CSE546: Naïve Bayes
Winter 2012

Luke Zettlemoyer
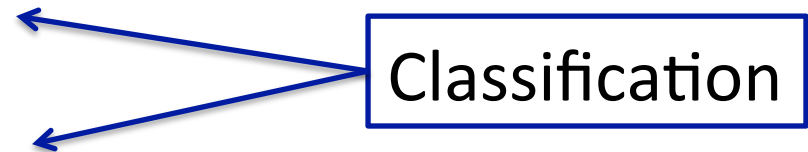
Slides adapted from Carlos Guestrin and Dan Klein

Supervised Learning: find f

- Given: Training set $\{(x_i, y_i) \mid i = 1 \ldots n\}$
- Find: A good approximation to $f : X \to Y$

Examples: what are X and Y ?

- Spam Detection
  - Map email to {Spam,Ham}
- Digit recognition
  - Map pixels to {0,1,2,3,4,5,6,7,8,9}
- Stock Prediction
  - Map new, historic prices, etc. to $\mathfrak{R}$ (the real numbers)

Classification

Example: Spam Filter

- **Input:** email
- **Output:** spam/ham
- **Setup:**
  - Get a large collection of example emails, each labeled "spam" or "ham"
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future emails

- **Features:** The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: SenderInContacts
  - …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES
  FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Example: Digit Recognition

- Input: images / pixel grids

- Output: a digit 0-9

- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future digit images

- Features: The attributes used to make the digit decision
  - Pixels: (6,8)=ON
  - Shape Patterns: NumComponents, AspectRatio, NumLoops
  - …

0

1

2

1

??

Other Classification Tasks

- In classification, we predict labels y (classes) for inputs x

- Examples:
    - Spam detection (input: document, classes: spam / ham)
    - OCR (input: images, classes: characters)
    - Medical diagnosis (input: symptoms, classes: diseases)
    - Automatic essay grader (input: document, classes: grades)
    - Fraud detection (input: account activity, classes: fraud / no fraud)
    - Customer service email routing
    - … many more

- Classification is an important commercial technology!

Lets take a probabilistic approach!!!

- Can we directly estimate the data distribution $P(X,Y)$?

- How do we represent these? How many parameters?

  - Prior, $P(Y)$:
    - Suppose Y is composed of k classes

  - Likelihood, $P(\mathbf{X}|Y)$:
    - Suppose $\mathbf{X}$ is composed of n binary features

- Complex model $\rightarrow$ High variance with limited data!!!

mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europe
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europe
bad	5	medium	medium	medium	medium	75to78	europe

Conditional Independence

- 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

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

- e.g.,

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

- Equivalent to:

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

Naïve Bayes

- ## Naïve Bayes assumption:
  - Features are independent given class:

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

  - More generally:

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

- ## How many parameters now?
  - Suppose **X** is composed of n binary features

The Naïve Bayes Classifier

- Given:
  - Prior P(Y)
  - *n* conditionally independent features **X** given the class Y
  - For each X_i, we have likelihood $P(X_i|Y)$
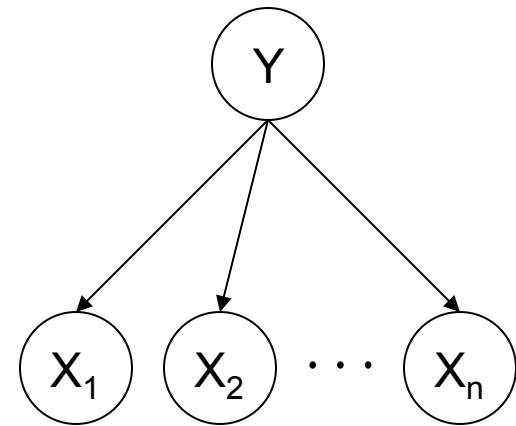
- Decision rule:

$$y^* = h_{NB}(\mathbf{x}) = \arg\max_y P(y)P(x_1, \ldots, x_n \mid y)$$
$$= \arg\max_y P(y) \prod_i P(x_i|y)$$

If certain assumption holds, NB is optimal classifier! Will discuss at end of lecture!

