MAP and Naïve Bayes Classifier

Dr. Jianlin Cheng

University of Missouri, Columbia

Slides Adapted from Book and CMU, Stanford Machine Learning Courses

Fall, 2011

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the coin parameter θ, within ε = 0.1, with probability at least 1-δ = 0.95.
  How many flips?

$$P(|\,\hat{\theta} - \theta^*\,| \geq \epsilon) \;\leq\; 2e^{-2n\epsilon^2}$$

Sample complexity

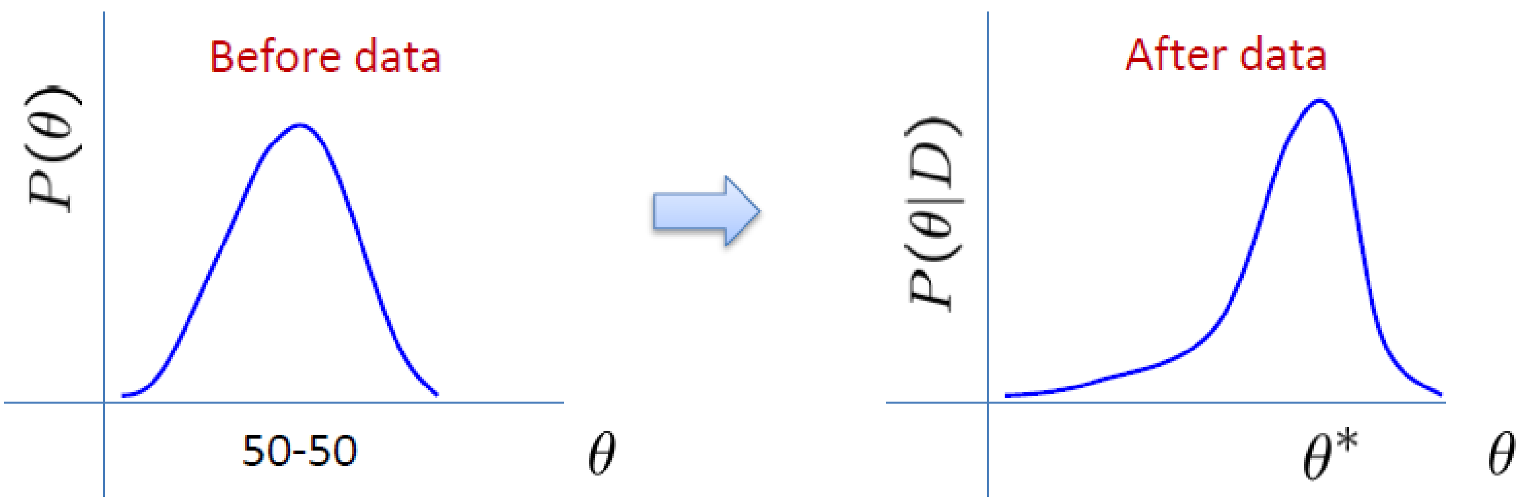
$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

Homework assignment 1: derive the bound for n given ϵ, δ

What about prior knowledge?

- Billionaire says: Wait, I know that the coin is "close" to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**

- Rather than estimating a single θ, we obtain a distribution over possible values of θ



Bayesian Learning

$$P(D,\theta) = P(D|\theta) * P(\theta)$$

- Use Bayes rule:

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

- Or equivalently:

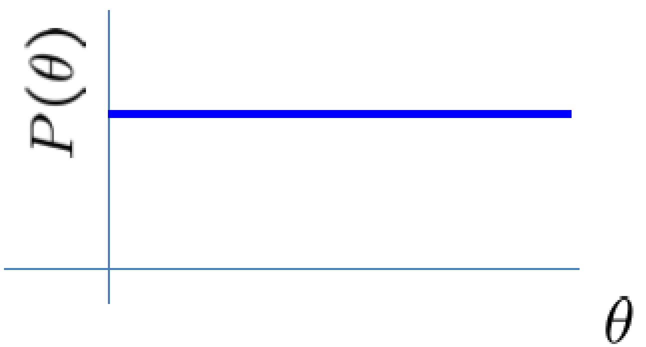
$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{P(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

Prior distribution

- What about prior?
  - Represents expert knowledge (philosophical approach)
  - Simple posterior form (engineer's approach)

- Uninformative priors:
  - Uniform distribution



- Conjugate priors:
  - Closed-form representation of posterior
  - $P(\theta)$ and $P(\theta|D)$ have the same form

