

BBM406

Fundamentals of Machine Learning

Lecture 8:
Maximum a Posteriori (MAP)
Naïve Bayes Classifier

Recap: MLE

- Maximum Likelihood estimation (MLE)

Choose value that maximizes the probability of observed data

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

Today

- Maximum a Posteriori (MAP)
- Bayes rule
 - Naïve Bayes Classifier
- Application
 - Text classification
 - “Mind reading” = fMRI data processing

What about prior knowledge? (MAP Estimation)

What about prior knowledge?

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

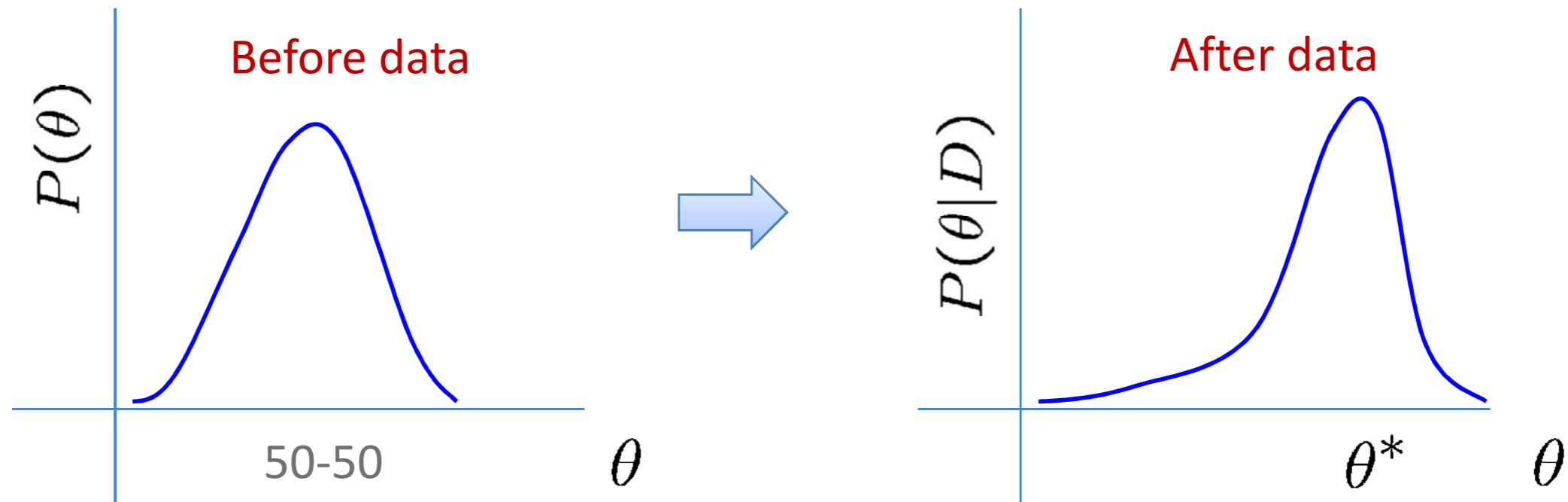
The Bayesian way...

What about prior knowledge?

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

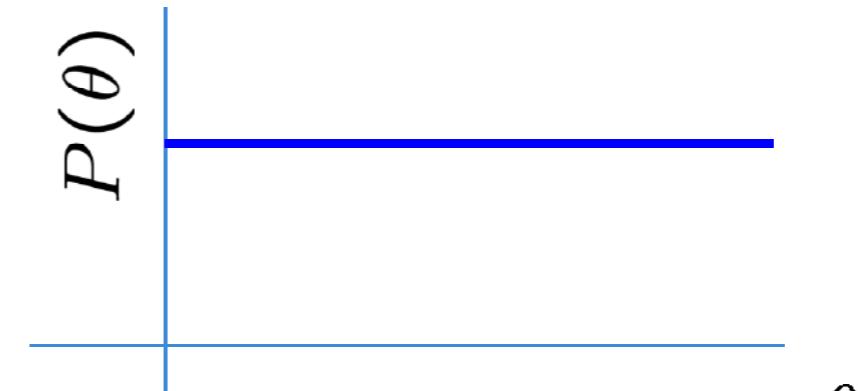
The Bayesian way...

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



Prior distribution

- What prior? What distribution do we want for a 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



In order to proceed we will need:

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:

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

- Or equivalently:

$$P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$$

posterior likelihood 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

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial

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

If the prior is Beta distribution,

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

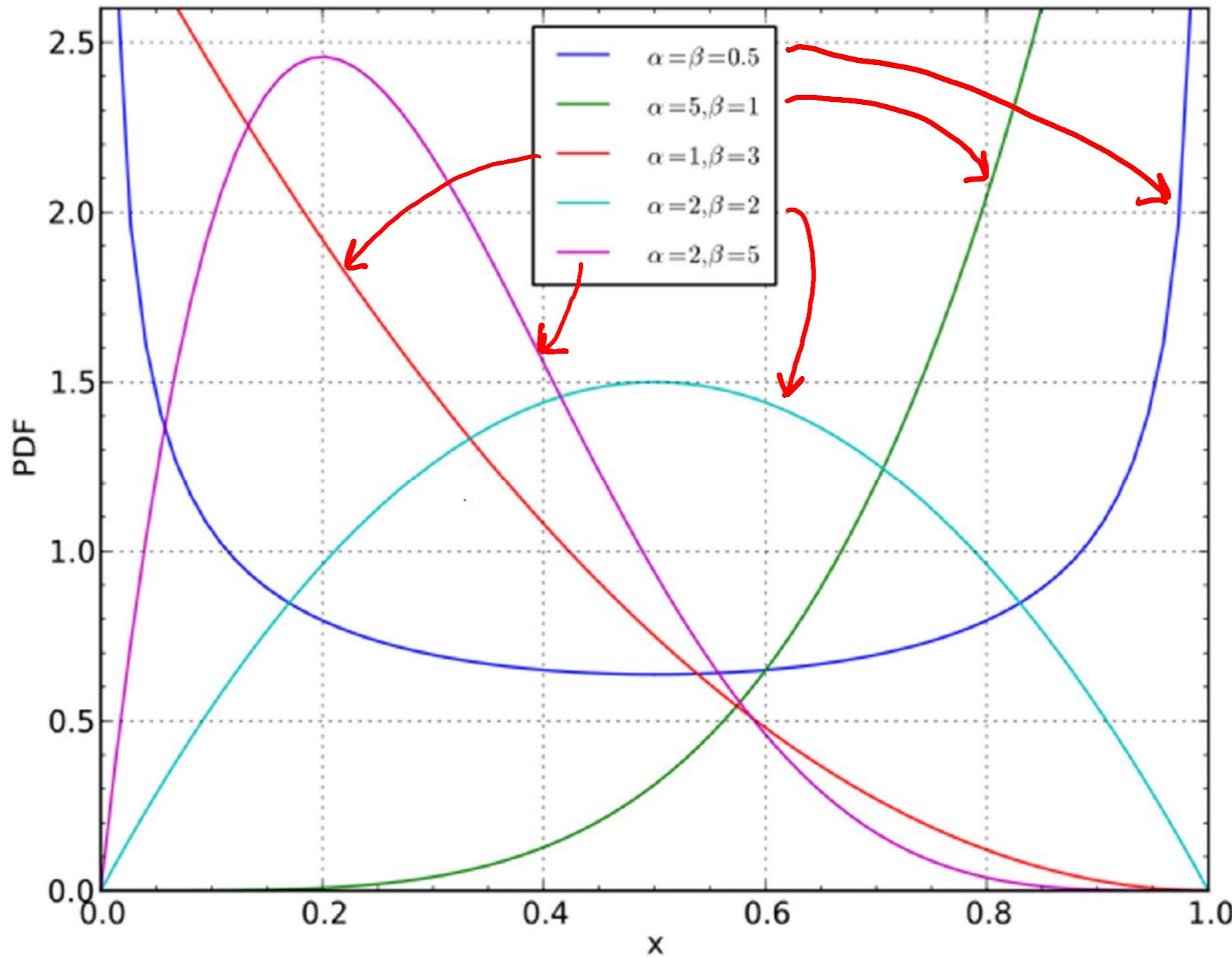
⇒ posterior is Beta distribution

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

$P(\theta)$ and $P(\theta|D)$ have the same form! [Conjugate prior]

$$\hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} P(D | \theta)P(\theta) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Beta distribution