A Digit Recognizer

- Input: pixel grids
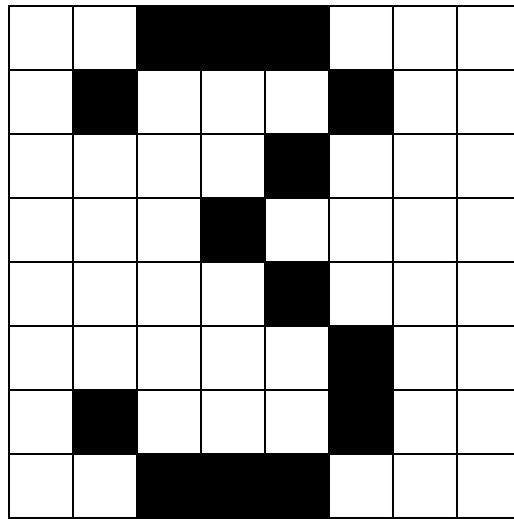
- Output: a digit 0-9

Naïve Bayes for Digits (Binary Inputs)

- Simple version:
  - One feature F_{ij} for each grid position <i,j>
  - Possible feature values are on / off, based on whether intensity is more or less than 0.5 in underlying image
  - Each input maps to a feature vector, e.g.

  $$\rightarrow \langle F_{0,0} = 0 \ \ F_{0,1} = 0 \ \ F_{0,2} = 1 \ \ F_{0,3} = 1 \ \ F_{0,4} = 0 \ \dots F_{15,15} = 0 \rangle$$

  - Here: lots of features, each is binary valued

- Naïve Bayes model:

$$P(Y|F_{0,0} \dots F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$$

- Are the features independent given class?
- What do we need to learn?

Example Distributions

$P(Y)$

1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1
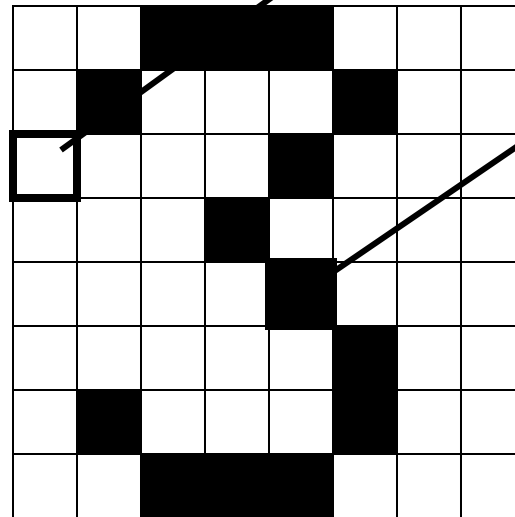
$P(F_{3,1} = on|Y)$

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

$P(F_{5,5} = on|Y)$

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

MLE for the parameters of NB

- Given dataset
  - Count(A=a,B=b) $\leftarrow$ number of examples where A=a and B=b

- MLE for discrete NB, simply:
  - Prior:

$$P(Y = y) = \frac{Count(Y = y)}{\sum_{y'} Count(Y = y')}$$

  - Likelihood:

$$P(X_i = x | Y = y) = \frac{Count(X_i = x, Y = y)}{\sum_{x'} Count(X_i = x', Y = y)}$$

Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:

$$P(X_1...X_n|Y) \neq \prod_i P(X_i|Y)$$

- Actual probabilities P(Y|**X**) often biased towards 0 or 1
- Nonetheless, NB is the single most used classifier out there
  - NB often performs well, even when assumption is violated
  - [Domingos & Pazzani '96] discuss some conditions for good performance

Subtleties of NB classifier 2: Overfitting

$P(\text{features}, C = 2)$

$P(C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.8$

$P(\text{on}|C = 2) = 0.1$

$P(\text{off}|C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.01$

$P(\text{features}, C = 3)$

$P(C = 3) = 0.1$

$P(\text{on}|C = 3) = 0.8$

$P(\text{on}|C = 3) = 0.9$

$P(\text{off}|C = 3) = 0.7$

$P(\text{on}|C = 3) = 0.0$
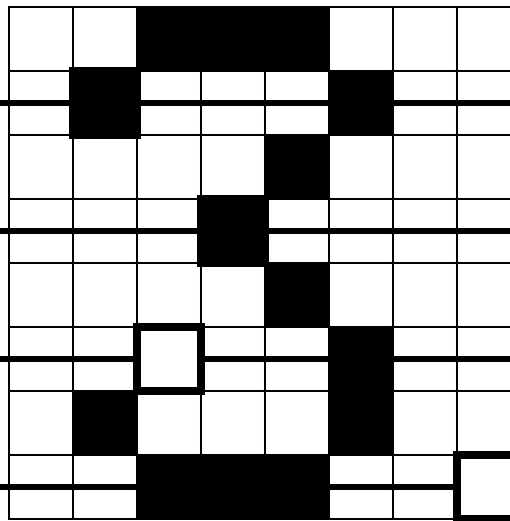
2 wins!!