Conjugate Prior

- $P(\theta)$ 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,

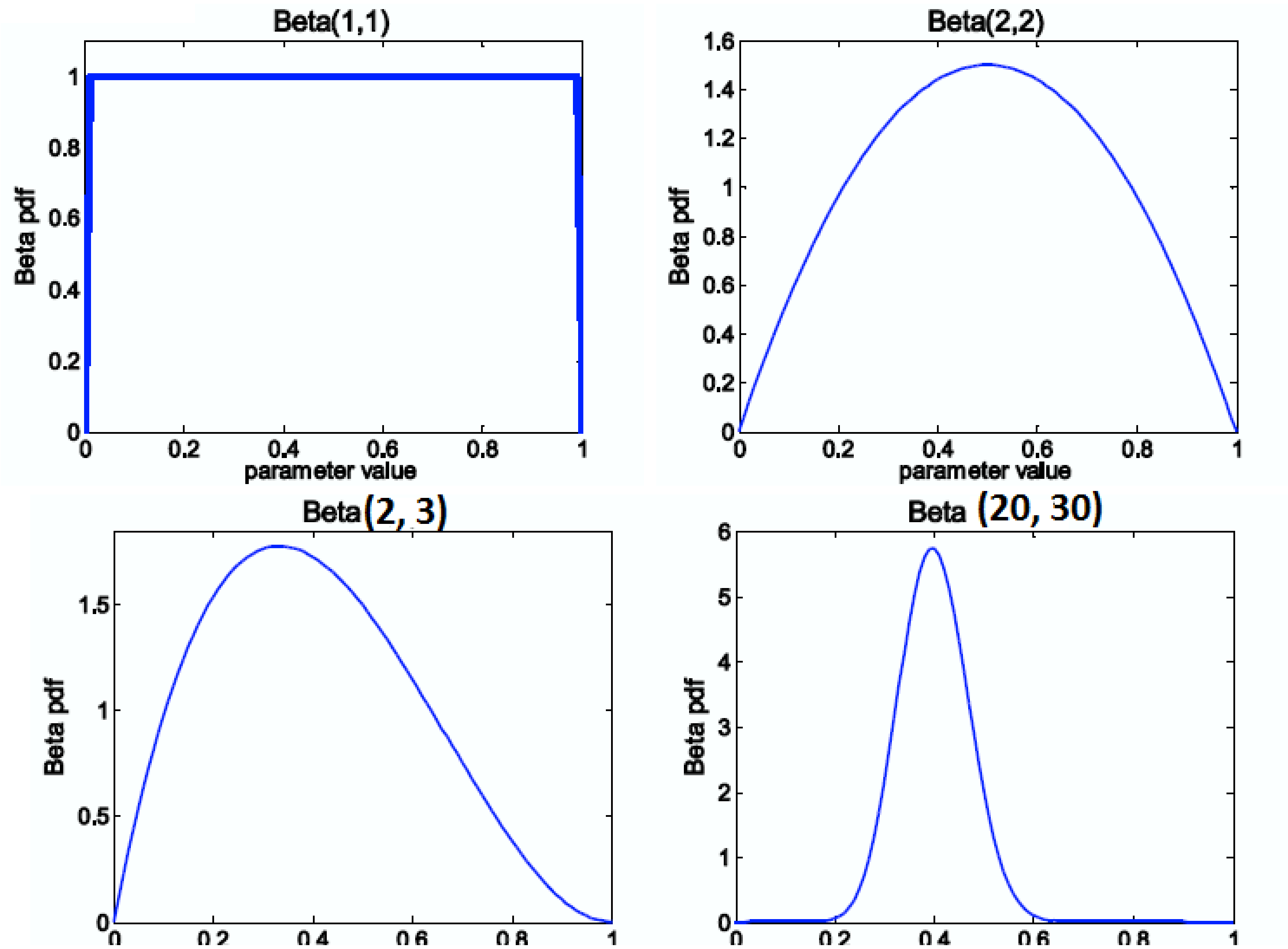
$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Then posterior is Beta distribution