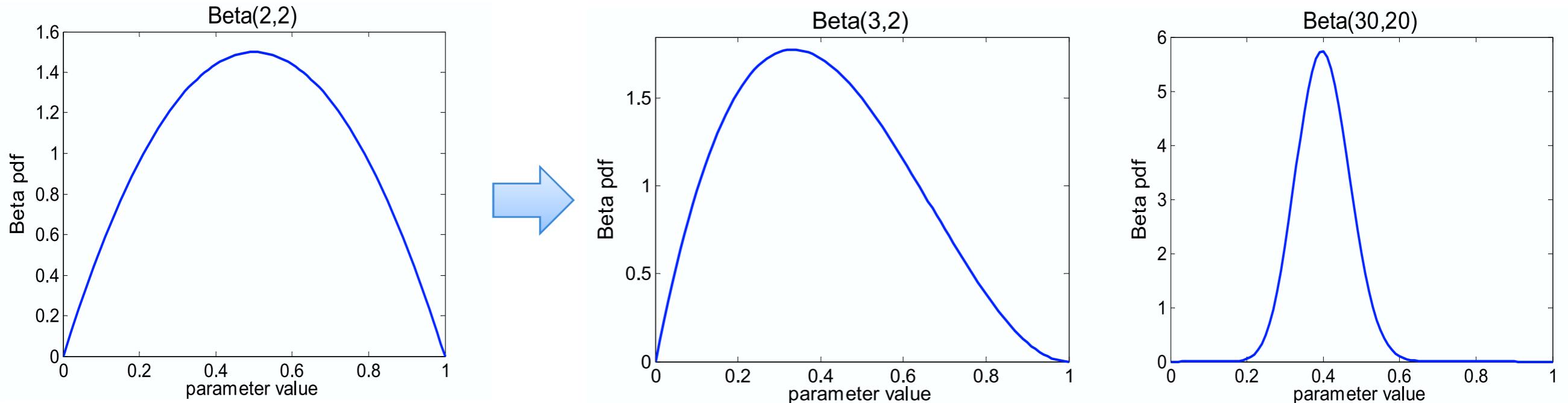


More concentrated as values of α, β increase

Beta conjugate prior

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

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



As $n = \alpha_H + \alpha_T$ increases

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



Han Solo and Bayesian Priors

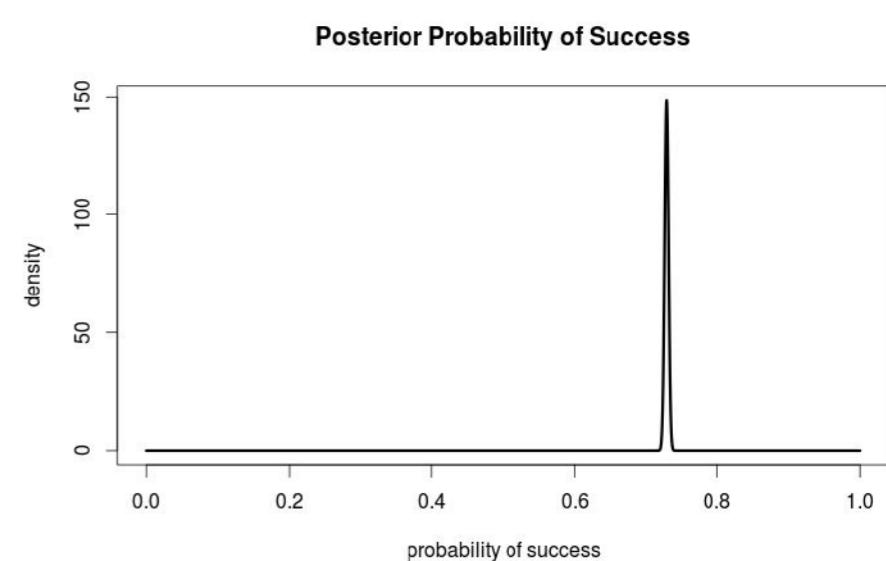
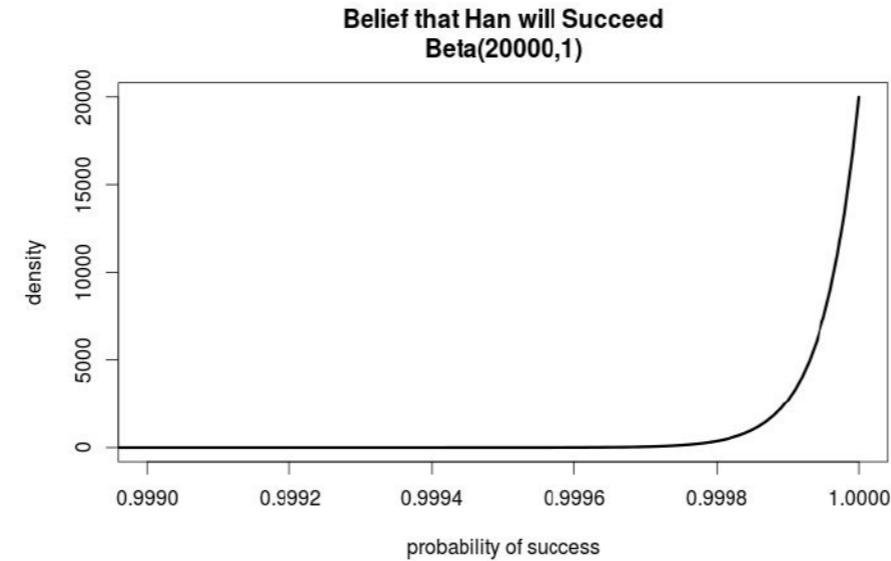
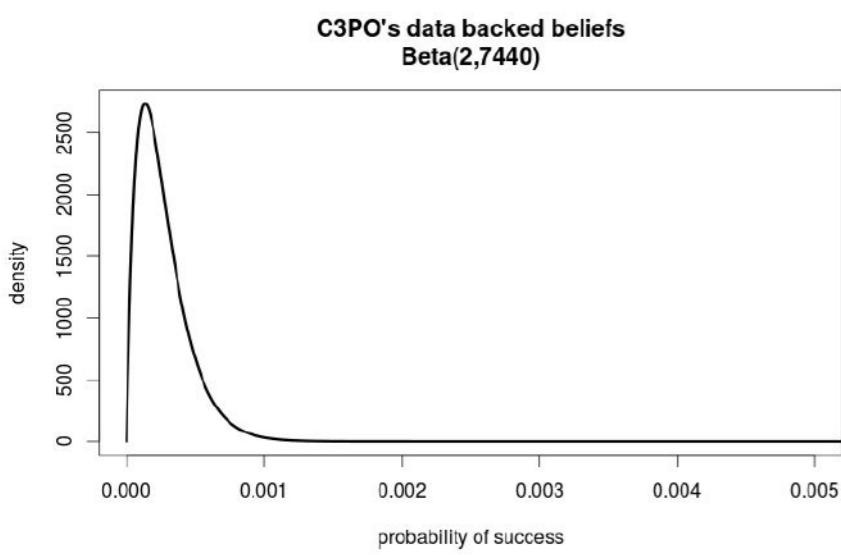


C3PO: Sir, the possibility of successfully navigating an asteroid field is approximately 3,720 to 1!

Han: Never tell me the odds!

$$P(\theta) \sim Beta(\beta_H, \beta_T)$$

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



MLE vs. MAP

- Maximum Likelihood estimation (MLE)

Choose value that maximizes the probability of observed data

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

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

Choose value that maximizes the probability of observed data

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

- Maximum *a posteriori* (MAP) estimation

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

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

When is MAP same as MLE?

From Binomial to Multinomial

Example: Dice roll problem (6 outcomes instead of 2)

Likelihood is $\sim \text{Multinomial}(\theta = \{\theta_1, \theta_2, \dots, \theta_k\})$

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

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^k \theta_i^{\beta_i - 1}}{B(\beta_1, \dots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \dots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \dots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

http://en.wikipedia.org/wiki/Dirichlet_distribution



Bayesians vs. Frequentists

You are no
good when
sample is
small



You give a
different
answer for
different
priors

Application of Bayes Rule

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

$$\begin{aligned} P(a = 1|t = 1) &= \frac{P(t = 1|a = 1)P(a = 1)}{P(t = 1)} \\ &= \frac{P(t = 1|a = 1)P(a = 1)}{P(t = 1|a = 1)P(a = 1) + P(t = 1|a = 0)P(a = 0)} \\ &= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091 \end{aligned}$$

Only 9%!

Improving the diagnosis

Use a weaker follow-up test!

- Approximately 0.1% are infected
- Test 2 reports positive for 90% infections
- Test 2 reports positive for 5% healthy people

$$\begin{aligned} P(a = 0|t_1 = 1, t_2 = 1) &= \frac{P(t_1 = 1, t_2 = 1|a = 0)P(a = 0)}{P(t_1 = 1, t_2 = 1|a = 1)P(a = 1) + P(t_1 = 1, t_2 = 1|a = 0)P(a = 0)} \\ &= \frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357 \end{aligned}$$

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

64%!...

AIDS test (Bayes rule)

Why can't we use Test 1 twice?

- Outcomes are **not** independent,
- but tests 1 and 2 **conditionally independent (by assumption):**

$$p(t_1, t_2 | a) = p(t_1 | a) \cdot p(t_2 | a)$$

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=WCbdZ5sXac25dpH02XcRyD0dts993hKwsAVXpGrFh0w=;
b=WK2B2+ExWnf/gvTkw6uUvKuP4XeoKn1Jq3USYTm0RARK8dSFjy0QsIHeAP9Yssxp60
7ngGoTzYqd+ZsyJfvQcLAWp1PCJhG8AMcnqWkx0NMeoFvIp2HQooZwxSOCx5ZRgY+7qX
uIbbdn4lUDXj6UFe16SpLDCKptd80Z3gr7+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: 6bw16D17HjZIkx0Eo138NZzyeHs
Message-ID: <CAFJJHDGPBW+SdZg0MdAABiAKydDk9tpeMoDijYGjoGO-WC7osg@mail.gmail.com>
Subject: CS 281B. Advanced Topics in Learning and Decision Making

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:

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

Naïve Bayes Assumption, Example

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $\mathbf{X} = (X_1 \quad X_2 \quad X_3 \quad \dots \quad \dots \quad X_d)$ Y

Sky	Temp	Humid	Wind	Water	Forecast	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

n rows



Naïve Bayes Assumption, Example

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $\mathbf{X} = (X_1 \quad X_2 \quad X_3 \quad \dots \quad \dots \quad X_d) \quad Y$

Sky	Temp	Humid	Wind	Water	Forecast	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

Naïve Bayes assumption:

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

Naïve Bayes Assumption, Example

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $\mathbf{X} = (X_1 \quad X_2 \quad X_3 \quad \dots \quad \dots \quad X_d) \quad Y$

n rows

Sky	Temp	Humid	Wind	Water	Forecast	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

Naïve Bayes assumption:

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

How many parameters to estimate?