For Binary Features: We already know the answer!

Beta(30,20)
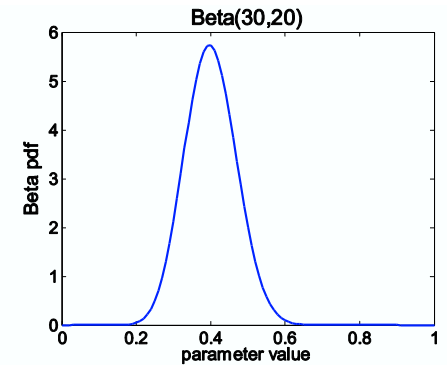
$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- **MAP:** use most likely parameter

$$\widehat{\theta} = \arg \max_\theta P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

- **Beta prior** equivalent to extra observations for each feature
- As $N \rightarrow \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

Multinomials: Laplace Smoothing

- Laplace's estimate:
  - Pretend you saw every outcome k extra times
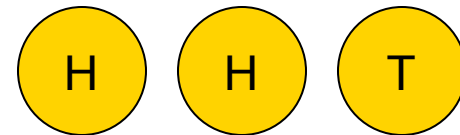
$$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$

  - What's Laplace with k = 0?
  - k is the strength of the prior
  - Can derive this as a MAP estimate for multinomial with *Dirichlet priors*

$$P_{LAP,0}(X) =$$

$$P_{LAP,1}(X) =$$

$$P_{LAP,100}(X) =$$

- Laplace for conditionals:
  - Smooth each condition independently:

$$P_{LAP,k}(x|y) = \frac{c(x, y) + k}{c(y) + k|X|}$$

Text classification

- Classify e-mails
  - Y = {Spam,NotSpam}
- Classify news articles
  - Y = {what is the topic of the article?}
- Classify webpages
  - Y = {Student, professor, project, …}

- What about the features **X**?
  - The text!