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

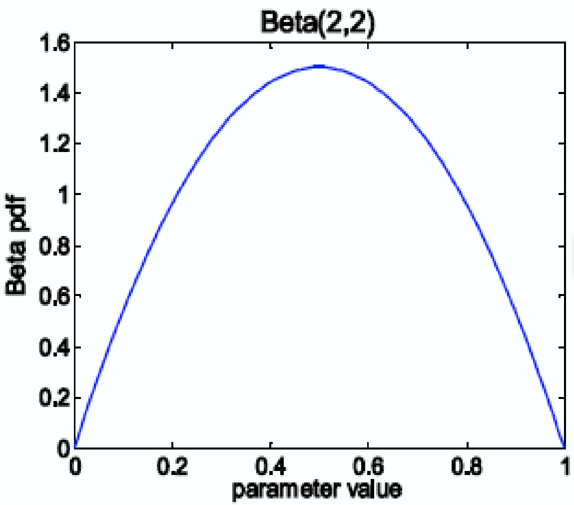
Beta distribution

$Beta(\beta_H, \beta_T)$ More concentrated as values of β_H, β_T increase

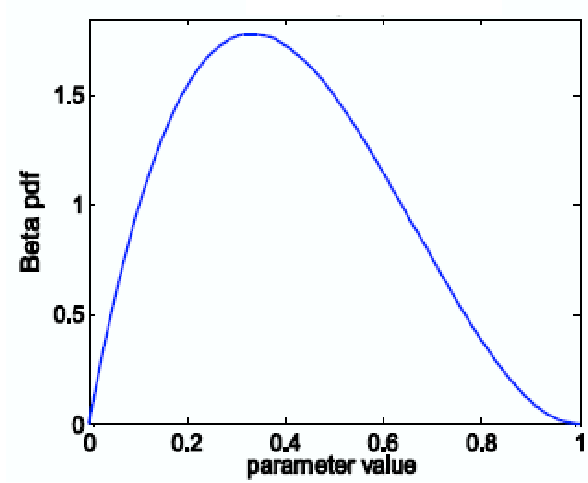


Beta conjugate prior

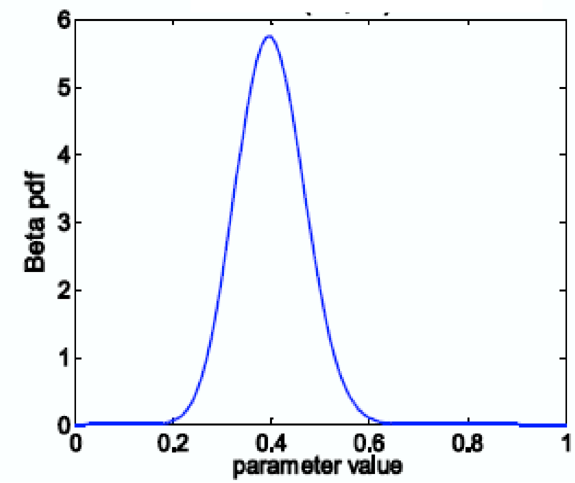
$$P(\theta) \sim Beta(\beta_H, \beta_T) \qquad P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Beta(2, 3)



Beta(20, 30)



As $n = \alpha_H + \alpha_T$ increases

As we get more samples, effect of prior is “washed out”

Conjugate Prior

- $P(\theta)$ 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.

Maximum A Posteriori Estimation

Choose θ that maximizes a posterior probability

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} \quad P(\theta \mid D) \\ &= \arg\max_{\theta} \quad P(D \mid \theta) P(\theta)\end{aligned}$$

MAP estimate of probability of head:

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

$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

When is MAP same as MLE?

MAP using Conjugate Prior

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

Coin flip problem

Likelihood is ~ Binomial

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

If prior is Beta distribution,

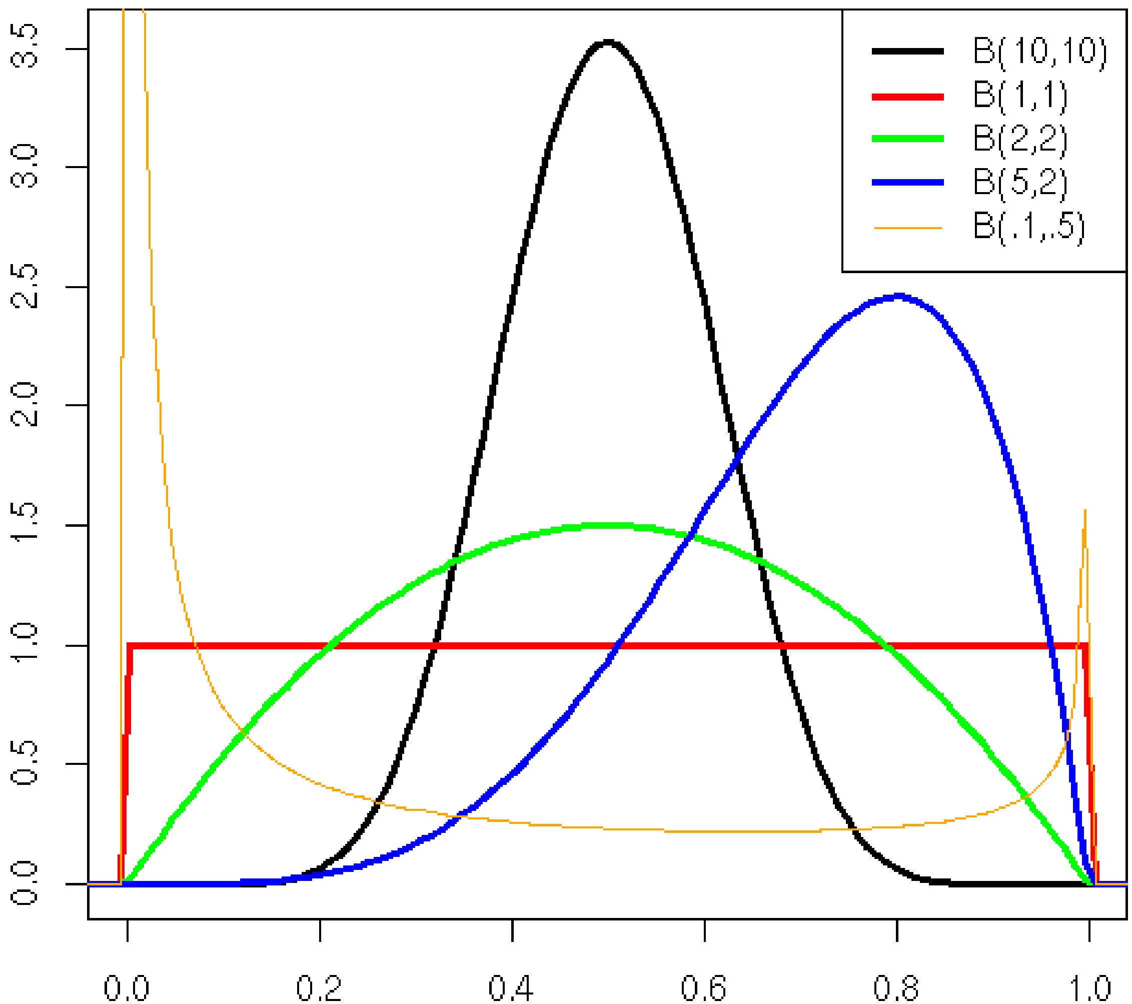
$$P(\theta) \propto \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1} \sim Beta(\beta_H, \beta_T)$$

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 few beta probability distributions



MLE vs. MAP

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

What if we toss the coin too few times?

