Machine Learning

CSE 6363 (Fall 2019)

Lecture 8 Probability Distribution, Naïve Bayes

Dajiang Zhu, Ph.D.

Department of Computer Science and Engineering

Slides of this course (CSE6363) courtesy: Dr. Heng Huang, Dr. Aarti Singh

MLE for Gaussian

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Probility of i.i.d. samples $x_1, x_2, \ldots, x_N$:

$$P(D|\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

How to find μ and σ?

Log-likelihood of data:

$$\ln P(D|\mu,\sigma) = \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right]$$

Ref: Kevin Murphy

Unsupervised Learning

- **Data:** $D = \{d_1, d_2, .., d_n\}$
  $d_i = \mathbf{x}_i$ vector of values
  No target value (output) y

- **Objective:**
  – learn relations between samples, components of samples

Types of problems:
- **Clustering**
  Group together "similar" examples, e.g. patient cases
- **Density estimation**
  – Model probabilistically the population of samples

Ref: Milos Hauskrecht

Unsupervised Learning

Learning distributions/densities – Unsupervised learning

- Task: Learn $P(X;\theta) \equiv \text{Learn } \theta$ (know form of P, except θ)
- Experience: D = $\{X_i\}_{i=1}^{n} \sim P(X;\theta)$
- Performance: $\max_\theta P(D|\theta)$

$$= \min_\theta -\log P(D|\theta)$$

$$= \min_\theta \frac{1}{n}\sum_{i=1}^{n} \underbrace{-\log P(X_i|\theta)}_{\text{loss}(X_i,\theta)}$$ **Negative log Likelihood loss!**

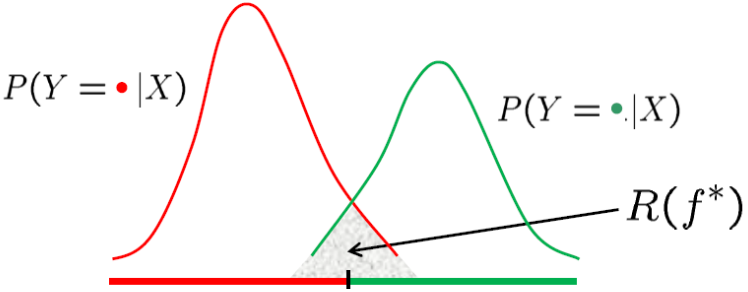
Ref: Milos Hauskrecht

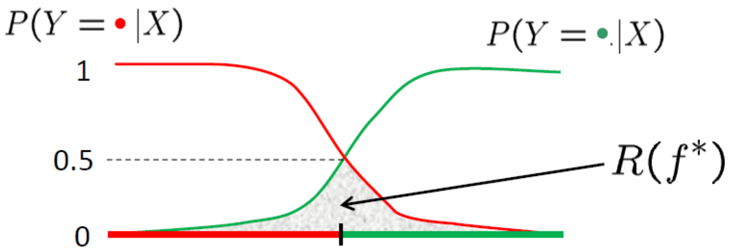
Optimal Classification

Optimal predictor (Bayes classifier):

$$f^* = \arg\min_f P(f(X) \neq Y)$$

$$f^*(x) = \arg\max_{Y=y} P(Y = y | X = x)$$





- Even the optimal classifier makes mistakes – $R(f^*)>0$
- Optimal classifier depends on **unknown** distribution - P_{xy}

Ref: Milos Hauskrecht

Optimal Classifier

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

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

Optimal classifier :

