Advanced Introduction to Machine Learning CMU-10715

MLE, MAP, Bayes classification

Barnabás Póczos

Sept 9, 2015

Outline

Theory:

- Probabilities:
  - Dependence, Independence, Conditional Independence
- Parameter estimation:
  - Maximum Likelihood Estimation (MLE)
  - Maximum aposteriori (MAP)
- Bayes rule
  - Naïve Bayes Classifier

Application:

Naive Bayes Classifier for
- Spam filtering
- “Mind reading” = fMRI data processing

Independence

Independence

Independent random variables:

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

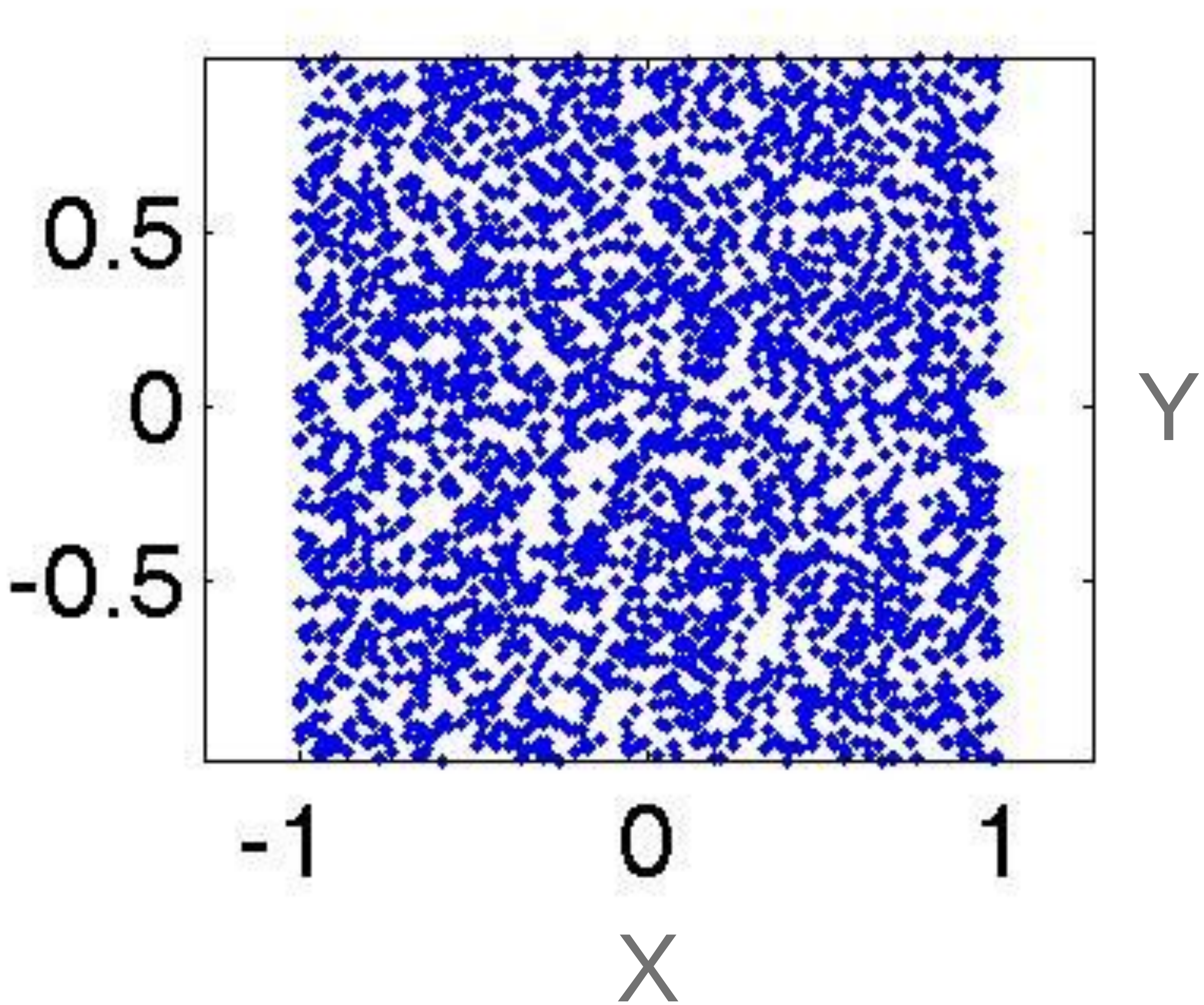
$$P(X|Y) = P(X)$$

Y and X don't contain information about each other.
Observing Y doesn't help predicting X.
Observing X doesn't help predicting Y.

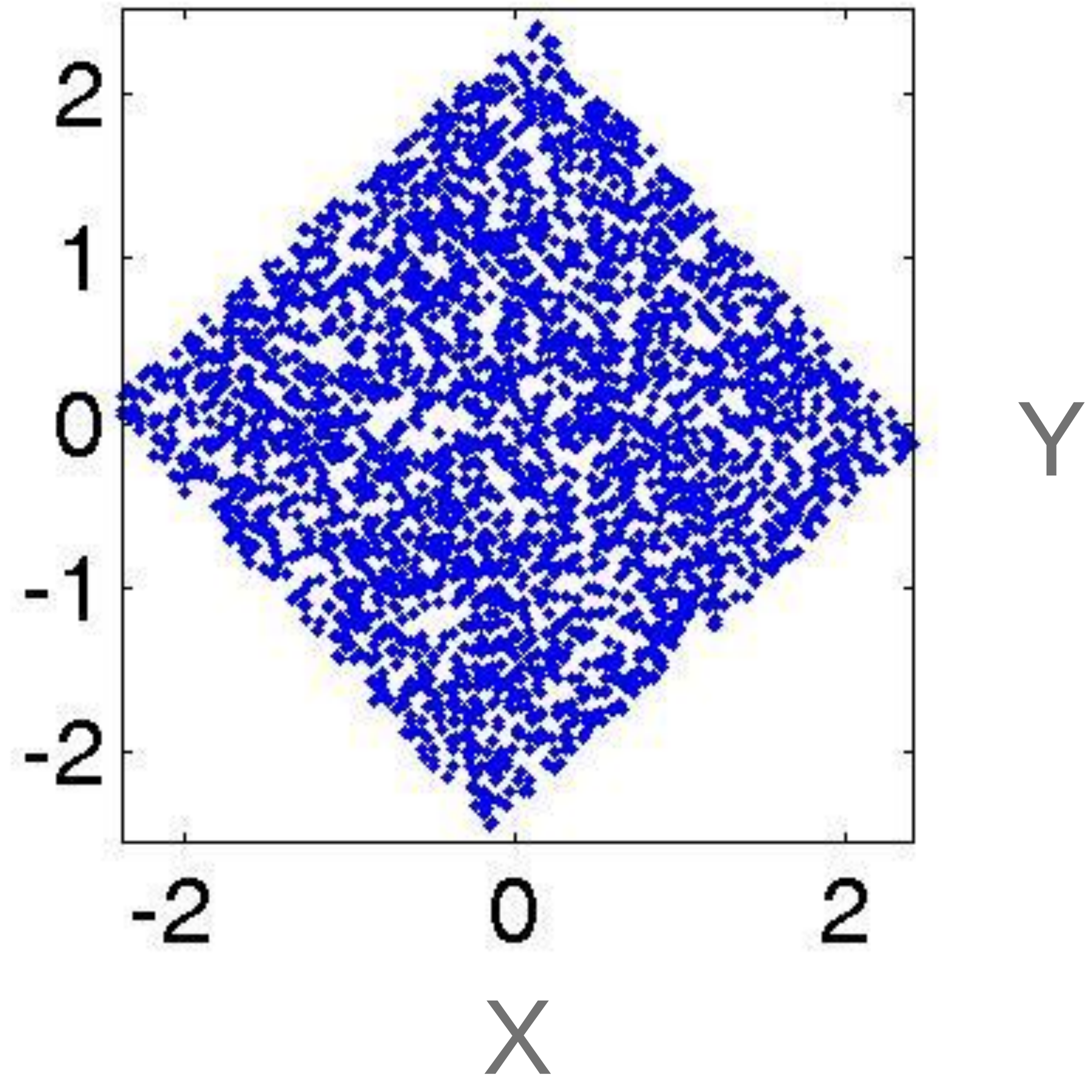
Examples:

Independent: Winning on roulette this week and next week.
Dependent: Russian roulette

Dependent / Independent



Independent X,Y



Dependent X,Y

Conditionally Independent

Conditionally independent:

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

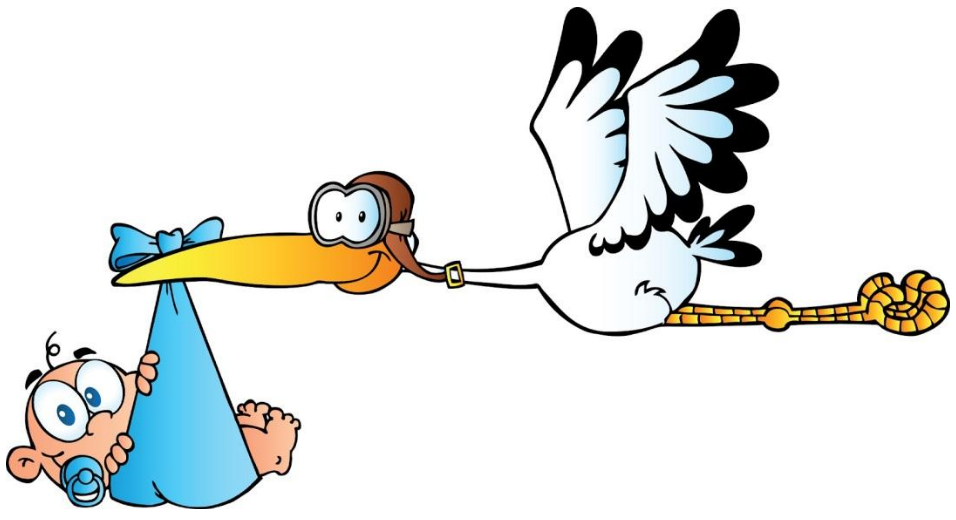
Knowing Z makes X and Y independent

Examples:

Dependent: shoe size and reading skills
Conditionally independent: shoe size and reading skills given age

Storks deliver babies:

Highly statistically significant correlation exists between stork populations and human birth rates across Europe.

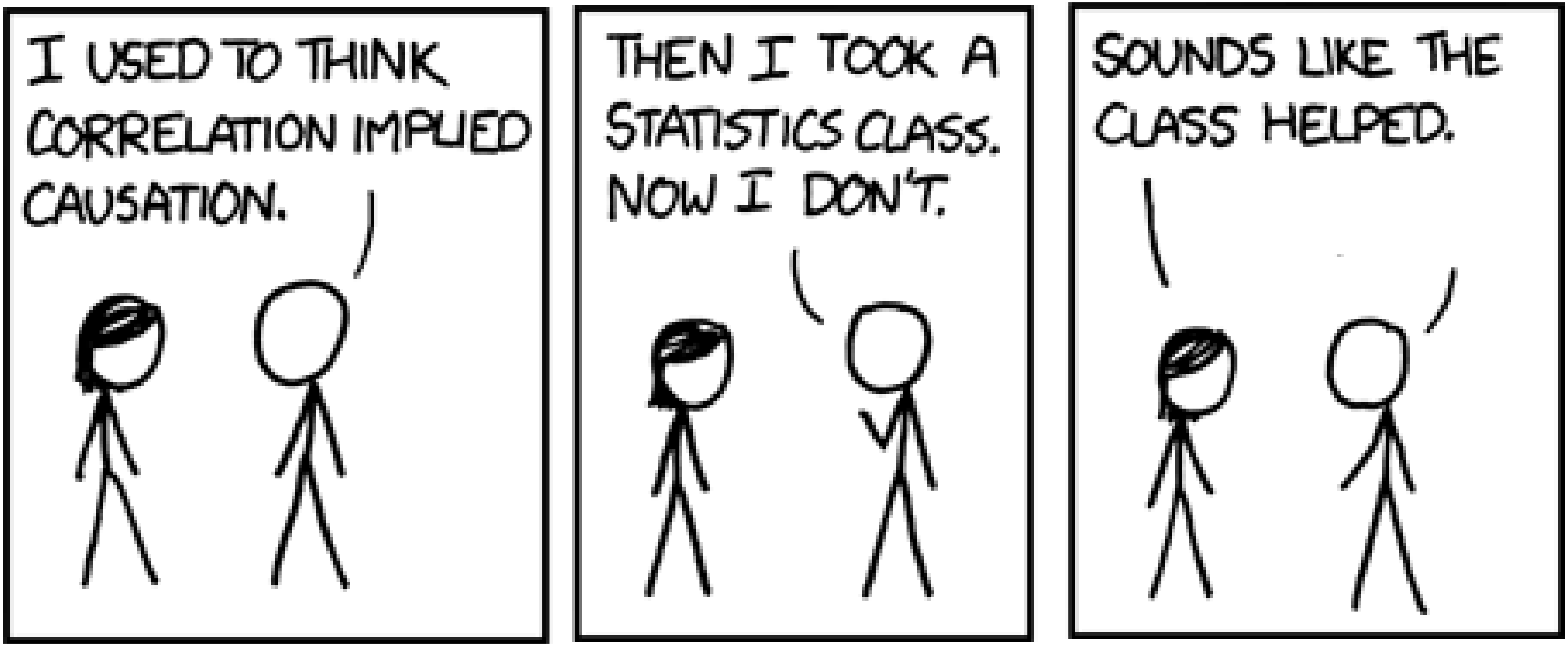


Conditionally Independent

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains…

Correlation ≠ Causation



xkcd.com

Our first machine learning problem:

Parameter estimation: MLE, MAP

Estimating Probabilities

Flipping a Coin

I have a coin, if I flip it, what's the probability that it will fall with the head up?

Let us flip it a few times to estimate the probability:

The estimated probability is: **3/5** "Frequency of heads"

Flipping a Coin

The estimated probability is: **3/5** "Frequency of heads"

Questions:

(1) Why frequency of heads???

(2) How good is this estimation???

(3) Why is this a machine learning problem???

We are going to answer these questions

Question (1)

Why frequency of heads???

- Frequency of heads is exactly the *maximum likelihood estimator* for this problem
- MLE has nice properties

Maximum Likelihood Estimation

MLE for Bernoulli distribution

Data, D =

$$D = \{X_i\}_{i=1}^n, \ X_i \in \{\mathrm{H}, \mathrm{T}\}$$

P(Heads) = θ, *P(Tails)* = $1-\theta$

Flips are **i.i.d.**:
- **Independent** events
  - **Identically distributed** according to Bernoulli distribution

MLE: Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$

Identically distributed

$$= \arg\max_{\theta} \; \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

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

$$= \arg\max_{\theta} \; \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H}(1-\theta)^{\alpha_T - 1}\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\alpha_H(1-\theta) - \alpha_T\theta\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

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

That's exactly the "**Frequency of heads**"

Question (2)

How good is this MLE estimation???

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

How many flips do I need?

I flipped the coins 5 times: 3 heads, 2 tails

$$\widehat{\theta}_{MLE} = \frac{3}{5}$$

What if I flipped 26 heads and 24 tails?

$$\widehat{\theta}_{MLE} = \frac{26}{50}$$

- **Which estimator should we trust more?**

Simple bound

Let θ^* be the true parameter.