- You say: Probability next toss is a head = 0
- Billionaire says: You're fired! ...with prob 1 ☺

$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

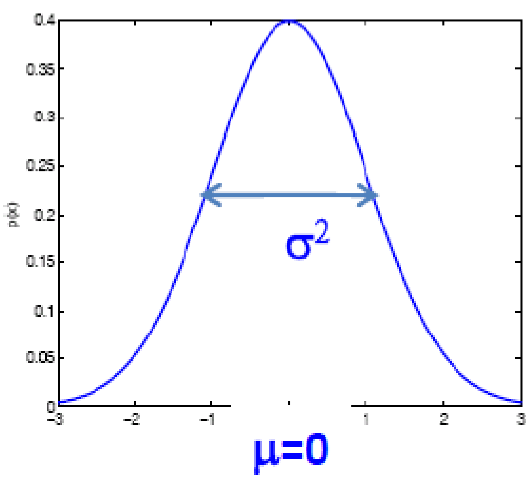
- Beta prior equivalent to extra coin flips **(regularization)**
- As $n \to \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

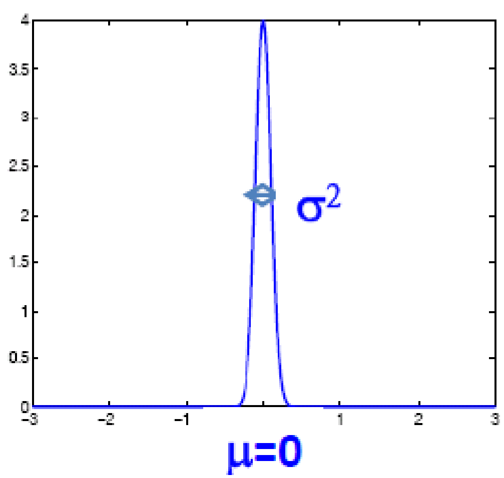
Bayesians vs. Frequentists

What about continuous variables?

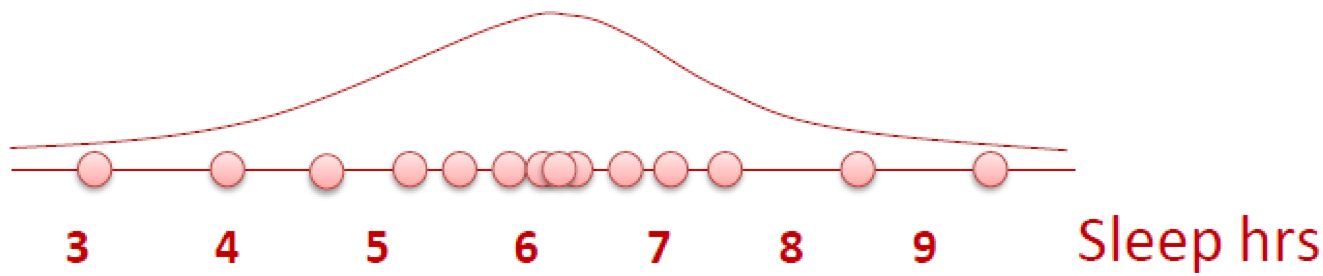
- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**

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





Gaussian distribution

Data, D = 

- Parameters: μ – mean, σ^2 - variance
- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

MLE Estimate of Gaussian

- Find u, σ^2 that maximize $P(D|u, \sigma^2)$

MLE Estimate of Gaussian

- Find u, σ^2 that maximize P(D|u, σ^2)

$$P(D|u,\sigma^2) = \prod_{i=1}^{n} P(x_i|u,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-u)^2}{2\sigma^2}}$$

$$\log P(D|u,\sigma^2) = \sum_{i=1}^{n} (-\log(\sqrt{2\pi}\sigma) - \frac{(x_i-u)^2}{2\sigma^2}) = -n\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(x_i-u)^2}{2\sigma^2}$$

$$\frac{\partial \log P(D|u,\sigma^2)}{\partial u} = -\sum_{i=1}^{n} \frac{2(x_i-u)}{2\sigma^2} = \frac{-\sum_{i=1}^{n} x_i + nu}{\sigma^2} = 0$$

$$-\sum_{i=1}^{n} x_i + nu = 0$$

MLE for Gaussian mean and variance

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

Note: MLE for the variance of a Gaussian is **biased**

- Expected result of estimation is **not** true parameter!
- Unbiased variance estimator:

$$\widehat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

MAP for Gaussian mean and variance

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}} \quad = N(\eta, \lambda^2)$$

MAP for Gaussian Mean

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\eta}{\lambda^2}}{\frac{n}{\sigma^2} + \frac{1}{\lambda^2}}$$

(Assuming known variance σ^2)

What you should know…

- Learning parametric distributions: form known, parameters unknown
  - Bernoulli (θ, probability of flip)
  - Gaussian (μ, mean and σ^2, variance)
- MLE
- MAP

What loss function are we minimizing?