(X is composed of d binary features,
 Y has K possible class labels)

Naïve Bayes Assumption, Example

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $\mathbf{X} = (X_1 \quad X_2 \quad X_3 \quad \dots \quad \dots \quad X_d) \quad Y$

Sky	Temp	Humid	Wind	Water	Forecast	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

Naïve Bayes assumption:

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

How many parameters to estimate?

(X is composed of d binary features,
 Y has K possible class labels)

($2^d - 1$) K vs $(2-1)dK$

Naïve Bayes Classifier

Given:

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

Naïve Bayes Decision rule:

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

Naïve Bayes Algorithm for discrete features

Training data: $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$

$$X^{(j)} = (X_1^{(j)}, \dots, X_d^{(j)})$$

n d -dimensional discrete features + K class labels

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

We need to estimate these probabilities!



Estimate them with MLE (Relative Frequencies)!

Naïve Bayes Algorithm for discrete features

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

We need to estimate these probabilities!

Estimators

For Class Prior

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

For Likelihood

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

NB Prediction for test data:

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

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

Subtlety: Insufficient training data

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

For example,

there is no $X_1 = \text{'Earn'}$ when $Y = \text{'SpamEmail'}$ in our dataset.

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

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

Thus, no matter what the values X_2, \dots, X_d take:

$$P(Y = b | X_1 = a, X_2, \dots, X_d) = 0$$

What now???

Naïve Bayes Alg – Discrete features

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

Use your expert knowledge & apply prior distributions:

- Add m “virtual” examples
- Same as assuming conjugate priors

Assume priors: $Q(Y = b) \quad Q(X_i = a, Y = b)$

MAP Estimate:

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


virtual examples
with $Y = b$

called Laplace smoothing

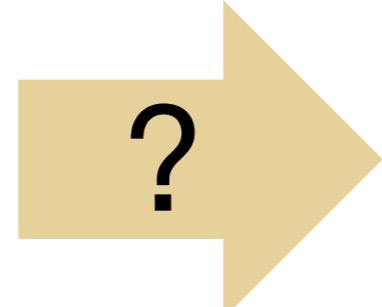
Case Study: Text Classification

Positive or negative movie review?

-  • unbelievably disappointing
-  • Full of zany characters and richly applied satire, and some great plot twists
-  • this is the greatest screwball comedy ever filmed
-  • It was pathetic. The worst part about it was the boxing scenes.

What is the subject of this article?

MEDLINE Article



MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- ...

Text Classification: definition

- Input:
 - a document d
 - a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$
- Output: a predicted class $c \in C$

Hand-coded rules

- Rules based on combinations of words or other features
 - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
 - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Text Classification and Naive Bayes

- Classify emails
 - $Y = \{\text{Spam}, \text{NotSpam}\}$
- Classify news articles
 - $Y = \{\text{what is the topic of the article?}\}$

What are the features X ?

The text!

Let X_i represent i^{th} word in the document

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

A problem: The support of $P(\mathbf{X}|\mathbf{Y})$ is huge!

- Article at least 1000 words, $\mathbf{X}=\{X_1, \dots, X_{1000}\}$
- X_i represents i^{th} word in document, i.e., the domain of X_i is the entire vocabulary, e.g., Webster Dictionary (or more).



$$X_i \in \{1, \dots, 50000\} \Rightarrow K(1000^{50000} - 1)$$

parameters to estimate without the NB assumption....

$$h_{MAP}(\mathbf{x}) = \arg \max_{1 \leq k \leq K} P(Y = k) P(X_1 = x_1, \dots, X_{1000} = x_{1000} | Y = k)$$

NB for Text Classification

$X_i \in \{1, \dots, 50000\} \Rightarrow K(1000^{50000} - 1)$ parameters to estimate....

NB assumption helps a lot!!!

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

$\Rightarrow 1000K(50000-1)$ parameters to estimate with NB assumption

NB assumption helps, but still lots of parameters to estimate.

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

Bag of words model

Typical additional assumption:

Position in document doesn't matter:

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

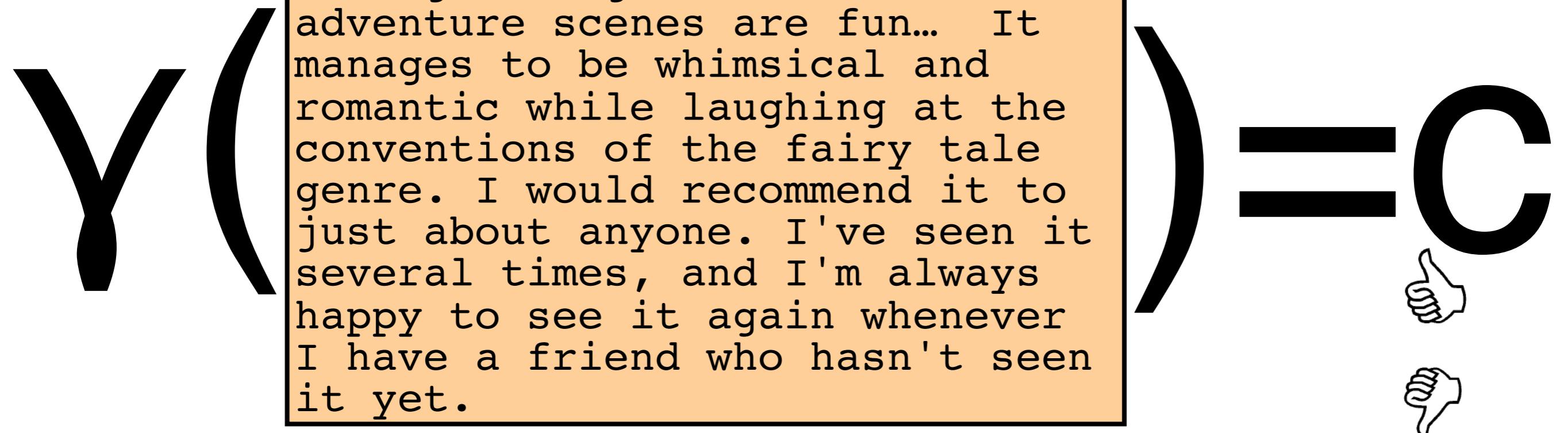
- “Bag of words” model – order of words on the page ignored
The document is just a bag of words: i.i.d. words
- Sounds really silly, but often works very well!