$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

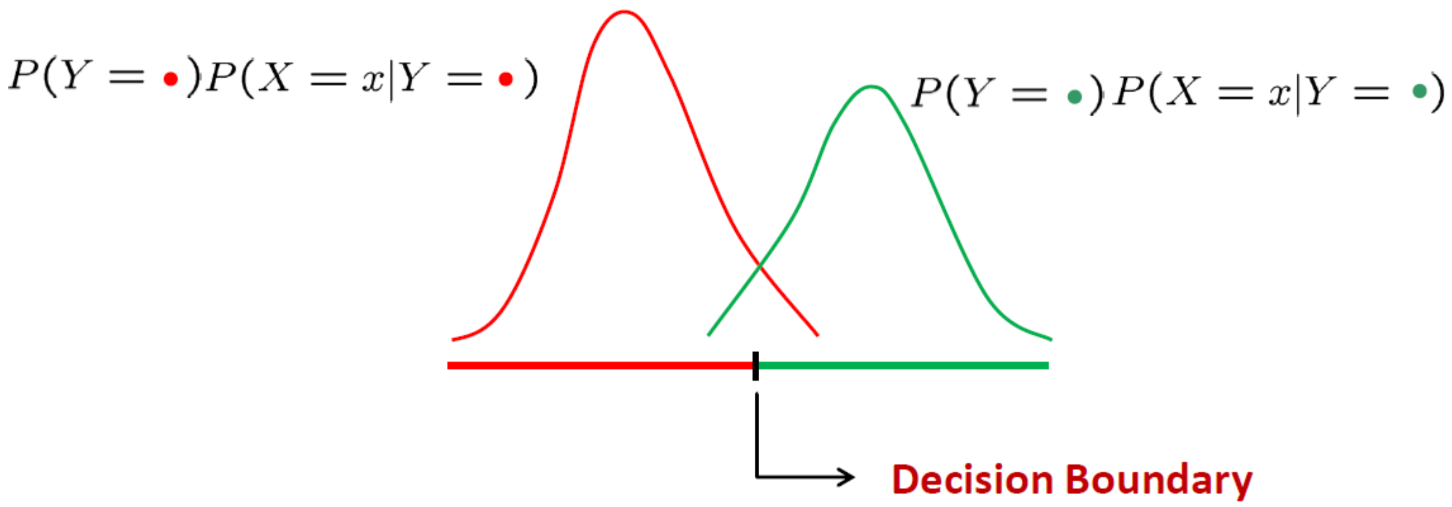
$$= \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\boxed{\text{Class conditional density}}}\underbrace{P(Y = y)}_{\boxed{\text{Class prior}}}$$

Ref: Milos Hauskrecht

Example Decision Boundaries

Gaussian class conditional densities (1-dimension/feature)

$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$



Ref: Milos Hauskrecht

Learning the optimal classifier

Optimal classifier :

$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

$$= \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional density}} \underbrace{P(Y = y)}_{\text{Class prior}}$$

Need to know Prior P(Y = y) for all y
Likelihood P(X=x|Y= y) for all x,y

Ref: Milos Hauskrecht

How to Learn the Classifier?

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d) \quad Y$

n rows

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Lets learn P(Y|X)

- How do we represent these? How many parameters?
  - Prior, P(Y): **K-1**
    - Suppose Y is composed of *k* classes
  - Likelihood, P(**X**|Y): $(2^d - 1)K$
    - Suppose **X** is composed of *n* binary features

Ref: Carlos Guestrin

How to Learn the Classifier?

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d)$ $\quad Y$

n rows

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Lets learn P(Y|X) –how many parameters?

$2^dK – 1$ (K classes, d binary features)

Need n >> 2^dK –1 number of training data to learn all parameters

Ref: Carlos Guestrin

Marginal Independence

When your knowledge of Y's value doesn't affect your belief in the value of X…

Random variable X is **marginal independent** of random variable Y if

$$P(X = x_i \mid Y = y_k) = P(X = x_i)$$

X and Y are **independent** iff:

$$P(X \mid Y) = P(X) \text{ or } P(Y \mid X) = P(Y) \text{ or } P(X, Y) = P(X)\, P(Y)$$

That is new evidence Y(or X) does not affect current belief in X (or Y)

Ref: Milos Hauskrecht

Conditional Independence

X is **conditionally independent** of Y given Z:

- probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

- Equivalent to:

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

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

Note: does NOT mean Thunder is independent of Rain

Ref: Milos Hauskrecht

Conditional vs. Marginal Independence

- C calls A and B separately and tells them a number n ϵ {1,…,10}
- Due to noise in the phone, A and B each imperfectly (and independently) draw a conclusion about what the number was.
- A thinks the number was n_a and B thinks it was n_b.
- Are n_a and n_b marginally independent?
- Are n_a and n_b conditionally independent given n?