- Learning distributions/densities
- **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

Learn a Probabilistic 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?

Prior: $P(Y = y)$ for all y **K-1 if K labels**

Likelihood: $P(X=x|Y = y)$ for all x,y **$(2^d - 1)K$ if d binary features**

Learning the Optimal 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?

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

Need n >> $2^d K - 1$ number of training data to learn all parameters

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:

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

- e.g., $P(Thunder | Rain, Lightning) = P(Thunder | Lightning)$

  Note: does NOT mean Thunder is independent of Rain

Conditional vs. Marginal Independence

- C calls A and B separately and tells them a number $n \in \{1,\ldots,10\}$
- 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?
  - No, we expect e.g. $P(n_a = 1 \mid n_b = 1) > P(n_a = 1)$
- Are n_a and n_b conditionally independent given n?
  - Yes, because if we know the true number, the outcomes n_a and n_b are purely determined by the noise in each phone.

$$P(n_a = 1 \mid n_b = 1, n = 2) = P(n_a = 1 \mid n = 2)$$

Prediction using Conditional Independence

- Predict Lightening
- From two **conditionally Independent** features
  - Thunder
  - Rain

\# parameters needed to learn likelihood given L

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

With conditional independence assumption

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

Naïve Bayes Assumption

- 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:

$$P(X_1 \ldots X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

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

Naïve Bayes Assumption

- 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:

$$P(X_1 ... X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

- How many parameters now? **(2-1)dK vs. (2^d-1)K**
  - Suppose **X** is composed of *d* binary features

Naïve Bayes Classifier

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

$P(y \mid x_1, \ldots, x_d)$

- Decision rule:

$$
\begin{aligned}
f_{NB}(\mathbf{x}) &= \arg\max_y \; P(x_1, \ldots, x_d \mid y) P(y) \\
&= \arg\max_y \prod_{i=1}^{d} P(x_i|y) P(y)
\end{aligned}
$$

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

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- Maximum Likelihood Estimates
  - For Class Prior

$$\widehat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$

  - For Likelihood

$$\widehat{P} = (x_i|y) = \frac{\widehat{P}(x_i, y)}{\widehat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$

- NB Prediction for test data $X = (x_1, \ldots, x_d)$

$$Y = \arg\max_y \widehat{P}(y) \prod_{i=1}^{d} \frac{\widehat{P}(x_i, y)}{\widehat{P}(y)}$$

Subtlety 1 – Violation of NB Assumption

- Usually, features are not conditionally independent:

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

- 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

Subtlety 2 – 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:
  - $P(Y=b \mid X_1=a, X_2,\ldots,X_d) = 0$

$$\boxed{P(X_1 = a, X_2 \ldots X_n | Y) = P(X_1 = a | Y) \prod_{i=2}^{d} P(X_i | Y)}$$

- What now???

MLE vs. MAP

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

What if we toss the coin too few times?

- You say: Probability next toss is a head = 0
- Billionaire says: You're fired! ...with prob 1 ☺

$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

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

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $\quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- Maximum A Posteriori Estimates – add m "virtual" examples

Assume priors

$$Q(Y = b) \qquad\qquad Q(X_i = a, Y = b)$$

MAP Estimate

$$\widehat{P}(X_i = a | Y = b) = \frac{\{\#j : X_i^{(j)} = a, Y^{(j)} = b\} + mQ(X_i = a, Y = b)}{\{\#j : Y^{(j)} = b\} + \underbrace{mQ(Y = b)}_{\text{\# virtual examples with Y = b}}}$$