⇒ $K(50000-1)$ parameters to estimate

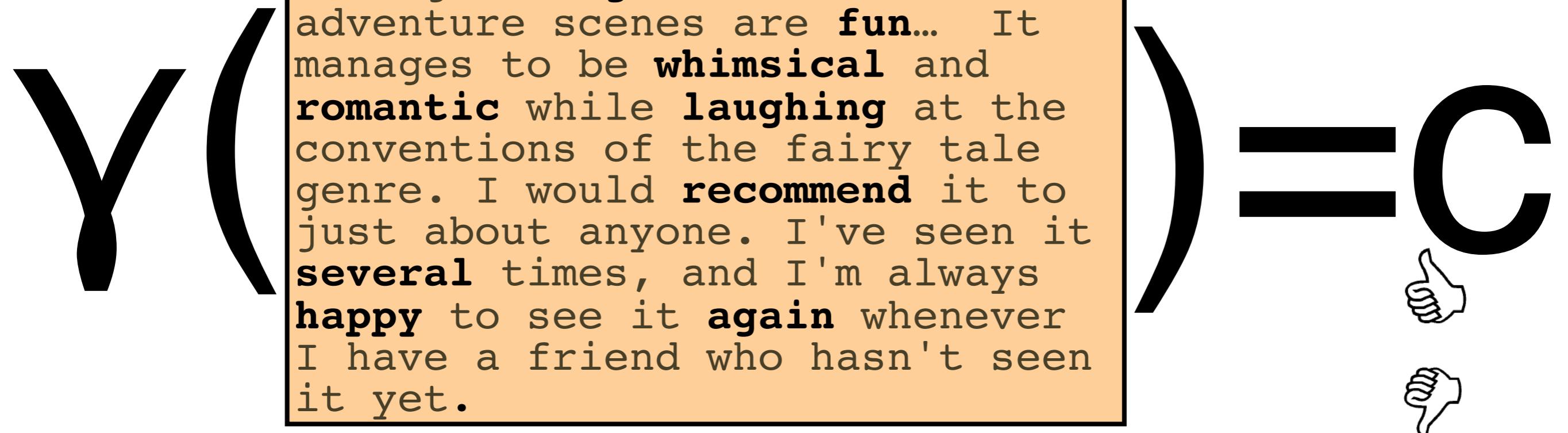
The probability of a document with words x_1, x_2, \dots