Features **X** are entire document — X_i for i^{th} word in article

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
```

NB for Text classification

- P(**X**|Y) is huge!!!
  - Article at least 1000 words, **X**={X_1,...,X_{1000}}
  - X_i represents i^{th} word in document, i.e., the domain of X_i is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.

- NB assumption helps a lot!!!
  - P(X_i=x_i|Y=y) is just the probability of observing word x_i in a document on topic y

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Bag of words model

- Typical additional assumption –

  - **Position in document doesn't matter**:
    - $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$ (all position have the same distribution)

  - "Bag of words" model – order of words on the page ignored

  - Sounds really silly, but often works very well!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

Bag of words model

- Typical additional assumption –

  - **Position in document doesn't matter:**

    - $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$ (all position have the same distribution)

  - "Bag of words" model – order of words on the page ignored

  - Sounds really silly, but often works very well!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

in is lecture lecture next over person remember room sitting the the the to to up wake when you

Bag of Words Approach

NB with Bag of Words for text classification

- ## Learning phase:
  - Prior P(Y)
    - Count how many documents from each topic (prior)
  - $P(X_i|Y)$
    - For each topic, count how many times you saw word in documents of this topic (+ prior); remember this dist'n is shared across all positions i

- ## Test phase:
  - For each document
    - Use naïve Bayes decision rule

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Twenty News Groups results

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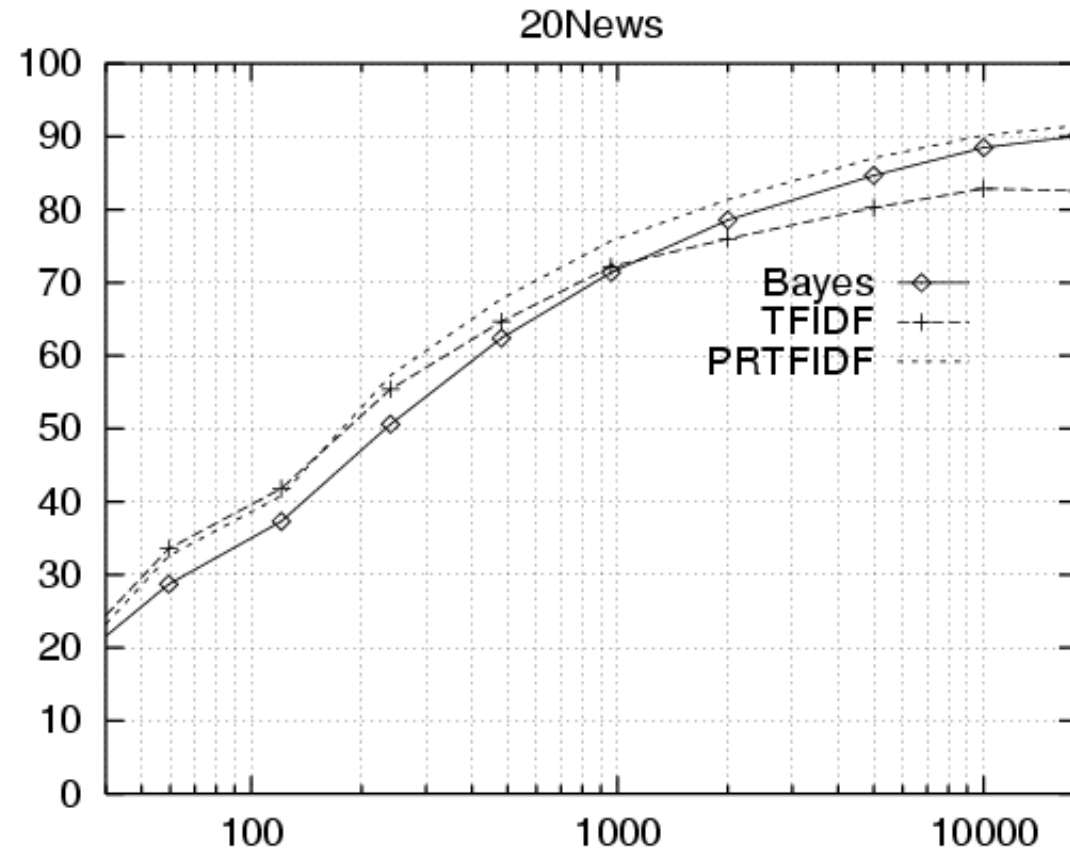
Accuracy vs. Training set size (1/3 withheld for test)

What if we have continuous X_i ?

Eg., character recognition: X_i is ith pixel
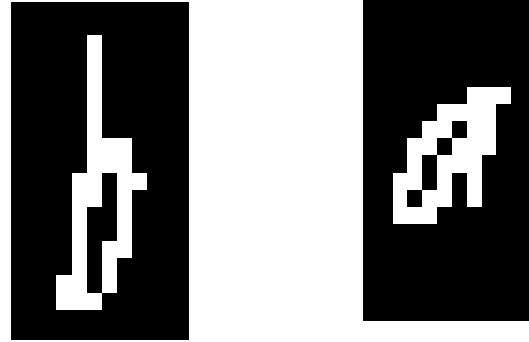
Gaussian Naïve Bayes (GNB):

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

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)

Estimating Parameters: *Y* discrete, *X*_i continuous

Maximum likelihood estimates:

- Mean:

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

jth training example

$\delta(x)$=1 if x true, else 0

- Variance:

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