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

Case Study: Text Classification

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

- 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
```

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 at the i^{th} position 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**: $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

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

Bag of words approach

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

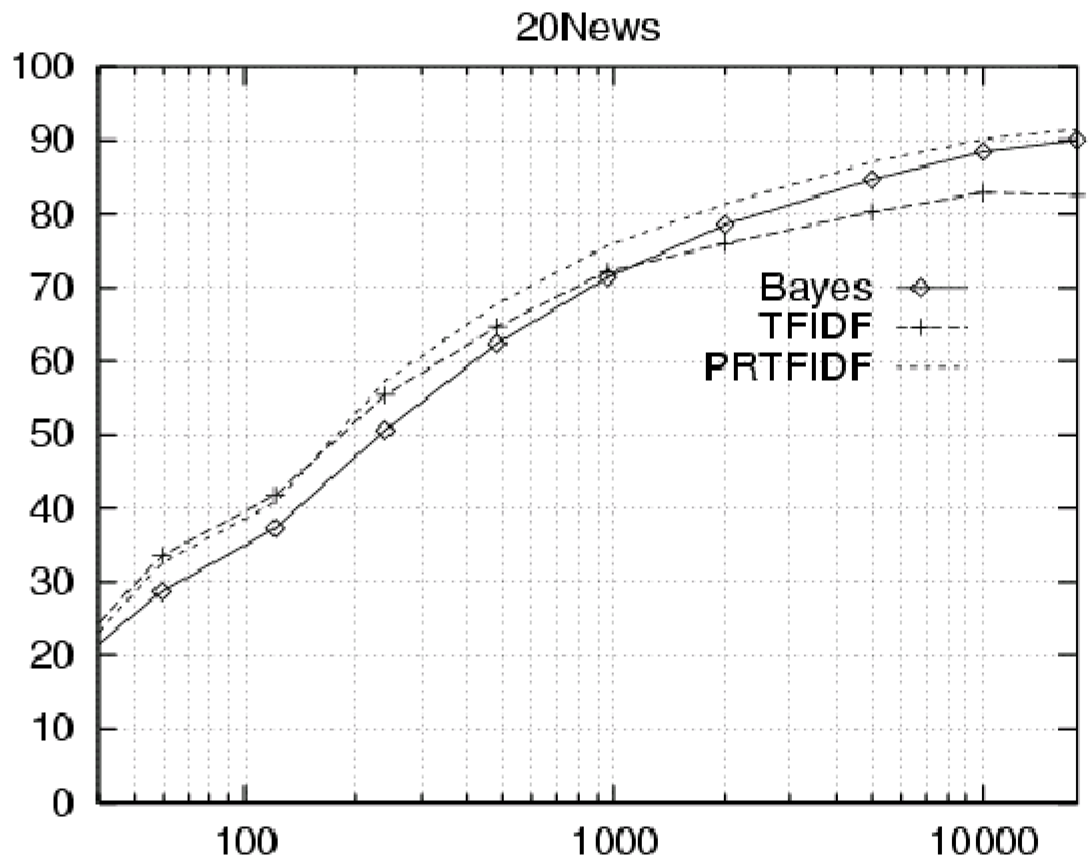
Twenty news groups results

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	

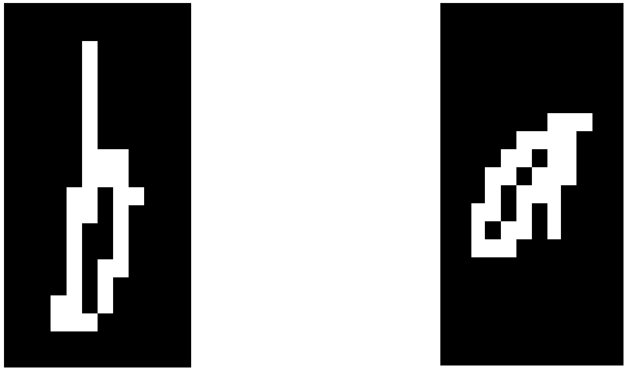
Learning curve for twenty news groups



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