For $n = \alpha_H + \alpha_T$, and $\quad \widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

For any $\varepsilon > 0$:

Hoeffding's inequality:

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

Probably Approximate Correct (PAC)Learning

I want to know the coin parameter θ, within $\varepsilon = 0.1$ error with probability at least $1\text{-}\delta = 0.95$.

How many flips do I need?

$$P(|\ \widehat{\theta} - \theta^*\ | \geq \epsilon) \ \leq \ 2e^{-2n\epsilon^2} \leq \delta$$

Sample complexity:

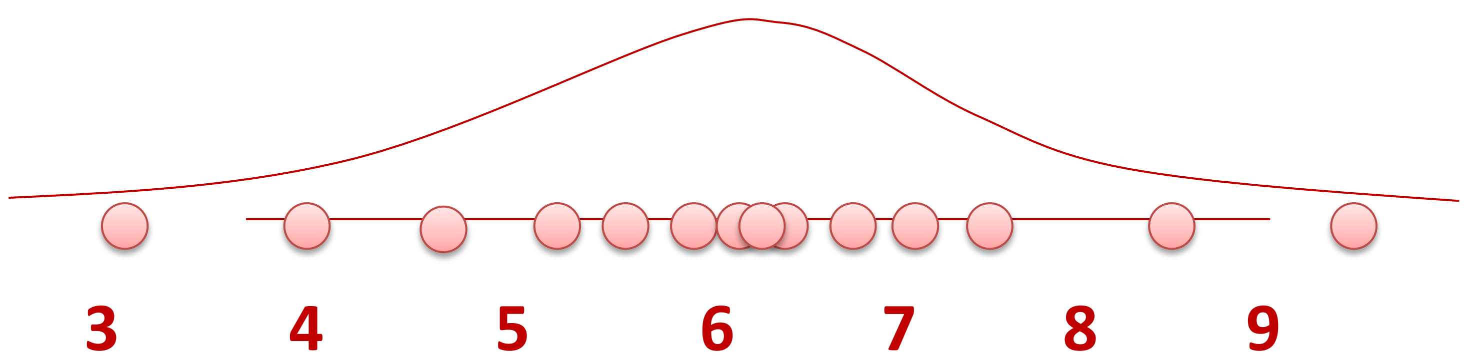
$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

Question (3)

Why is this a machine learning problem???

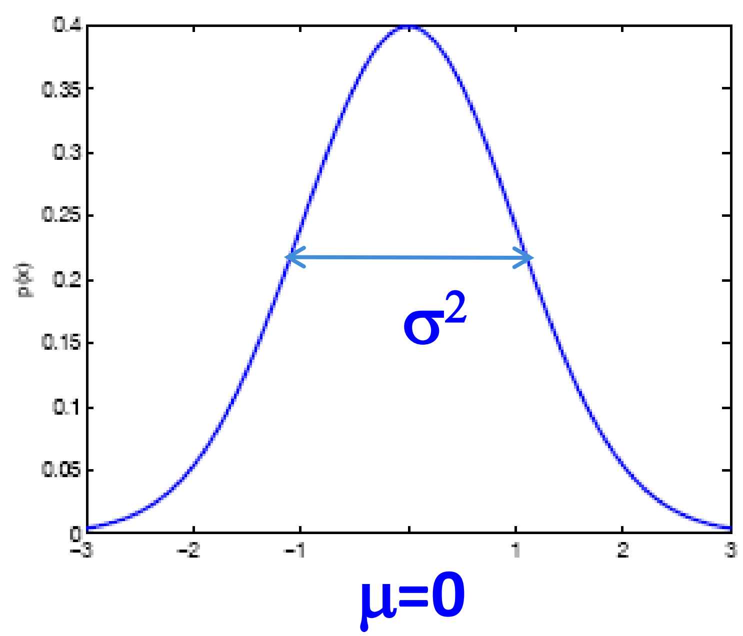
- improve their performance (accuracy of the predicted prob.)
- at some task (predicting the probability of heads)
- with experience (the more coins we flip the better we are)

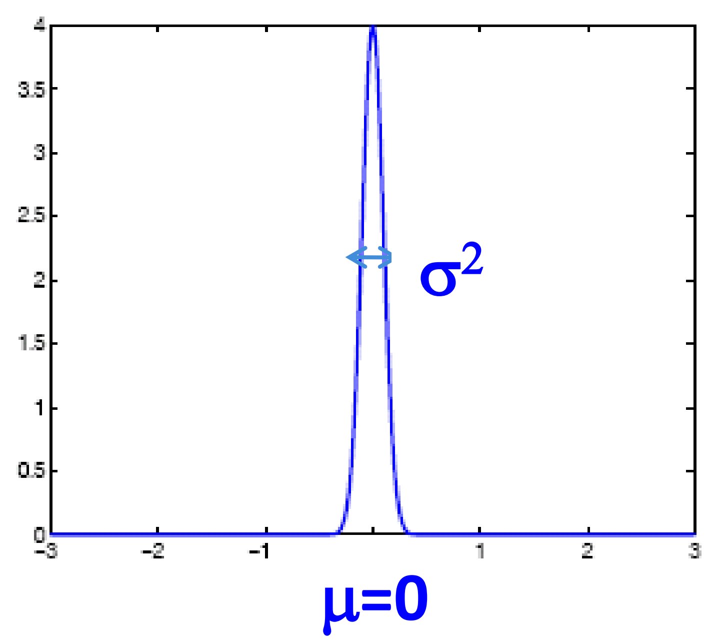
What about continuous features?



Let us try Gaussians…

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) = \mathcal{N}_x(\mu, \sigma)$$





MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

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

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-(X_i-\mu)^2/2\sigma^2}$$

Identically distributed

$$= \arg \max_{\theta=(\mu,\sigma^2)} \underbrace{\frac{1}{2\sigma^2} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}}_{J(\theta)}$$

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** the true parameter!]

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

What about prior knowledge?

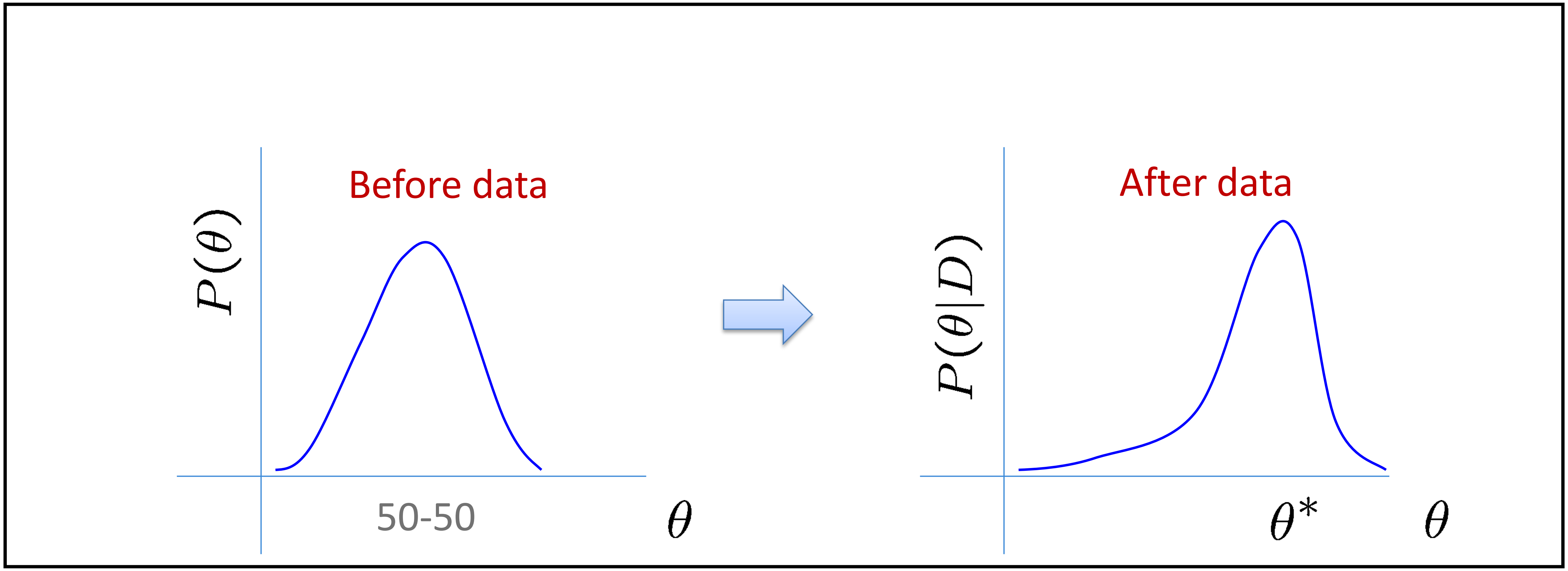
(MAP Estimation)