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

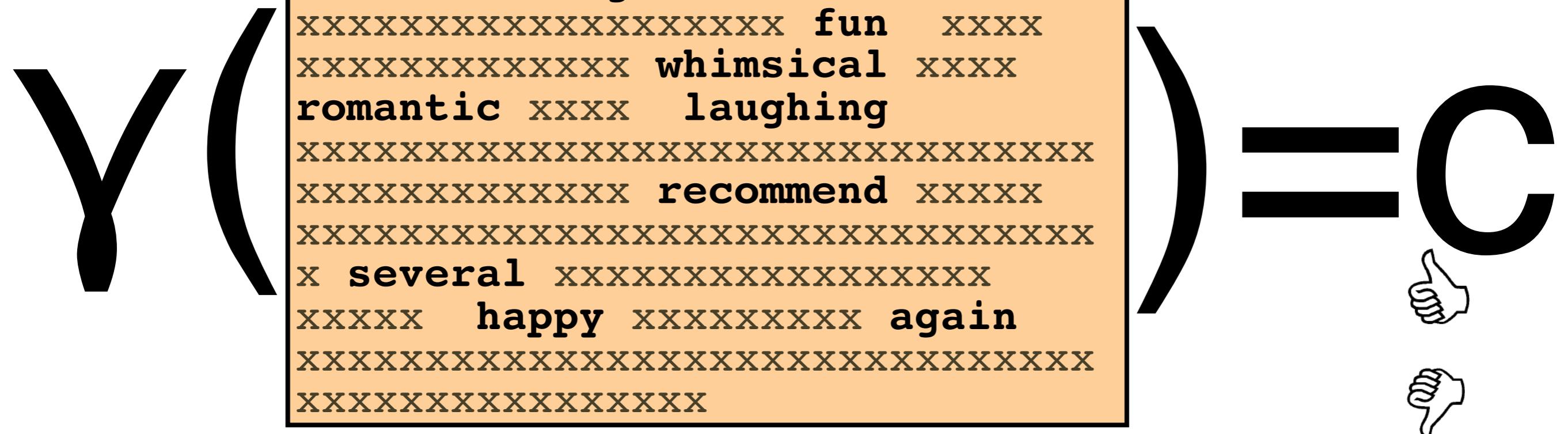
The bag of words representation



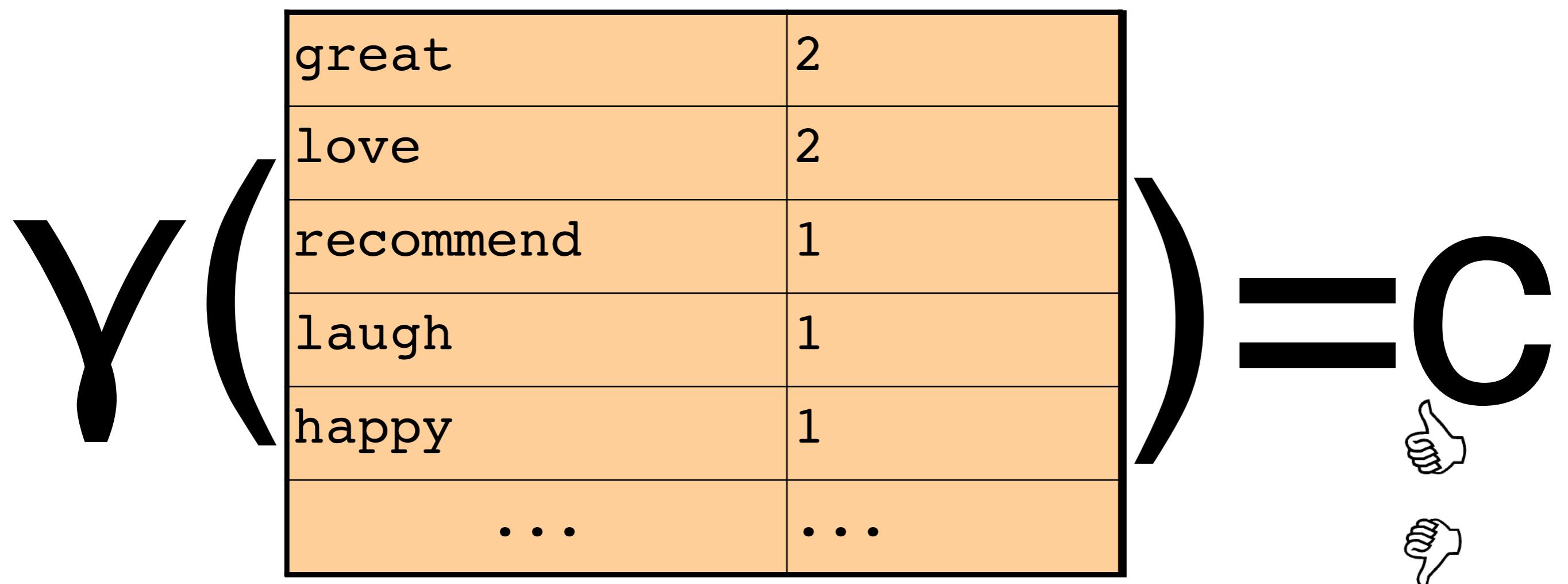
The bag of words representation



The bag of words representation: using a subset of words



The bag of words representation



$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w|c) = \frac{\text{count}(w,c) + 1}{\text{count}(c) + |V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w|c) = \frac{\text{count}(w,c) + 1}{\text{count}(c) + |V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$$P(c) = \frac{3}{4}$$

$$P(j) = \frac{1}{4}$$

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w|c) = \frac{\text{count}(w,c)+1}{\text{count}(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$$P(c) = \frac{3}{4}$$

$$P(j) = \frac{1}{4}$$

Conditional Probabilities:

$$P(\text{Chinese}|c) = (5+1) / (8+6) = 6/14 = 3/7$$

$$P(\text{Tokyo}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Japan}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Chinese}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Tokyo}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Japan}|j) = (1+1) / (3+6) = 2/9$$

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w|c) = \frac{\text{count}(w,c)+1}{\text{count}(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$$P(c) = \frac{3}{4}$$

$$P(j) = \frac{1}{4}$$

Conditional Probabilities:

$$P(\text{Chinese}|c) = (5+1) / (8+6) = 6/14 = 3/7$$

$$P(\text{Tokyo}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Japan}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Chinese}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Tokyo}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Japan}|j) = (1+1) / (3+6) = 2/9$$

Choosing a class:

$$P(\text{cl}d_5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \\ \approx 0.0003$$

$$P(\text{jl}d_5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \\ \approx 0.0001$$

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?

e.g., 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

$$\begin{aligned} h_{NB}(x) &= \arg \max_y P(y) \prod_i P(X_i = x_i | y) \\ &\approx \arg \max_k \hat{P}(Y = k) \prod_i \mathcal{N}(\hat{\mu}_{ik}, \hat{\sigma}_{ik}) \end{aligned}$$

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

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

Estimating parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

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

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

**ith pixel in
jth training image** ←

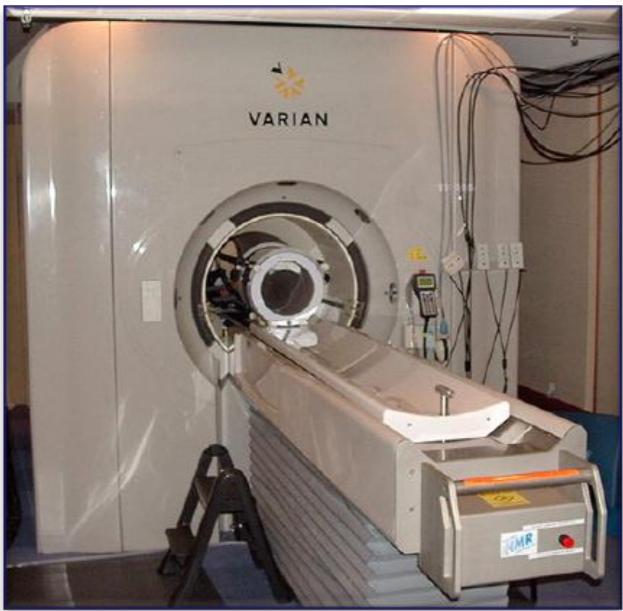
→ **kth class**
→ **jth training image**

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

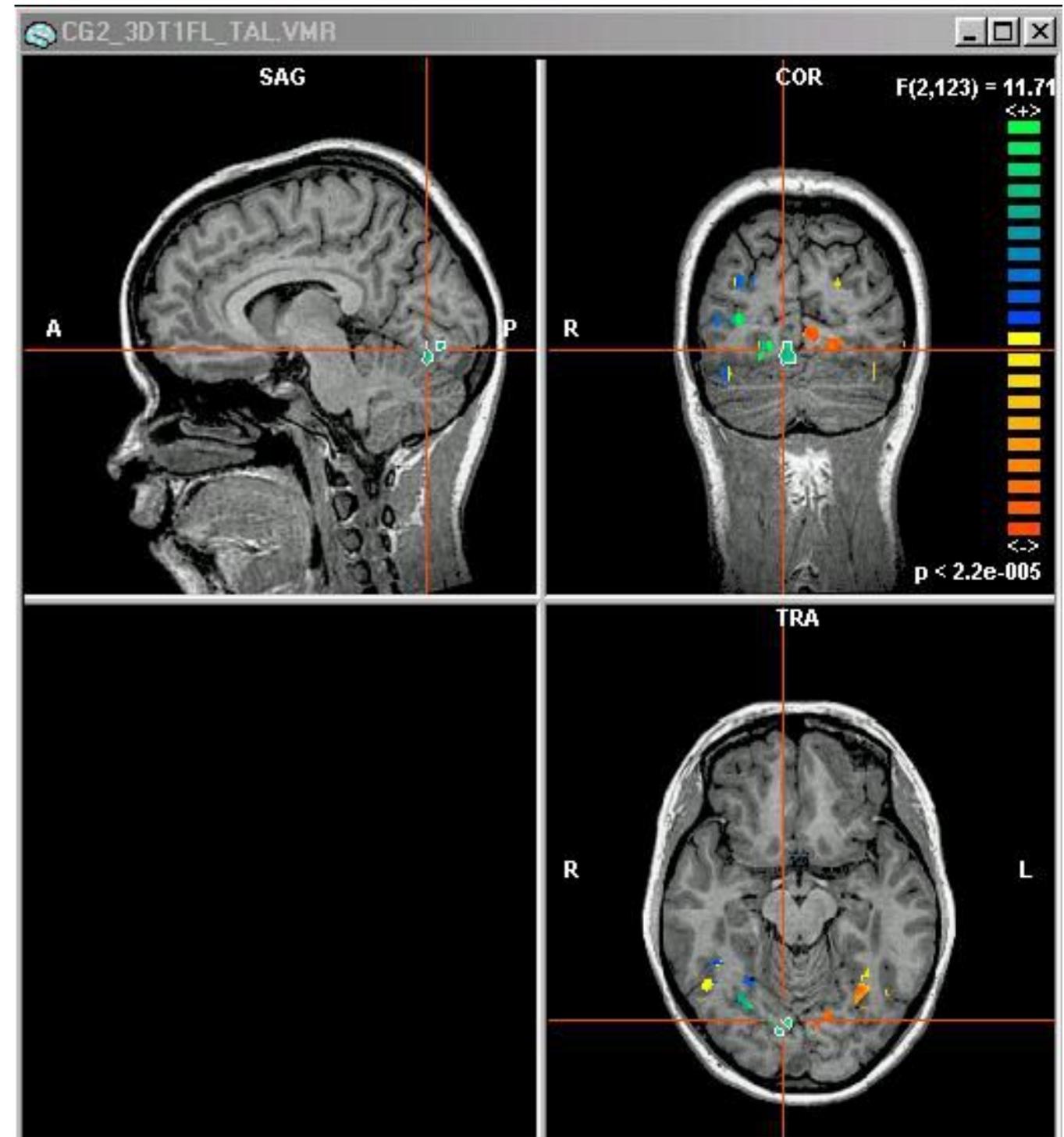
$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