Ref: Milos Hauskrecht

Prediction using Conditional Independence

- When predicting lightening, we use two conditionally independent features
  - ✓ Thunder
  - ✓ Rain

\# parameters needed to learn likelihood given L

$P(T,R|L)$ $\quad$ **$(2^2-1)2 = 6$**

With conditional independence assumption

$P(T,R|L) = P(T|L)\,P(R|L)$ $\quad$ **$(2-1)2 + (2-1)2 = 4$**

Ref: Milos Hauskrecht

The Naïve Bayes Assumption

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

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

  - More generally:

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

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

Ref: Carlos Guestrin

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:

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

If conditional independence assumption holds, NB is optimal classifier! But worse otherwise

Ref: Carlos Guestrin

The Naïve Bayes Algorithm

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- Maximum Likelihood Estimates
  - ✓ For class Prior
$$\widehat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$
  - ✓ For Likelihood $\frac{\widehat{P}(x_i, y)}{\widehat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$
- NB Prediction for test data $X = (x_1, \ldots, x_d)$
$$Y = \arg\max_y \widehat{P}(y) \prod_{i=1}^{d} \frac{\widehat{P}(x_i, y)}{\widehat{P}(y)}$$

Ref: Carlos Guestrin

Violation of NB Assumption

- Usually, features are not conditionally independent:

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

- 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

Ref: Carlos Guestrin

Insufficient training data

- What if you never see a training instance where X_1=a when Y=b?
  - e.g., Y={SpamEmail}, X_1={'Earn'}
  - $P(X_1=a \mid Y=b) = 0$

- Thus, no matter what the values $X_2,\ldots,X_d$ take:
  - $P(Y=b \mid X_1=a, X_2,\ldots,X_d) = 0$

What shall we do?

Ref: Carlos Guestrin

The Naïve Bayes Algorithm

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

- MAP - add m "virtual" examples

  ✓ MAP Estimate

$$\widehat{P}(X_i = a | Y = b) = \frac{\{\#j : X_i^{(j)} = a, Y^{(j)} = b\} \boxed{+ mQ(X_i = a, Y = b)}}{\{\#j : Y^{(j)} = b\} \boxed{+ mQ(Y = b)}}$$

 # virtual examples with Y = b

- Now, even if you never observe a class/feature posterior probability never zero.

Ref: Carlos Guestrin

Example

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes

outlook	temp.	humidity	windy	play
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

outlook	temp.	humidity	windy	play
sunny	cool	high	true	?

Ref: Carlos Guestrin

Example

Frequencies and probabilities for the weather data:

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Ref: Carlos Guestrin

Example

Now assume that we have to classify the following new instance:

outlook	temp.	humidity	windy	play
sunny	cool	high	true	?

Key idea: compute a probability for each class based on the probability distribution in the training data.

First take into account the the probability of each attribute. Treat all attributes **equally important**, i.e., multiply the probabilities:

$$P(\text{yes}) = 2/9 \cdot 3/9 \cdot 3/9 \cdot 3/9 = 0.0082$$
$$P(\text{no}) = 3/5 \cdot 1/5 \cdot 4/5 \cdot 3/5 = 0.0577$$

Ref: Carlos Guestrin

Example

Now take into account the **overall probability** of a given class. Multiply it with the probabilities of the attributes:

$$P(\text{yes}) = 0.0082 \cdot 9/14 = 0.0053$$
$$P(\text{no}) = 0.0577 \cdot 5/14 = 0.0206$$

Now choose the class so that it **maximizes** this probability. This means that the new instance will be classified as no.

Ref: Carlos Guestrin

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!