What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way…

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



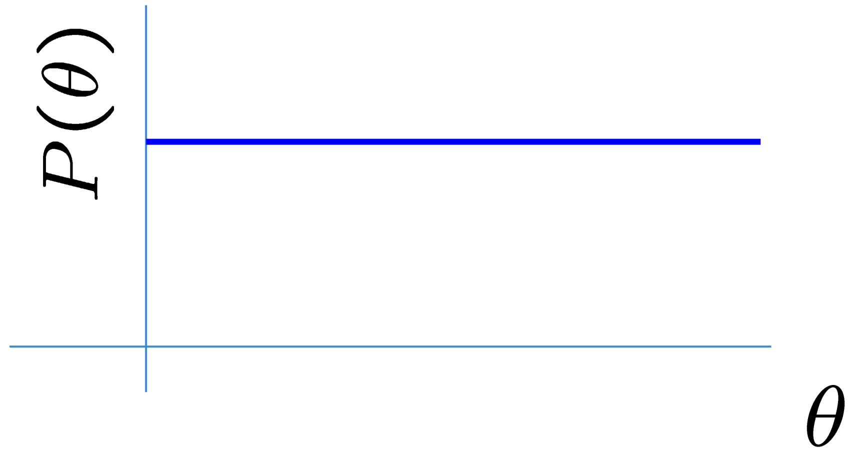
Prior distribution

What kind of prior distribution do we want to use?

- 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

Bayes Rule

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**

Chain Rule & Bayes Rule

Chain rule:

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

Bayes rule:

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

Bayes rule is important for reverse conditioning.

Bayesian Learning

- Use Bayes rule:

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

- Or equivalently:

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

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 estimation for Binomial distribution

Coin flip problem: Likelihood is Binomial

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

If the prior is Beta distribution,

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

$\Rightarrow$ posterior is Beta distribution

Beta function: $$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt$$

MAP estimation for Binomial distribution

Likelihood is Binomial: $P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$

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

$\Rightarrow$ posterior is Beta distribution

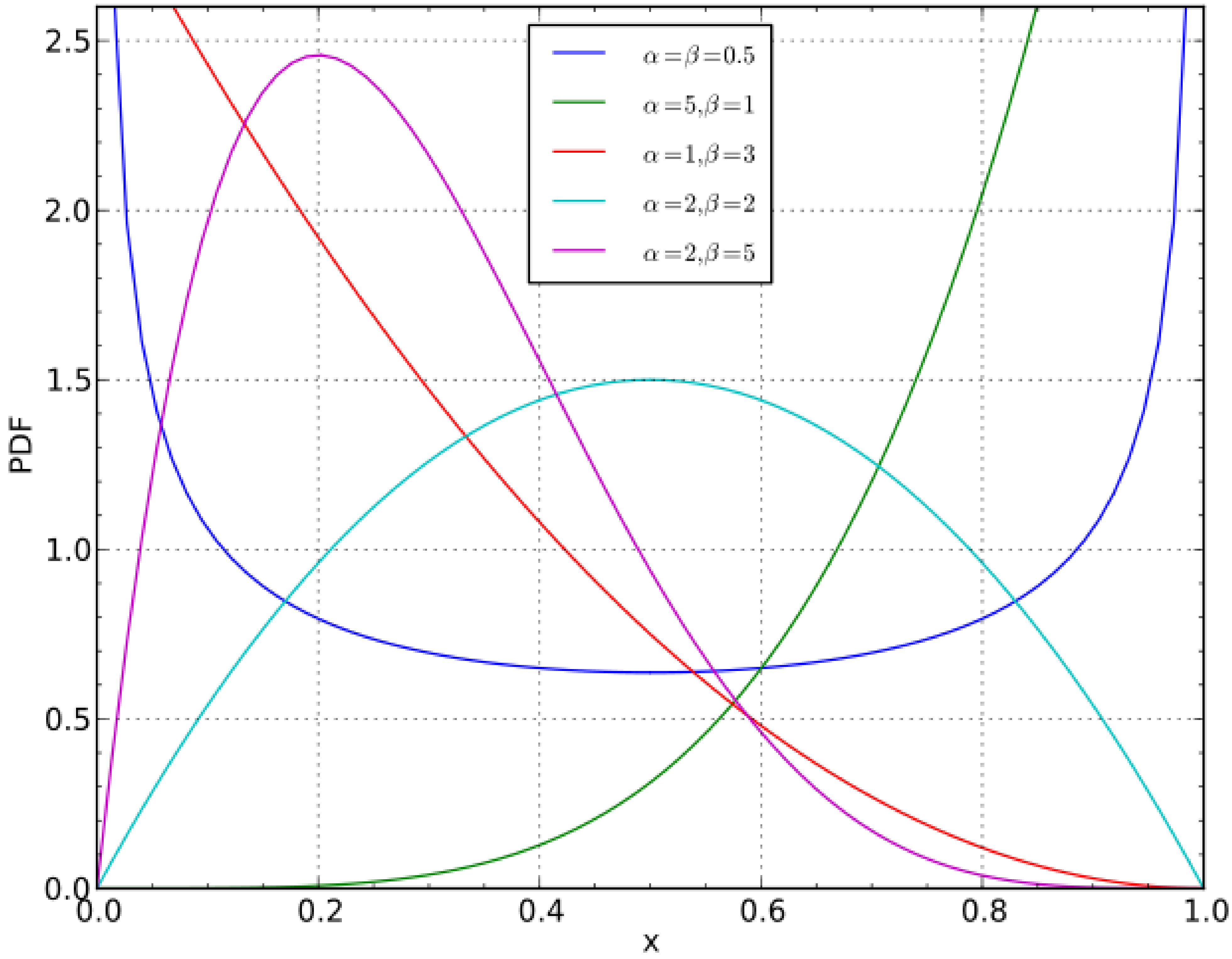
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

P(θ) and P(θ|D) have the same form! [Conjugate prior]