What if features are continuous?

Eg., character recognition: X_i is intensity at i^{th} pixel



Gaussian Naïve Bayes (GNB):

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

Different mean and variance for each class k and each pixel i.

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:

$$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

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

k^{th} class

j^{th} training image

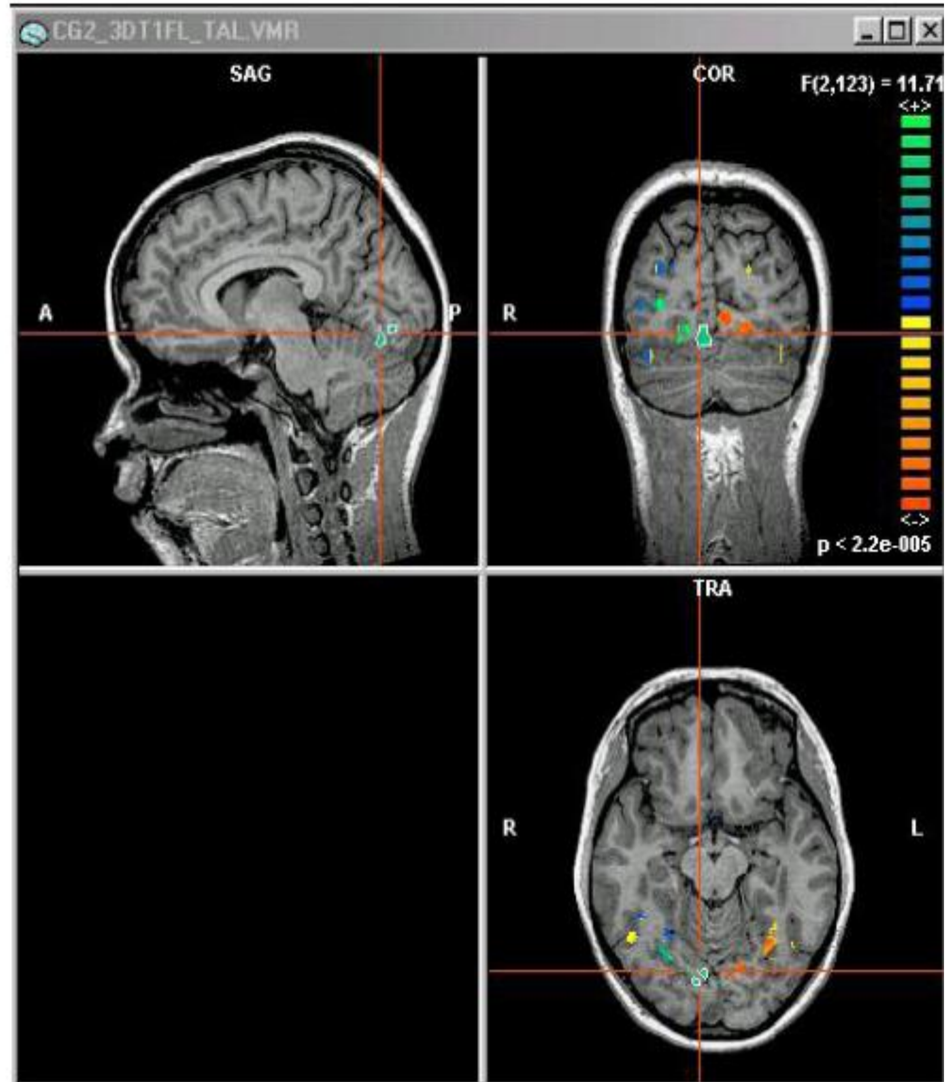
i^{th} pixel in j^{th} training image

$$\widehat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{j=1}^{N} (x_j - \widehat{\mu})^2$$

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

Example: GNB for classifying mental states