Ref: Carlos Guestrin

Features X Are Entire Document

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 decidec
```


Ref: Carlos Guestrin

NB for Text Classification

- $P(\mathbf{X}|Y)$ is huge!!!
  - Article at least 1000 words, $\mathbf{X}=\{X_1,\ldots,X_{1000}\}$
  - 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

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

Ref: Carlos Guestrin

Bag of Words Model

- Typical additional assumption – **Position in document doesn't matter**: $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

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

Ref: Carlos Guestrin

Bag of Words Model

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

Ref: Carlos Guestrin

NB with Bag of Words for Text Classification

- Learning phase:
  - Prior P(Y)
    - Count how many documents you have from each topic (+ prior)
  - $P(X_i|Y)$
    - For each topic, count how many times you saw word in documents of this topic (+ prior)
- Test phase:
  - For each document
    - Use naïve Bayes decision rule

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

Ref: Carlos Guestrin

Twenty News Groups Results

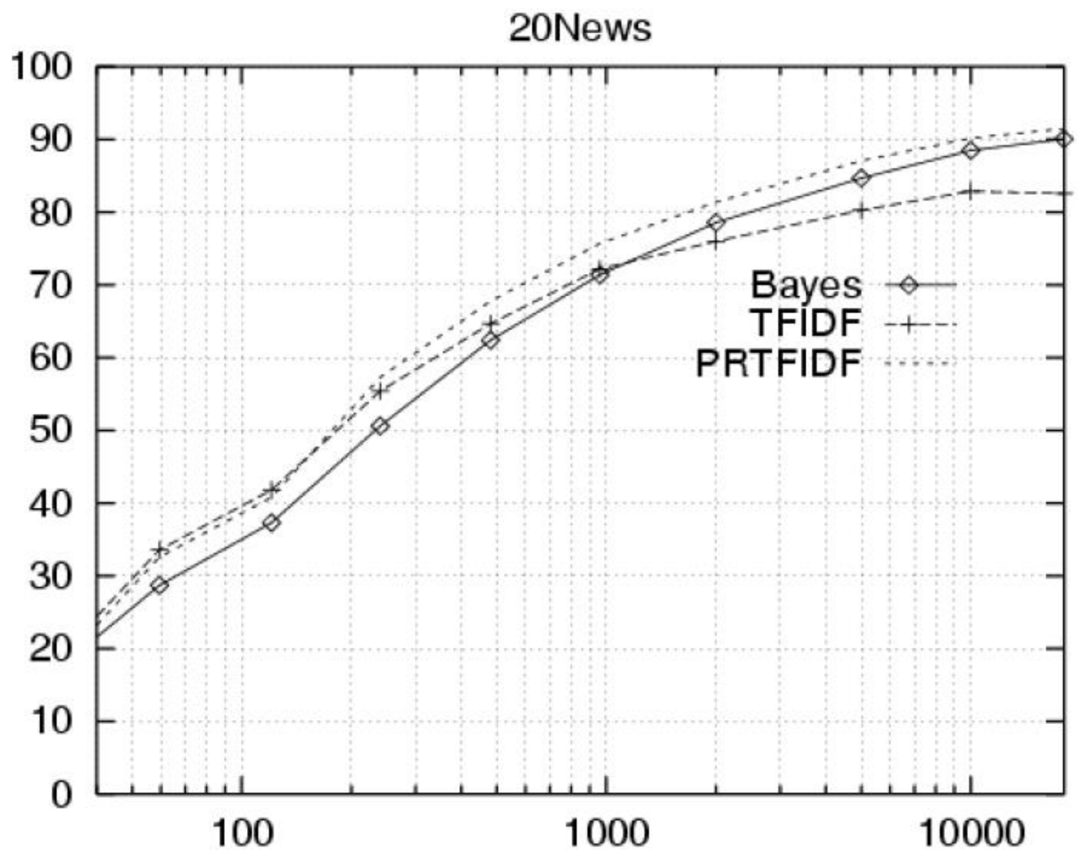
Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