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

$$= \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

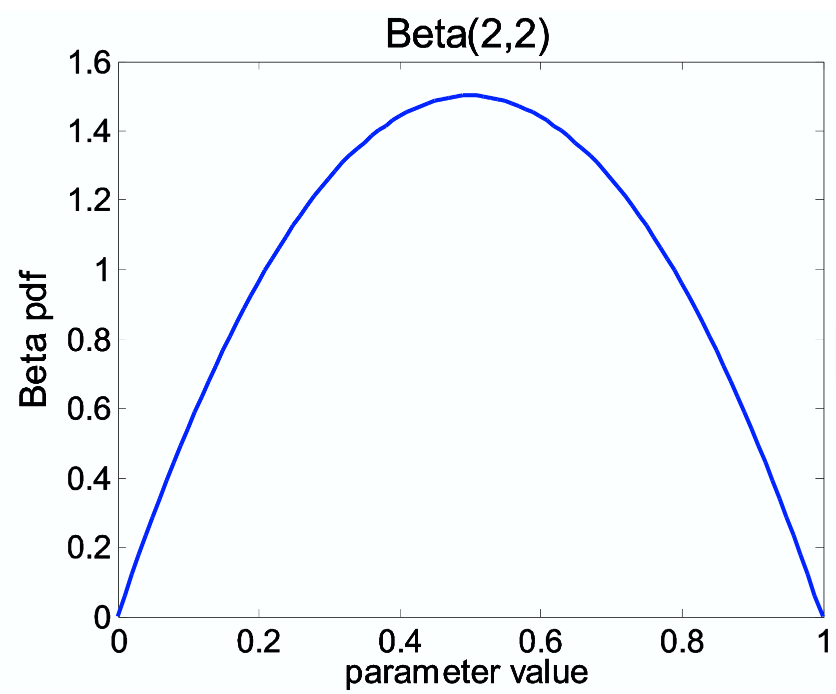
Beta distribution

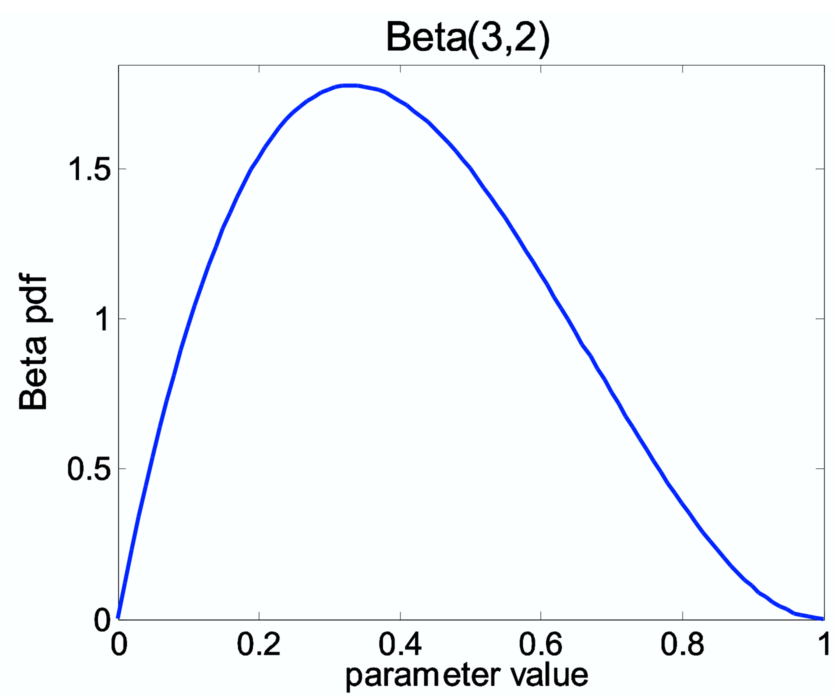


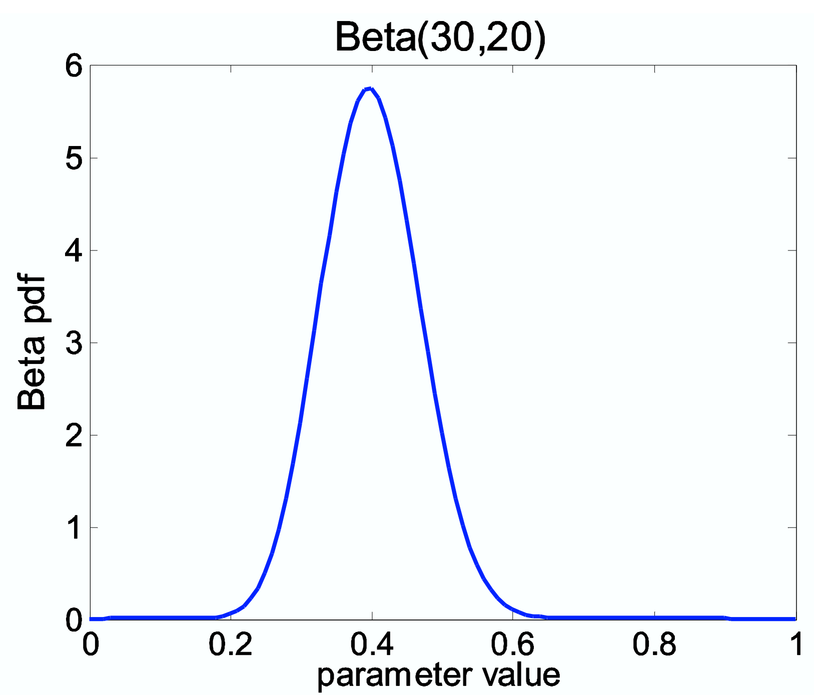
More concentrated as values of α, β increase

Beta conjugate prior

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







As $n = \alpha_H + \alpha_T$ increases

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

From Binomial to Multinomial

Example: 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.

Bayes Rule Application

AIDS test (Bayes rule)

Data

- Approximately 0.1% are infected
- Test detects all infections
- Test reports positive for 1% healthy people

Probability of having AIDS if test is positive:

$$P(a=1|t=1) = \frac{P(t=1|a=1)P(a=1)}{P(t=1)}$$

$$= \frac{P(t=1|a=1)P(a=1)}{P(t=1|a=1)P(a=1) + P(t=1|a=0)P(a=0)}$$

$$= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$$

Only 9%!...

Improving the diagnosis

Use a follow-up test!

- Test 2 reports positive for 90% infections
- Test 2 reports positive for 5% healthy people

$$P(a=0|t_1=1,t_2=1)=\frac{P(t_1=1,t_2=1|a=0)P(a=0)}{P(t_1=1,t_2=1|a=1)P(a=1)+P(t_1=1,t_2=1|a=0)P(a=0)}$$

$$=\frac{0.01\cdot 0.05\cdot 0.999}{1\cdot 0.9\cdot 0.001+0.01\cdot 0.05\cdot 0.999}=0.357$$

$$P(a=1|t_1=1,t_2=1)=0.643$$

Why can't we use Test 1 twice?

Outcomes are **not** independent but tests 1 and 2 are **conditionally independent** $p(t_1,t_2|a)=p(t_1|a)\cdot p(t_2|a)$