Case Study: Classifying Mental States

Example: GNB for classifying mental states



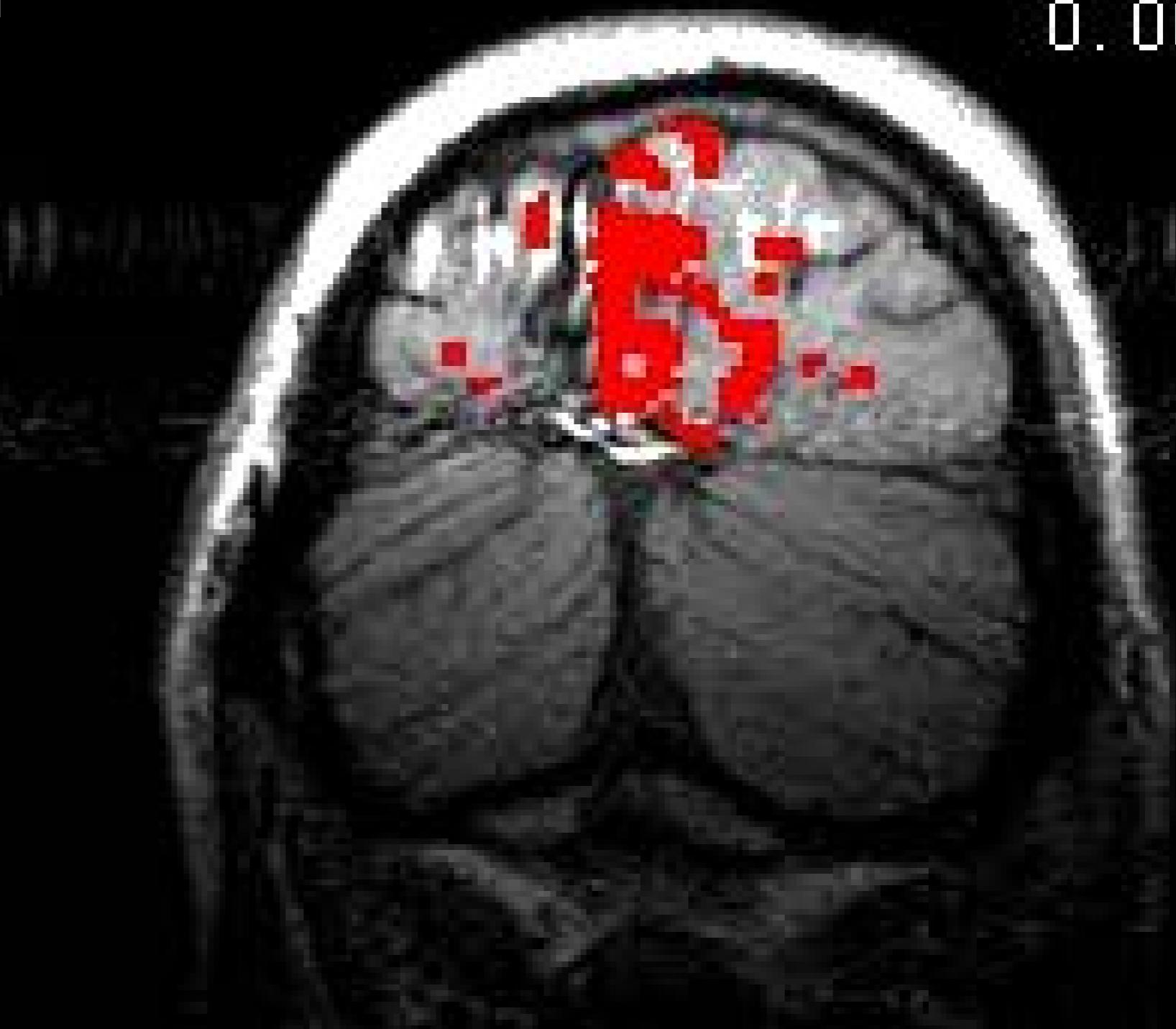
~1 mm resolution
~2 images per sec.
15,000 voxels/image
non-invasive, safe
measures Blood Oxygen Level Dependent (BOLD) response



[Mitchell et al.]