Example: GNB for classifying mental states

~1 mm resolution

~2 images per sec.
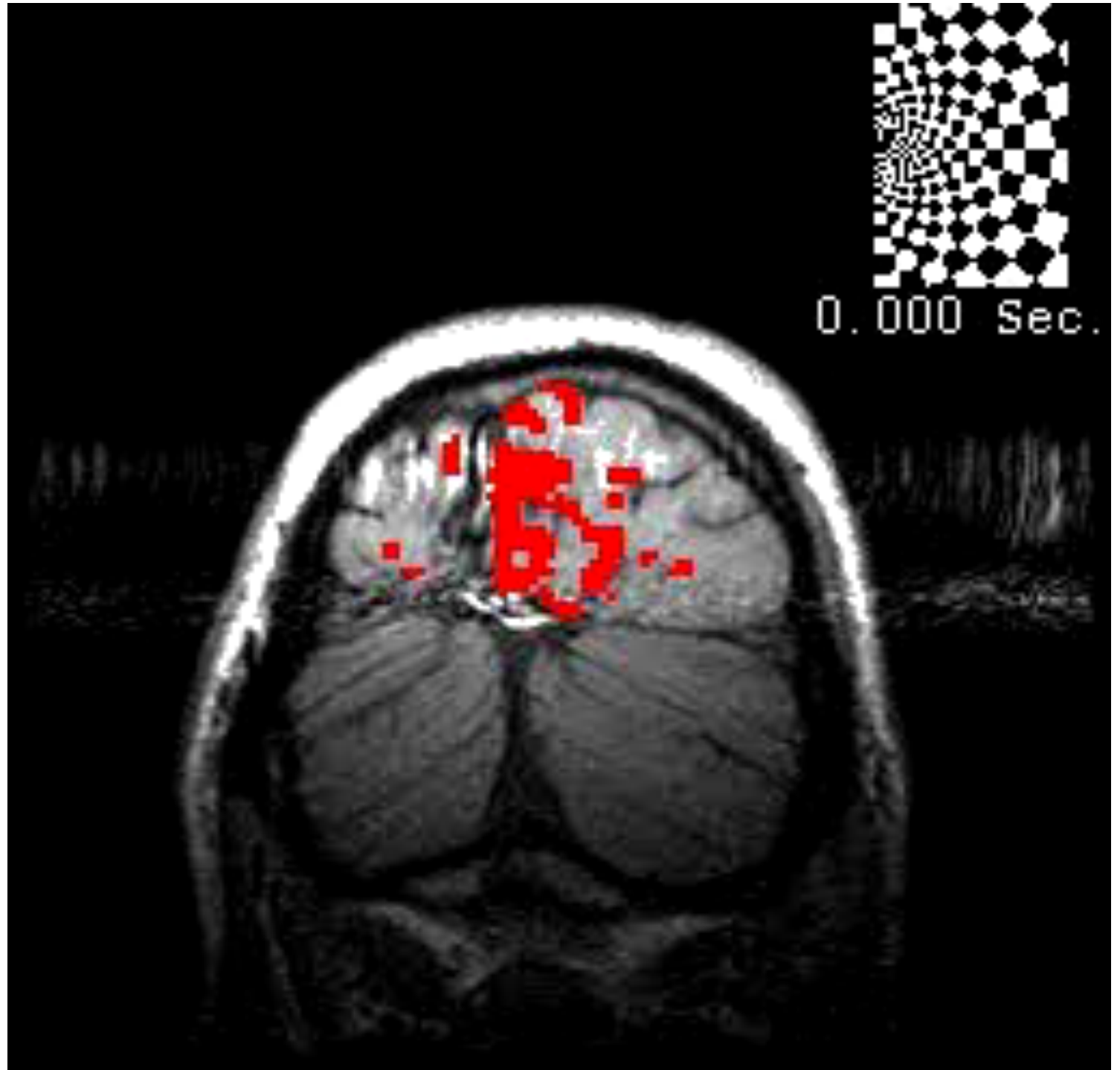
15,000 voxels/image

non-invasive, safe

measures Blood
Oxygen Level
Dependent (BOLD)
response
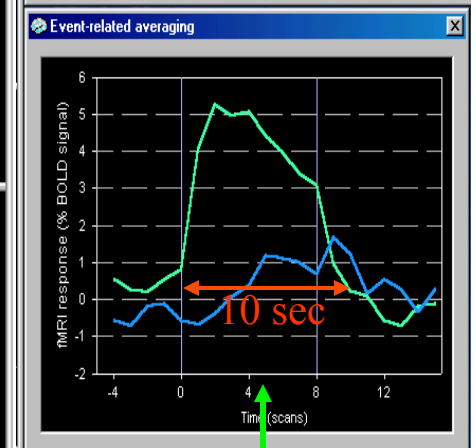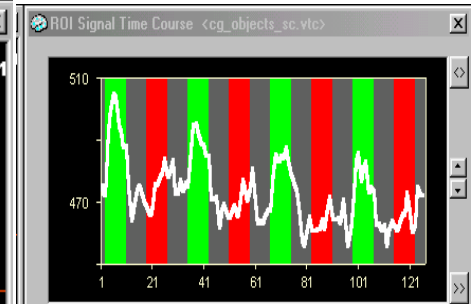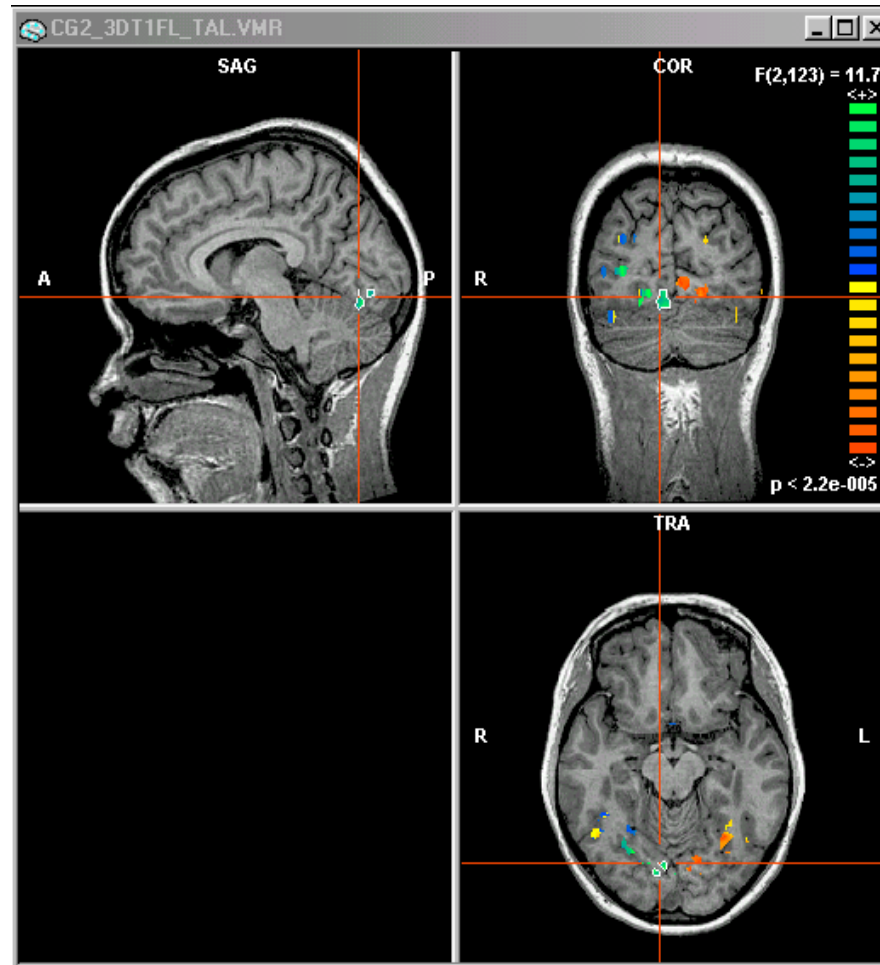
Typical
impulse
response

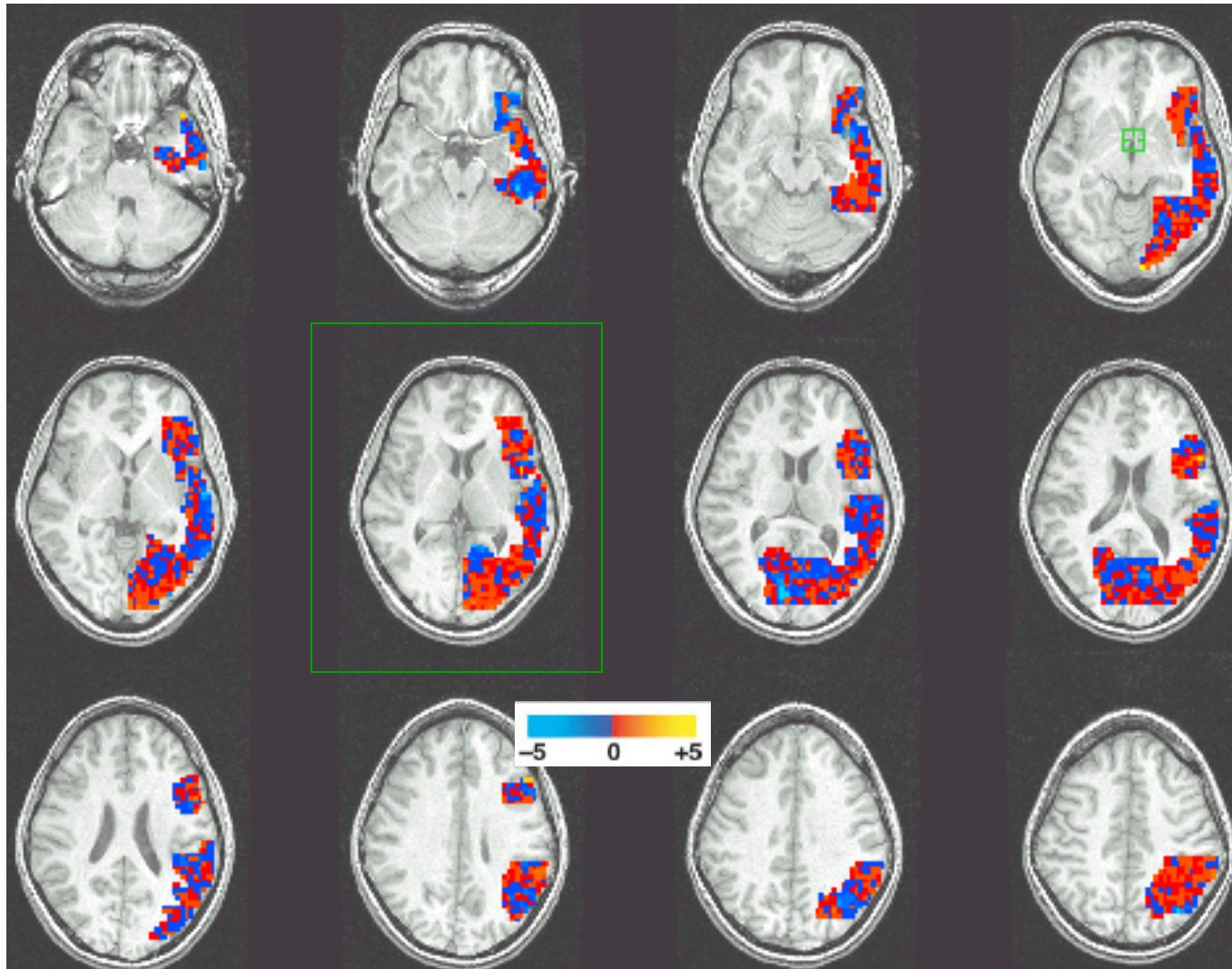Brain scans can track activation with precision and sensitivity

0.000 Sec.

[Mitchell et al.]

Gaussian Naïve Bayes: Learned $\mu_{voxel,word}$
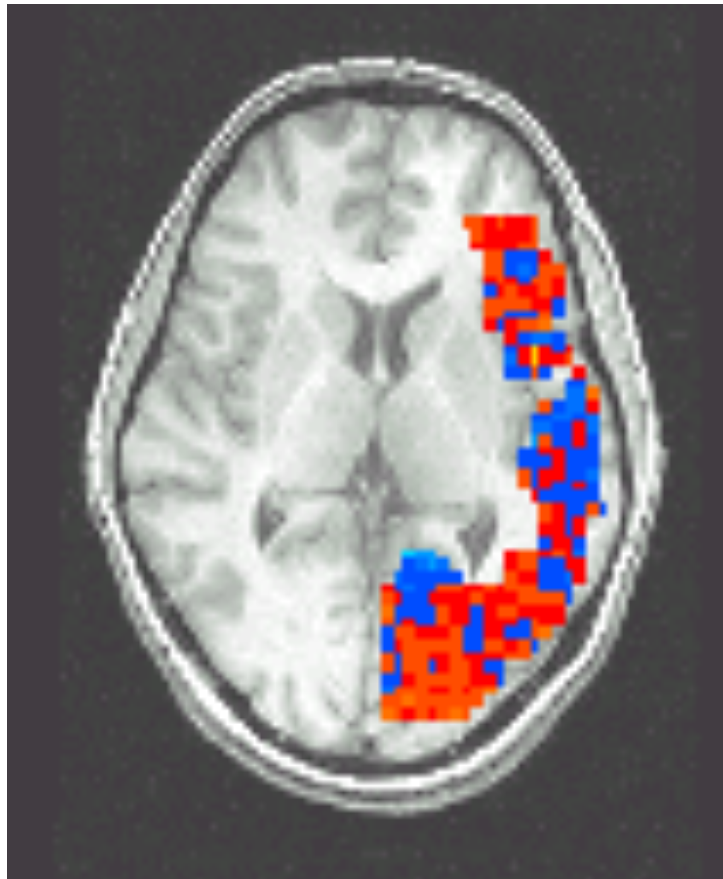## P(BrainActivity | WordCategory = {People,Animal})

[Mitchell et al.]
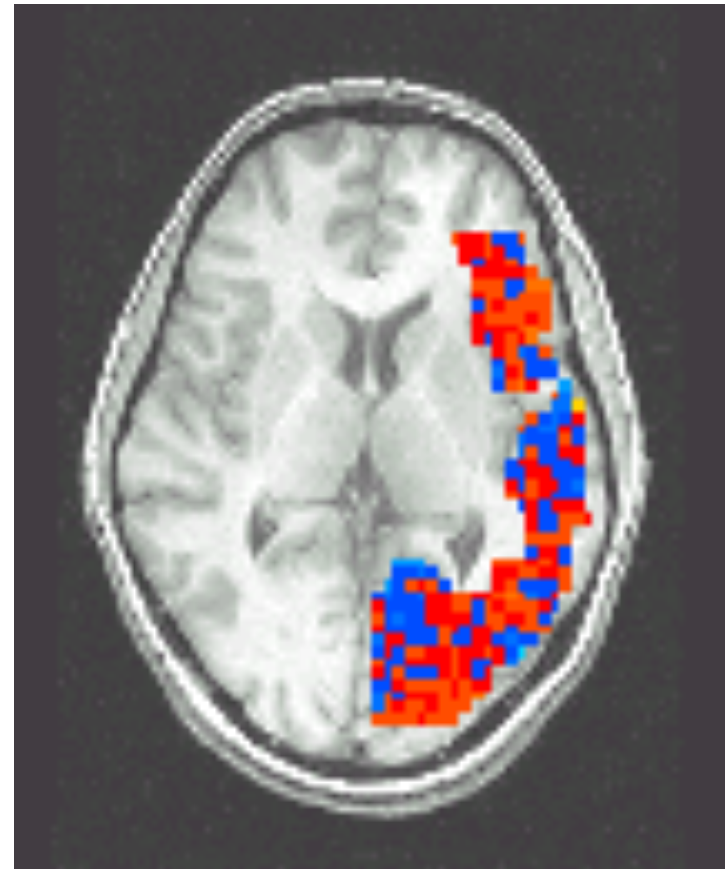
Learned Bayes Models – Means for P(BrainActivity | WordCategory)

Pairwise classification accuracy: 85%

People words

Animal words

Bayes Classifier is Optimal!

- **Learn**: h : $\mathbf{X} \mapsto Y$
  - $\mathbf{X}$ – features
  - Y – target classes

- Suppose: you know true P(Y|$\mathbf{X}$):
  - Bayes classifier:

$$h_{Bayes}(x) = \arg\max_y P(Y = y | \mathbf{X} = x)$$

- **Why?**

Optimal classification

- Theorem:
  - Bayes classifier h_{Bayes} is optimal!

$$error_{true}(h_{Bayes}) \leq error_{true}(h), \quad \forall h$$

  - Why?

$$p_h(error) = \int_x p_h(error|x)p(x)$$

$$= \int_x \int_y \delta(h(x),y)p(y|x)p(x)$$

What you need to know about Naïve Bayes

- Naïve Bayes classifier
  - What's the assumption
  - Why we use it
  - How do we learn it
  - Why is Bayesian estimation important
- Text classification
  - Bag of words model
- Gaussian NB
  - Features are still conditionally independent
  - Each feature has a Gaussian distribution given class
- Optimal decision using Bayes Classifier