The Naïve Bayes Classifier

Data for spam filtering

- date
- time
- recipient path
- IP number
- sender
- encoding
- many more features

```
Delivered-To: alex.smola@gmail.com
Received: by 10.216.47.73 with SMTP id s51cs361171web;
        Tue, 3 Jan 2012 14:17:53 -0800 (PST)
Received: by 10.213.17.145 with SMTP id s17mr2519891eba.147.1325629071725;
        Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Return-Path: <alex+caf_=alex.smola=gmail.com@smola.org>
Received: from mail-ey0-f175.google.com (mail-ey0-f175.google.com [209.85.215.175])
        by mx.google.com with ESMTPS id n4si29264232eef.57.2012.01.03.14.17.51
        (version=TLSv1/SSLv3 cipher=OTHER);
        Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Received-SPF: neutral (google.com: 209.85.215.175 is neither permitted nor denied by best guess record for domain of alex+caf_=alex.smola=gmail.com@smola.org) client-ip=209.85.215.175;
Authentication-Results: mx.google.com; spf=neutral (google.com: 209.85.215.175 is neither permitted nor denied by best guess record for domain of alex+caf_=alex.smola=gmail.com@smola.org) smtp.mail=alex+caf_=alex.smola=gmail.com@smola.org; dkim=pass (test mode) header.i=@googlemail.com
Received: by eaal1 with SMTP id l1so15092746eaa.6
        for <alex.smola@gmail.com>; Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Received: by 10.205.135.18 with SMTP id ie18mr5325064bkc.72.1325629071362;
        Tue, 03 Jan 2012 14:17:51 -0800 (PST)
X-Forwarded-To: alex.smola@gmail.com
X-Forwarded-For: alex@smola.org alex.smola@gmail.com
Delivered-To: alex@smola.org
Received: by 10.204.65.198 with SMTP id k6cs206093bki;
        Tue, 3 Jan 2012 14:17:50 -0800 (PST)
Received: by 10.52.88.179 with SMTP id bh19mr10729402vdb.38.1325629068795;
        Tue, 03 Jan 2012 14:17:48 -0800 (PST)
Return-Path: <althoff.tim@googlemail.com>
Received: from mail-vx0-f179.google.com (mail-vx0-f179.google.com [209.85.220.179])
        by mx.google.com with ESMTPS id dt4si11767074vdb.93.2012.01.03.14.17.48
        (version=TLSv1/SSLv3 cipher=OTHER);
        Tue, 03 Jan 2012 14:17:48 -0800 (PST)
Received-SPF: pass (google.com: domain of althoff.tim@googlemail.com designates 209.85.220.179 as permitted sender) client-ip=209.85.220.179;
Received: by vcbf13 with SMTP id f13so11295098vcb.10
        for <alex@smola.org>; Tue, 03 Jan 2012 14:17:48 -0800 (PST)
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed;
        d=googlemail.com; s=gamma;
        h=mime-version:sender:date:x-google-sender-auth:message-id:subject
         :from:to:content-type;
        bh=WCbdZ5sXac25dpH02XcRyDOdts993hKwsAVXpGrFh0w=;
        b=WK2B2+ExWnf/gvTkw6uUvKuP4XeoKnlJq3USYTm0RARK8dSFjyOQsIHeAP9Yssxp6O
         7ngGoTzYqd+ZsyJfvQcLAWp1PCJhG8AMcnqWkx0NMeoFvIp2HQooZwxSOCx5ZRgY+7qX
         uIbbdna4lUDXj6UFe16SpLDCkptd8OZ3gr7+o=
MIME-Version: 1.0
Received: by 10.220.108.81 with SMTP id e17mr24104004vcp.67.1325629067787;
 Tue, 03 Jan 2012 14:17:47 -0800 (PST)
Sender: althoff.tim@googlemail.com
Received: by 10.220.17.129 with HTTP; Tue, 3 Jan 2012 14:17:47 -0800 (PST)
Date: Tue, 3 Jan 2012 14:17:47 -0800
X-Google-Sender-Auth: 6bwi6D17HjZIkxOEol38NZzyeHs
Message-ID: <CAFJJHDGPBW+SdZg0MdAABiAKydDk9tpeMoDijYGjoGO-WC7osg@mail.gmail.com>
Subject: CS 281B. Advanced Topics in Learning and Decision Making
From: Tim Althoff <althoff@eecs.berkeley.edu>
To: alex@smola.org
Content-Type: multipart/alternative; boundary=f46d043c7af4b07e8d04b5a7113a

--f46d043c7af4b07e8d04b5a7113a
Content-Type: text/plain; charset=ISO-8859-1
```

Naïve Bayes Assumption

Naïve Bayes assumption: Features X_1 and X_2 are conditionally independent given the class label Y:

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

More generally:

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

How many parameters to estimate?

(X is composed of d binary features, e.g. presence of word "earn" in a text. Y has K possible class labels)

$\boxed{(2^d-1)K \text{ vs } (2-1)dK}$

Naïve Bayes Classifier

Given:
- Class prior $P(Y)$
- d conditionally independent features $X_1,\ldots,X_d$ given the class label Y
- For each X_i, we have the conditional likelihood $P(X_i|Y)$

Decision rule:

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

Naïve Bayes Algorithm for discrete features

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

n d dimensional features + class labels

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

We need to estimate these probabilities!

Estimate them with Relative Frequencies!

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

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

NB Prediction for test data:

$$X = (x_1, \ldots, x_d)$$

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

Subtlety: Insufficient training data

What if you never see a training instance where $X_1 = a$ when $Y = b$?

For example,

there is no X_1='Earn' when Y='SpamEmail' in our dataset.

$$\Rightarrow P(X_1 = a, Y = b) = 0 \Rightarrow P(X_1 = a | Y = b) = 0$$

$$\boxed{\Rightarrow P(X_1 = a, X_2 ... X_n | Y) = P(X_1 = a | Y) \prod_{i=2}^{d} P(X_i | Y) = 0}$$

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 now???

Case Study: Text Classification

Case Study: Text Classification

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

What about the features **X**?

The text!