Brain scans can
track activation
with precision and
sensitivity

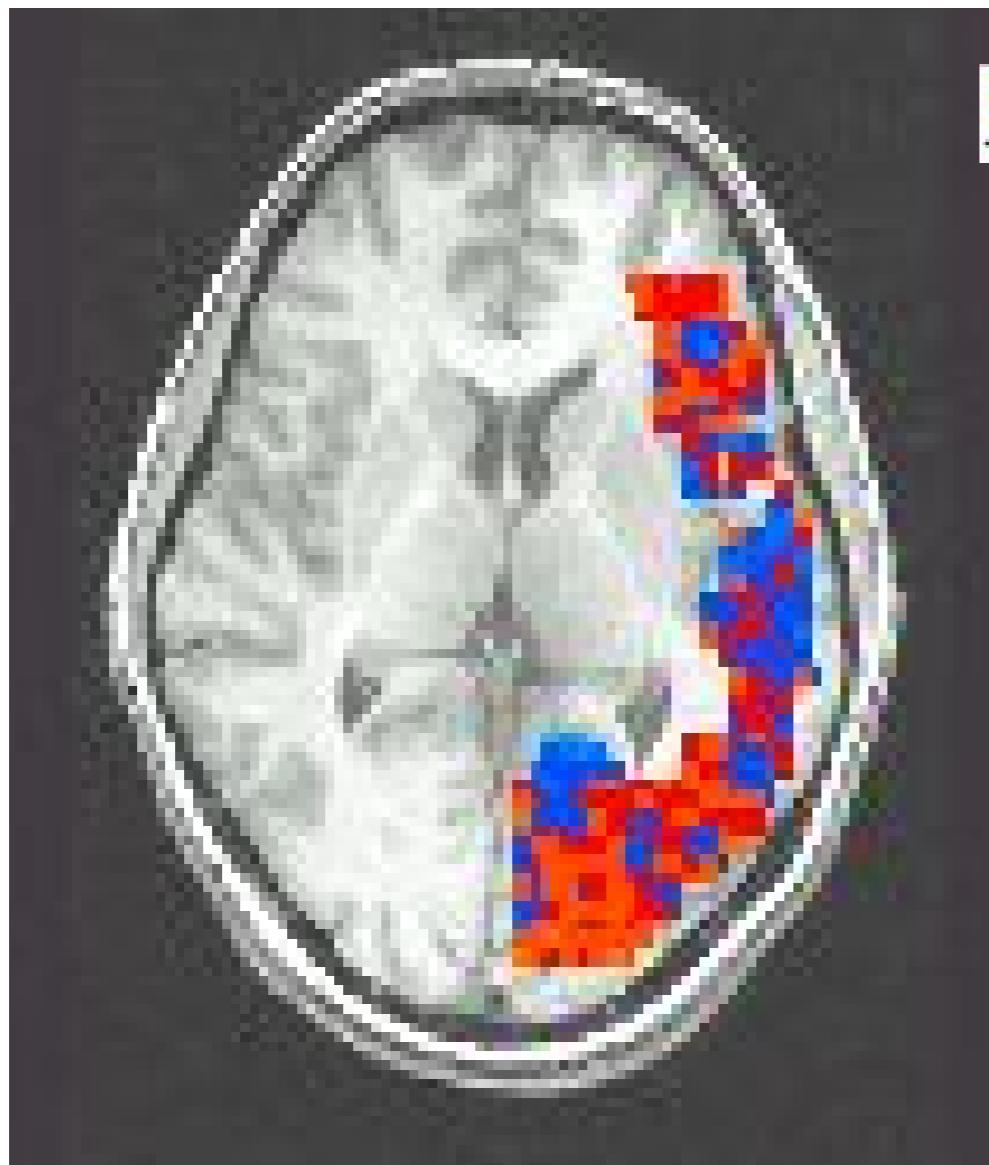


Learned Naïve Bayes Models

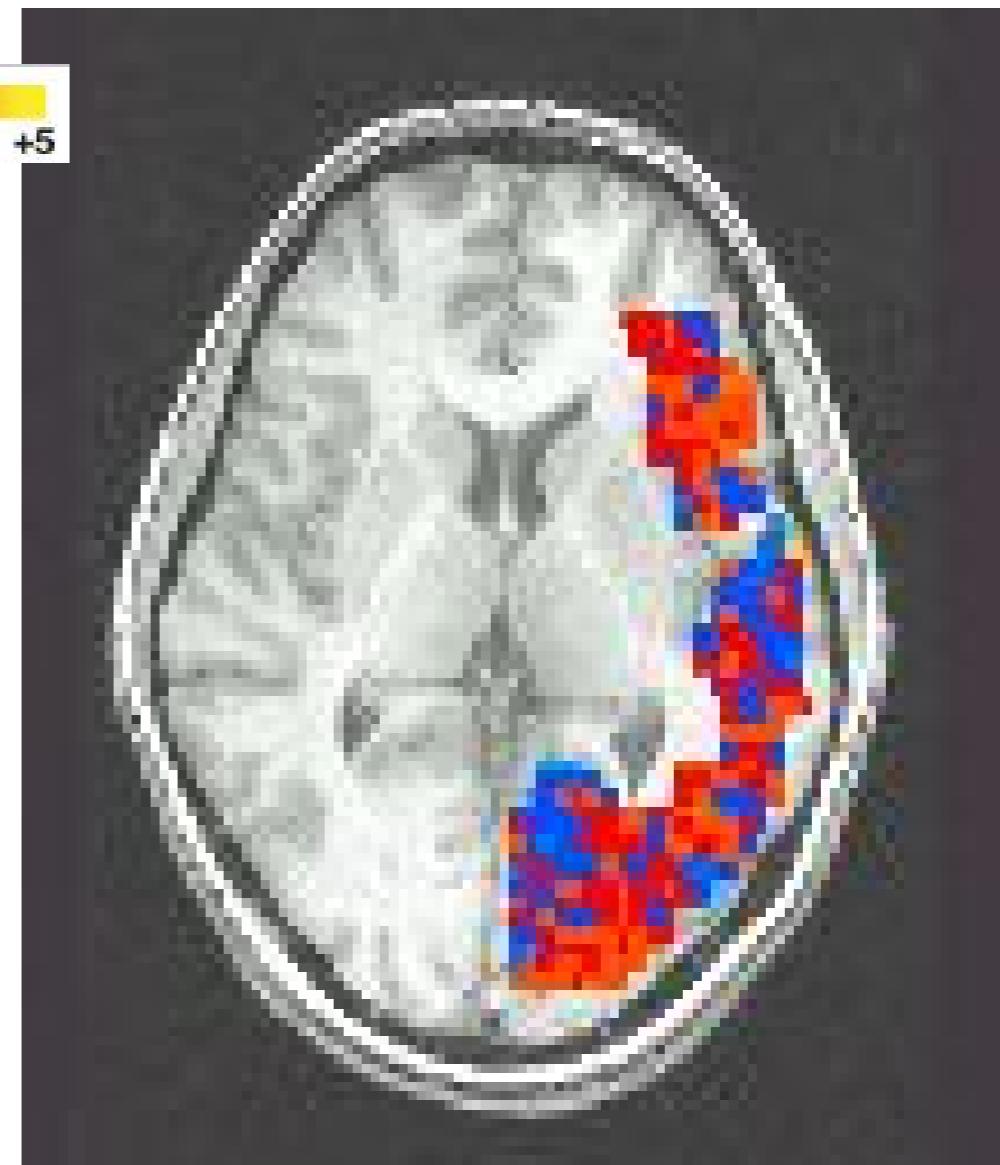
- Means for $P(\text{BrainActivity} \mid \text{WordCategory})$

Pairwise classification accuracy: [Mitchell et al.]
78-99%, 12 participants

Tool words



Building



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