[Mitchell et al.]



~1 mm resolution

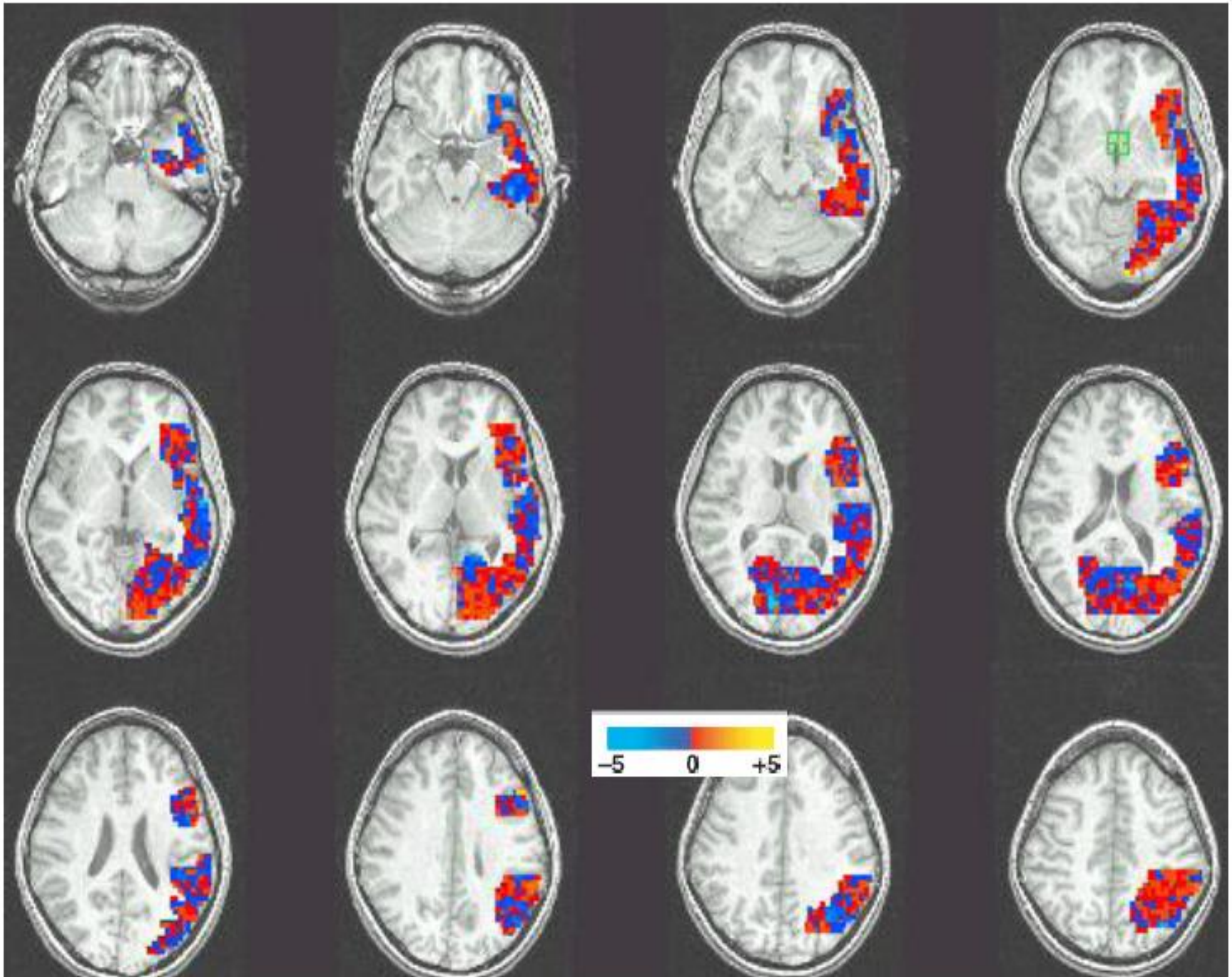
~2 images per sec.

15,000 voxels/image

non-invasive, safe

measures Blood Oxygen Level Dependent (BOLD) response

Gaussian Naïve Bayes: Learned $\mu_{voxel,word}$



[Mitchell et al.]

15,000 voxels or features

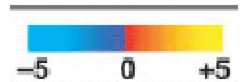
10 training examples or subjects per class

Learned Naïve Bayes Models –
Means for P(BrainActivity | WordCategory)

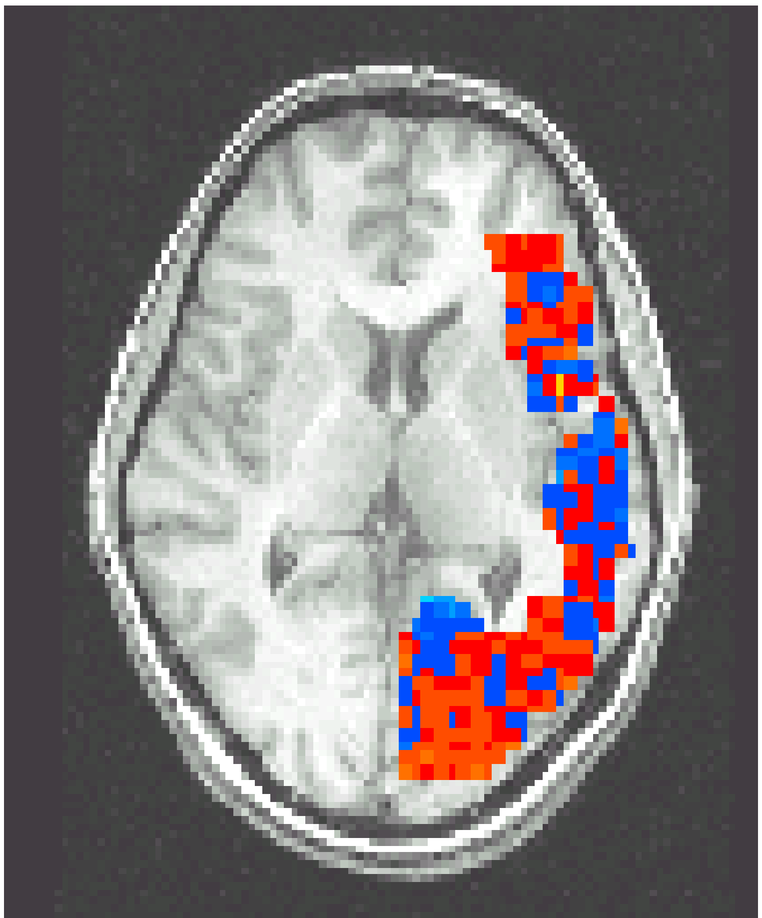
[Mitchell et al.]

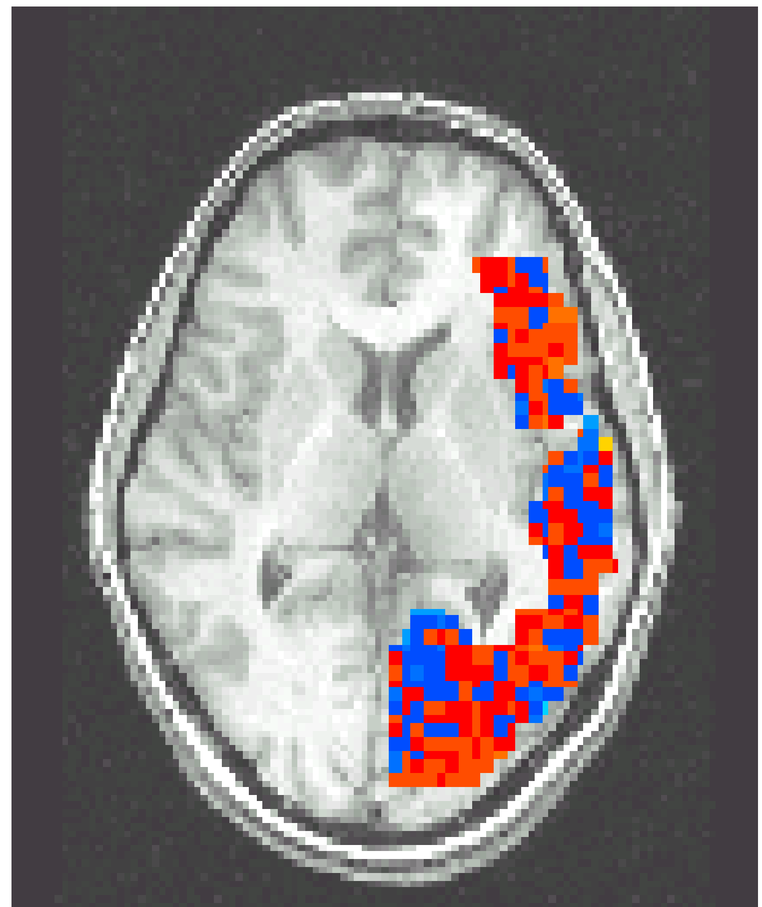
Pairwise classification accuracy: 85%

People words



Animal words





What you should know…

- Optimal decision using Bayes Classifier
- 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

Naïve Bayes Classifier Demo