X_i represents i^{th} word in 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 decided
```

NB for Text Classification

P(**X**|Y) is huge!!!

- Article at least 1000 words, **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). $X_i \in \{1,\ldots,50000\} \Rightarrow K50000^{1000}$ parameters….

NB assumption helps a lot!!!

- $P(X_i=x_i|Y=y)$ is the probability of observing word x_i at the i^{th} position in a document on topic y $\Rightarrow$ 1000K50000 parameters

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

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! $\Rightarrow$ K50000 parameters

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

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**: $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$

- "Bag of words" model – order of words on the page ignored
- Sounds really silly, but often works very well!

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

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

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

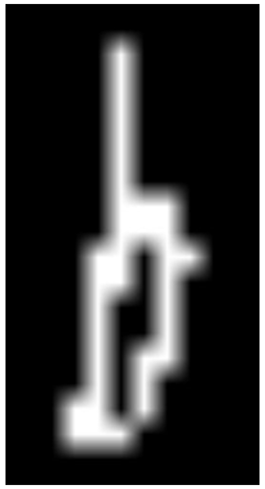
comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey

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	

Naïve Bayes: 89% accuracy

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)

Example: GNB for classifying mental states

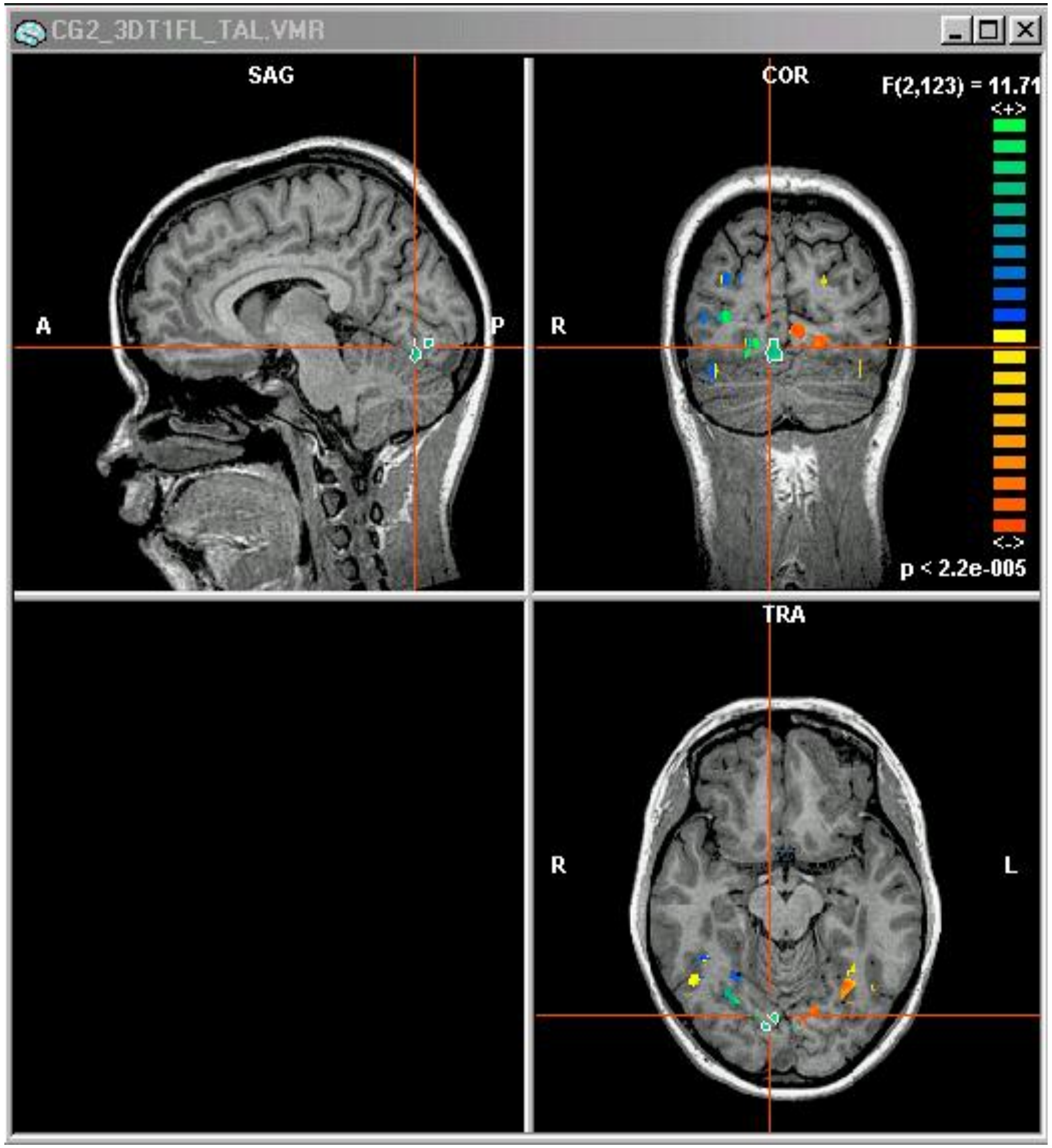
~1 mm resolution

