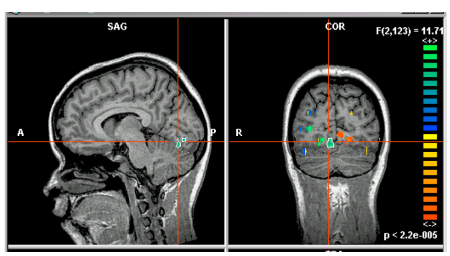
## What if we have continuous $X_i$ ?

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




## What if we have continuous $X_i$ ?

image classification: $X_i$ is i[th] pixel, Y = mental state




Still have:

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

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

## 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 $\sigma_{ik}$

- is independent of Y (i.e., $\sigma_i$),
- or independent of $X_i$ (i.e., $\sigma_k$)
- or both (i.e., $\sigma$)

## 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 $\mu_{ik}$ , variance $\sigma_{ik}$

- 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:

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

jth training example

ith feature

kth class

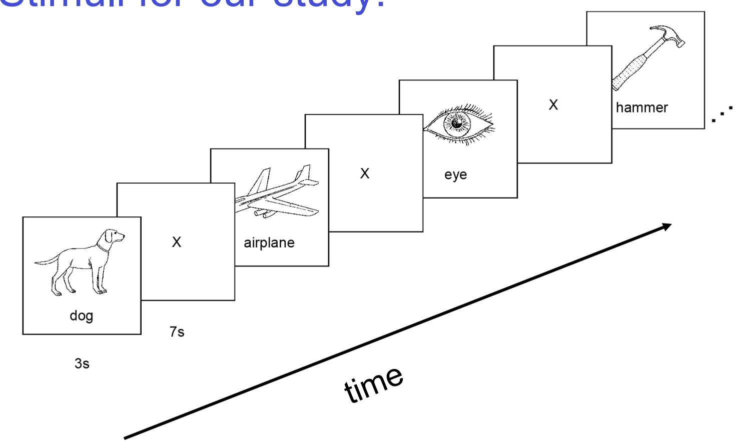
δ(z)=1 if z true, else 0

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

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

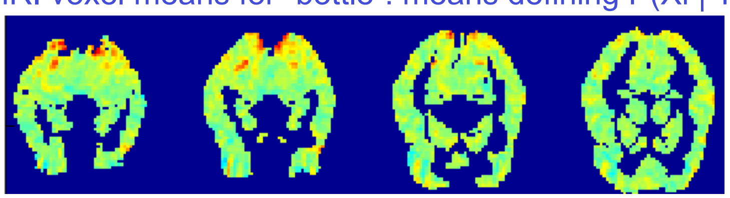
- are they reading a sentence or viewing a picture?
- reading the word “Hammer” or “Apartment”
- viewing a vertical or horizontal line?
- answering the question, or getting confused?

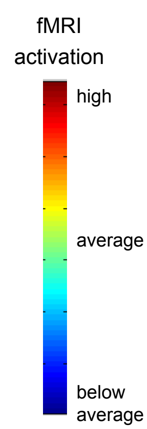
## Stimuli for our study:



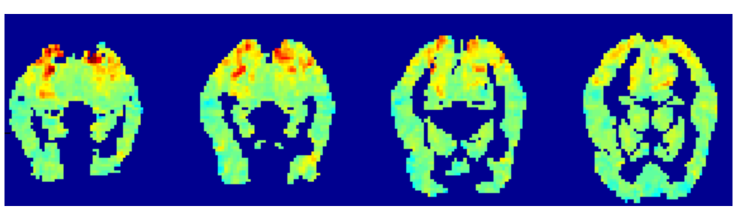

60 distinct exemplars, presented 6 times each

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

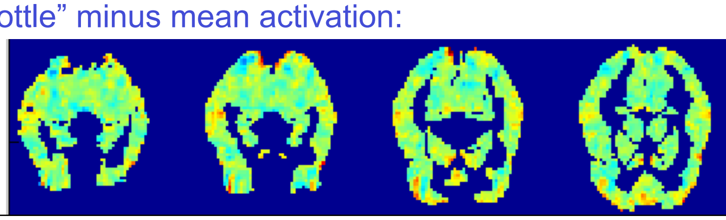




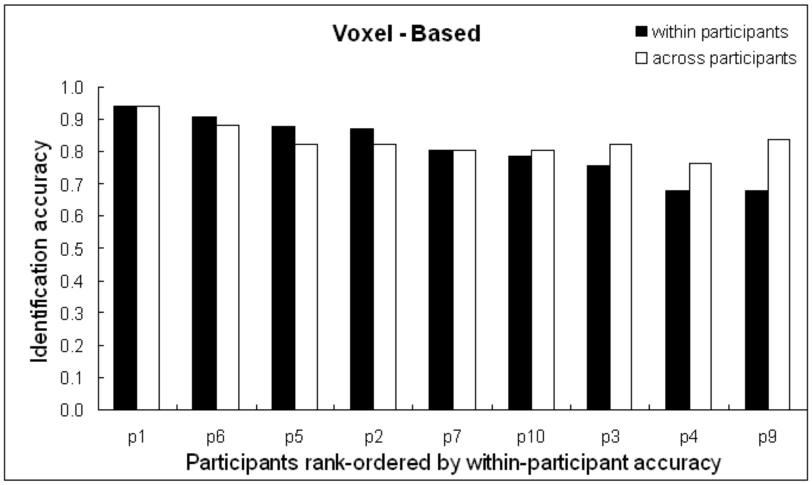

## Mean fMRI activation over all stimuli:



## “bottle” minus mean activation:

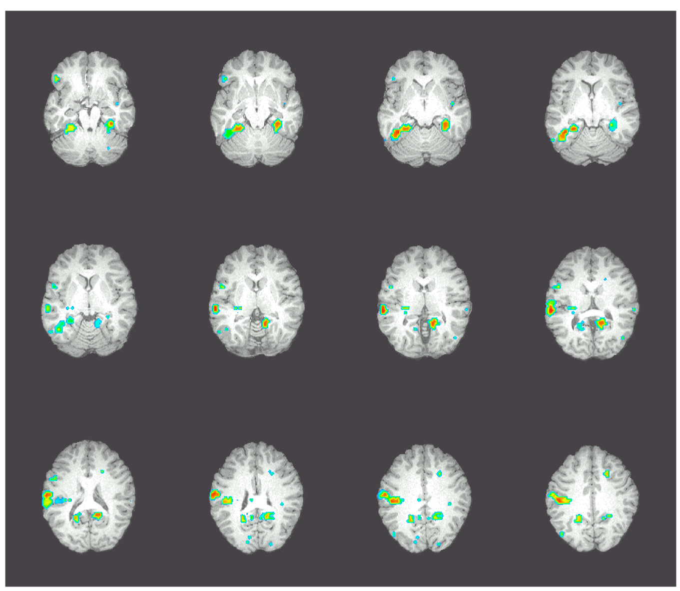


## Rank Accuracy Distinguishing among 60 words




## Tools vs Buildings: where does brain encode their word meanings?

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



## 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 error will the 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 extend Naïve Bayes if just 2 of the n $X_i$ are dependent?
- What does the decision surface of a Naïve Bayes classifier look like?

**Lejeune Dirichlet**


Johann Peter Gustav Lejeune Dirichlet

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