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

Ref: Carlos Guestrin

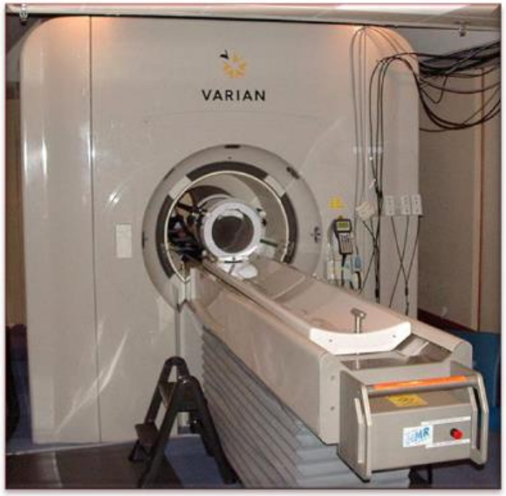
Twenty News Groups Results

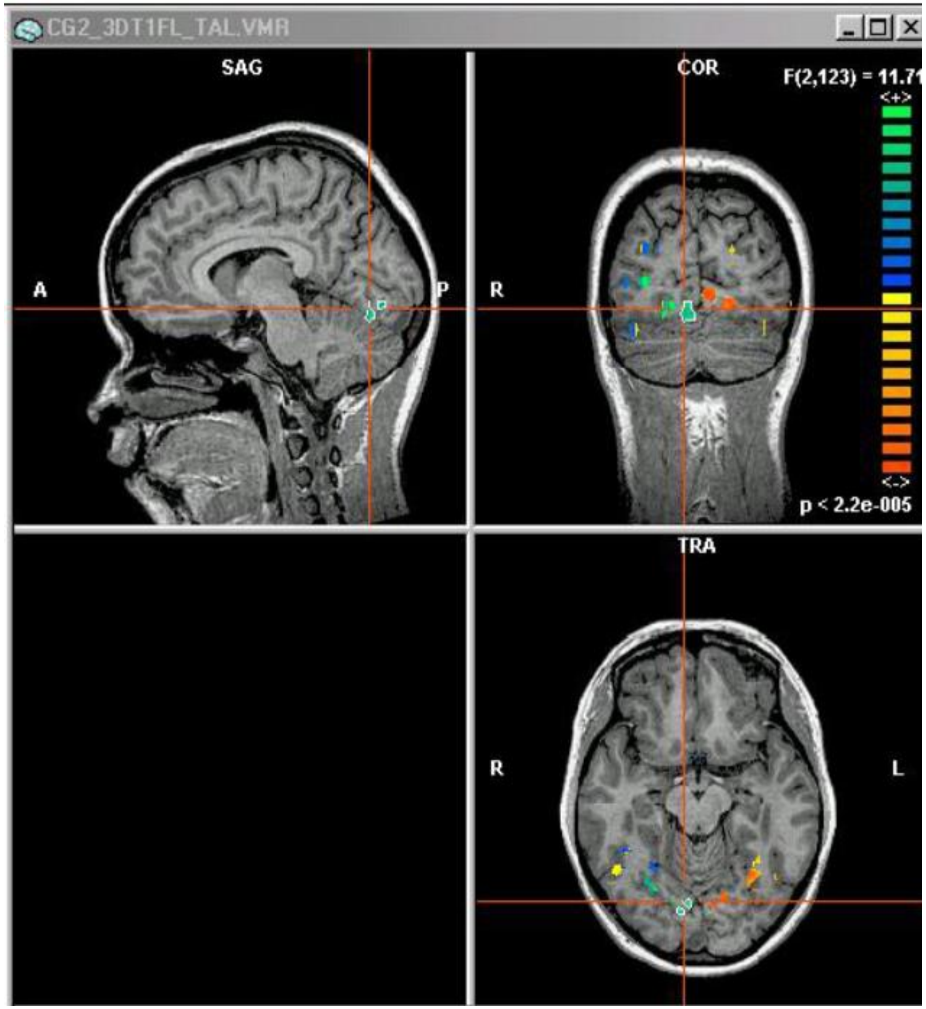


Accuracy vs. Training set size (1/3 withheld for test)

Ref: Carlos Guestrin

GNB Example





~1 mm resolution

~2 images per sec.

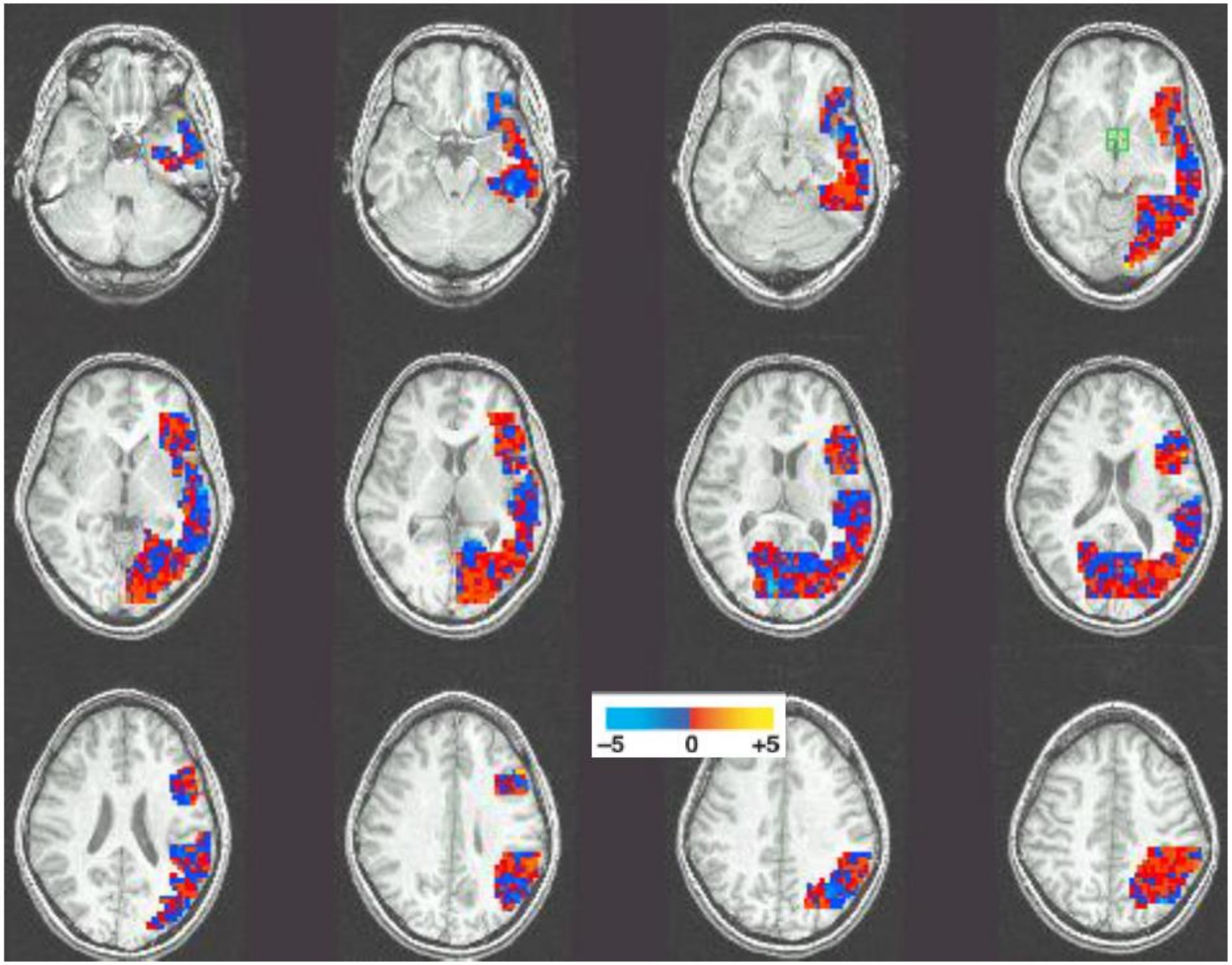
15,000 voxels/image

non-invasive, safe

measures Blood Oxygen Level Dependent (BOLD) response

[Mitchell et al.]

GNB Example



[Mitchell et al.]

15,000 voxels or features

10 training examples or subjects per class

GNB Example

Pairwise classification accuracy: 85%

[Mitchell et al.]