~2 images per sec.

15,000 voxels/image

non-invasive, safe

measures Blood Oxygen Level Dependent (BOLD) response



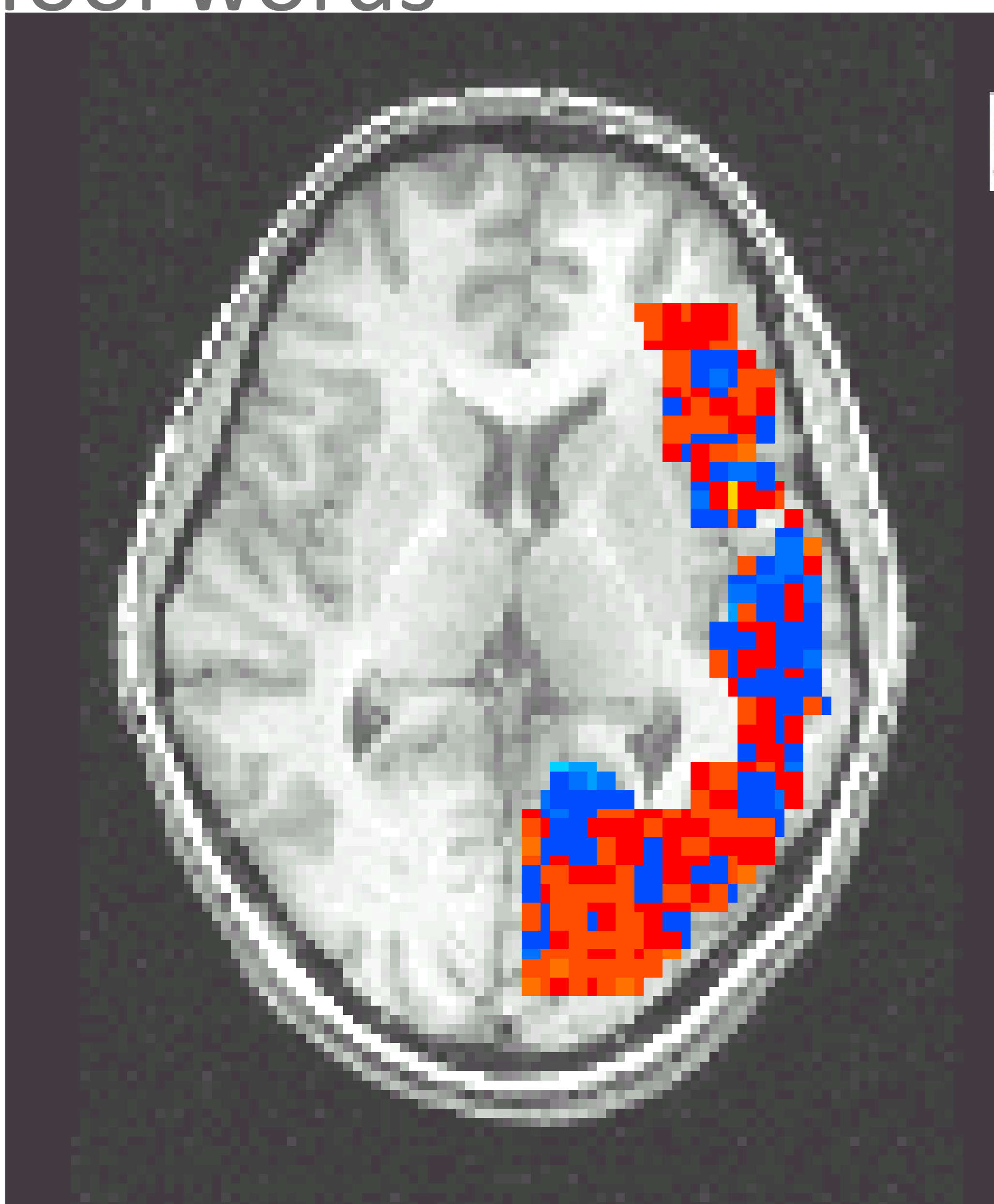
[Mitchell et al.]

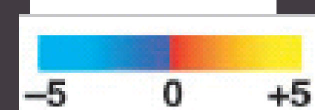
Learned Naïve Bayes Models – Means for P(BrainActivity | WordCategory)

Pairwise classification accuracy: 78-99%, 12 participants

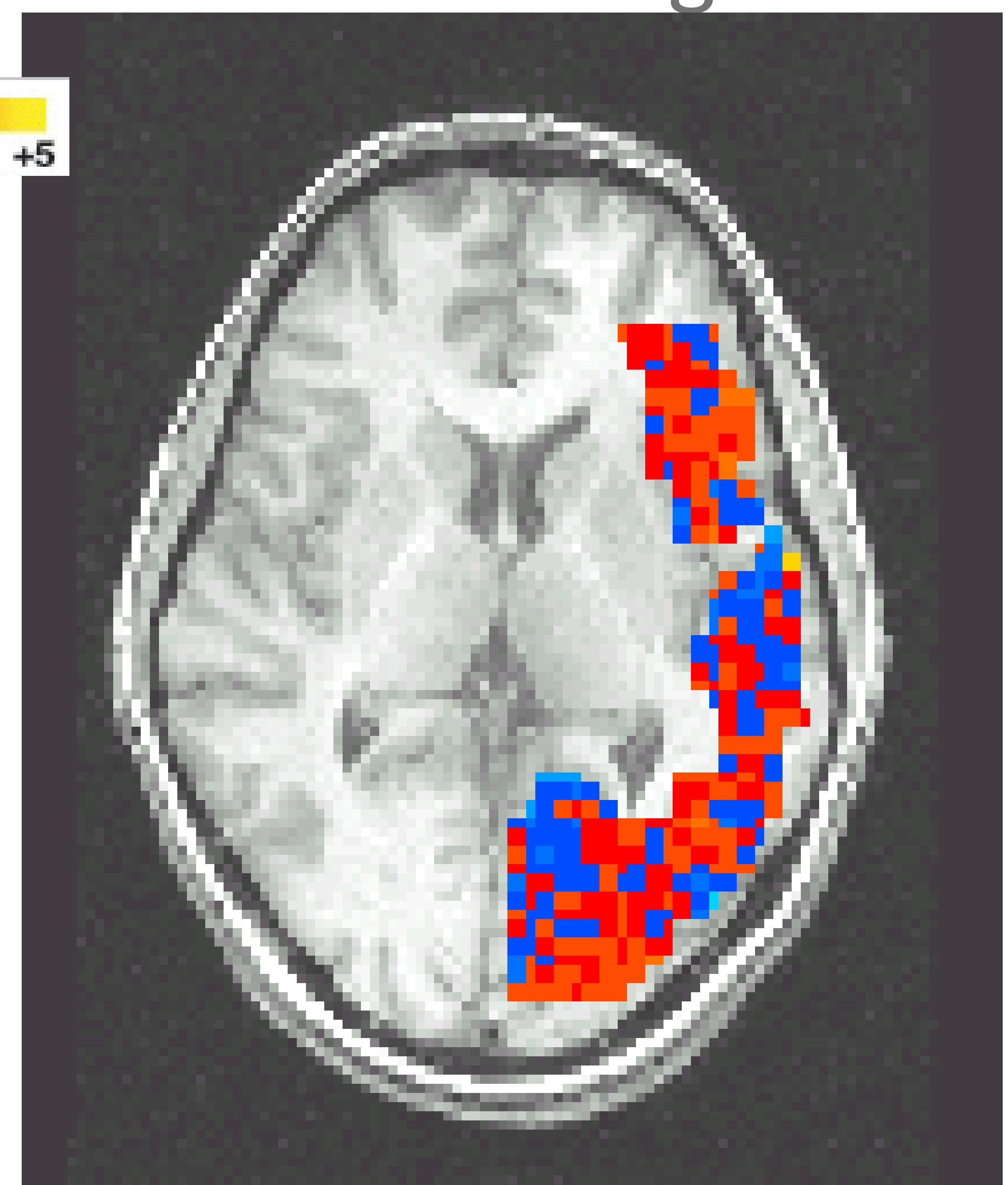
[Mitchell et al.]

Tool words





Building words



What you should know…

Naïve Bayes classifier

- What's the assumption
- Why we use it
- How do we learn it
- Why is Bayesian (MAP) estimation important

Text classification

- Bag of words model

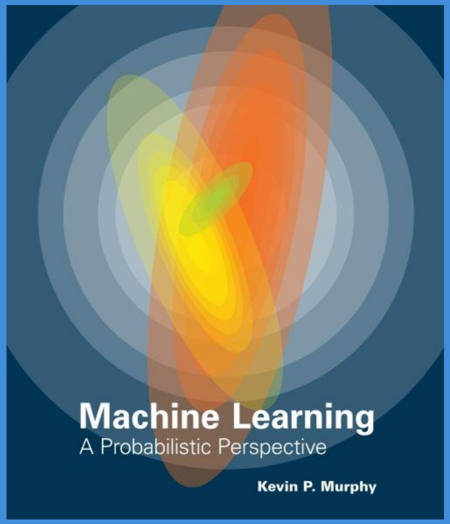
Gaussian NB

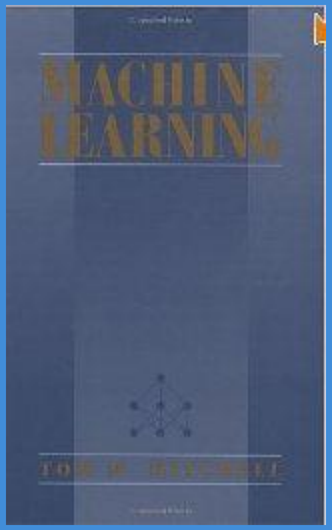
- Features are still conditionally independent
- Each feature has a Gaussian distribution given class

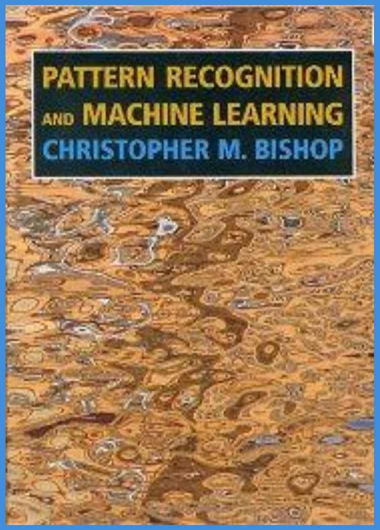
Further reading

Manuscript (book chapters 1 and 2)
http://alex.smola.org/teaching/berkeley2012/slides/chapter1_2.pdf

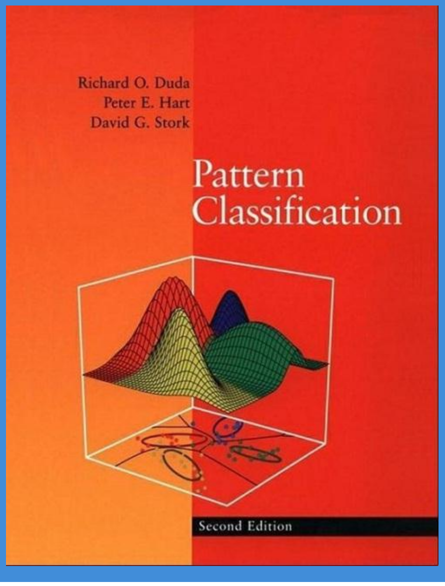
ML Books

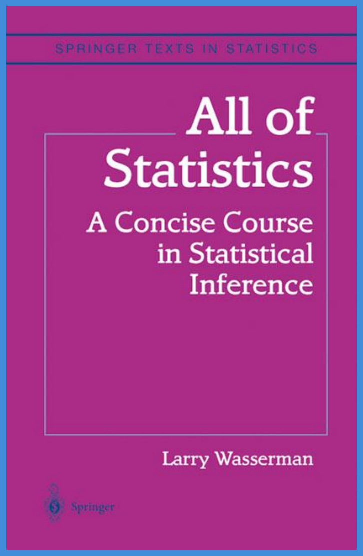


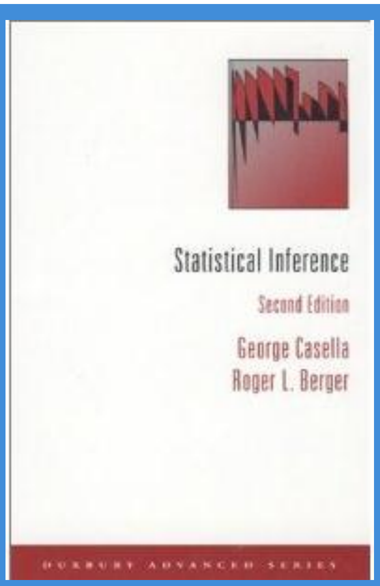




Statistics 101







Thanks for your attention ☺

References

Many slides are taken from

- Tom Mitchel
  http://www.cs.cmu.edu/~tom/10701_sp11/slides
- Alex Smola
- Aarti Singh
- Eric Xing
- Xi Chen
- http://www.math.ntu.edu.tw/~hchen/teaching/StatInference/notes/lecture2.pdf
- Wikipedia