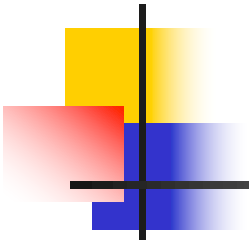
CSCI 5090/7090- Machine Learning

Spring 2018

Mehdi Allahyari
Georgia Southern University

Bayes Classifier

(slides borrowed from Tom Mitchell, Barnabás Póczos & Aarti Singh

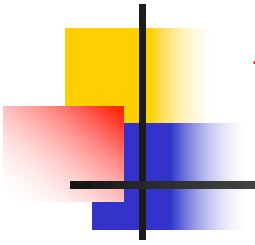


Joint Distribution:

sounds like the solution to
learning F: X → Y,
or P(Y | X).

Main problem: learning P(Y|X)
can require more data than we have

consider learning Joint Dist. with 100 attributes
# of rows in this table?
# of people on earth?
fraction of rows with 0 training examples?



What to do?

1. Be smart about how we estimate probabilities from sparse data
   - maximum likelihood estimates
   - maximum a posteriori estimates

2. Be smart about how to represent joint distributions
   - Bayes networks, graphical models

1. Be smart about how we estimate probabilities

Estimating Probability of Heads

X=1 X=0

- I show you the above coin X, and hire you to estimate the probability that it will turn up heads ($X = 1$) or tails ($X = 0$)
- You flip it repeatedly, observing
  - it turns up heads α_1 times
  - it turns up tails α_0 times
- Your estimate for $P(X = 1)$ is....?

Estimating θ = P(X=1)

X=1 X=0

Test A:

100 flips: 51 Heads (X=1), 49 Tails (X=0)

Test B:

3 flips: 2 Heads (X=1), 1 Tails (X=0)

Estimating θ = P(X=1)

X=1 X=0

Case C: (online learning)

- keep flipping, want single learning algorithm that gives reasonable estimate after each flip

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize **P(data | θ)**
- e.g., $$\hat{\theta}^{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Principle 2 (maximum a posteriori prob.):

- choose parameters θ that maximize **P(θ | data)**
- e.g.

$$\hat{\theta}^{MAP} = \frac{\alpha_1 + \#\text{hallucinated\_1s}}{(\alpha_1 + \#\text{hallucinated\_1s}) + (\alpha_0 + \#\text{hallucinated\_0s})}$$

Summary:
Maximum Likelihood Estimate

X=1 X=0

P(X=1) = θ
P(X=0) = 1-θ
(Bernoulli)

- Each flip yields boolean value for X

$$X \sim \text{Bernoulli: } P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros (Binomial)

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \text{argmax}_\theta \, P(D|\theta) = \frac{\alpha_1}{\alpha_1+\alpha_0}$$

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize P(data | θ)

Principle 2 (maximum a posteriori prob.):

- choose parameters θ that maximize

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)\, P(\theta)}{P(\text{data})}$$

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

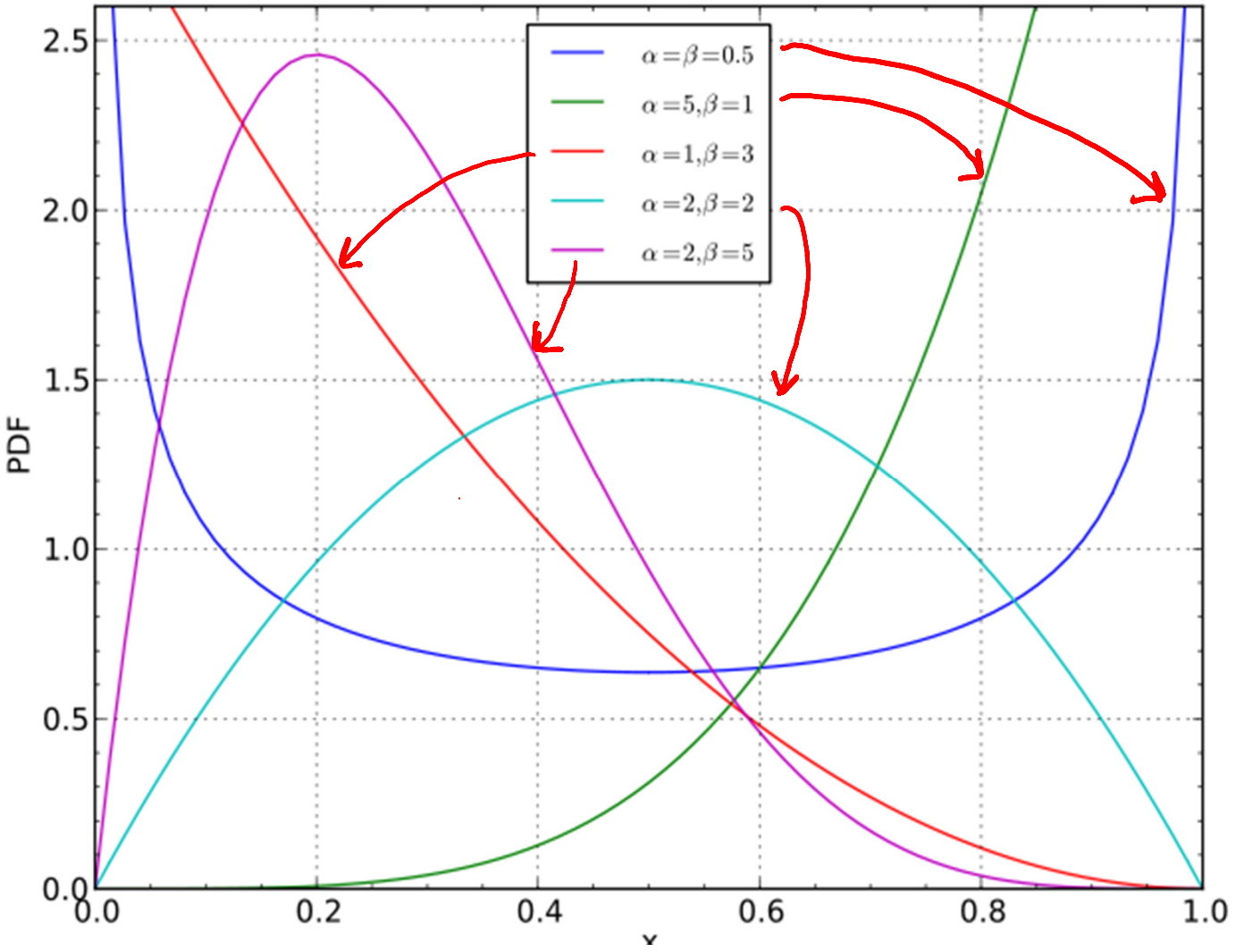
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

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

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

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

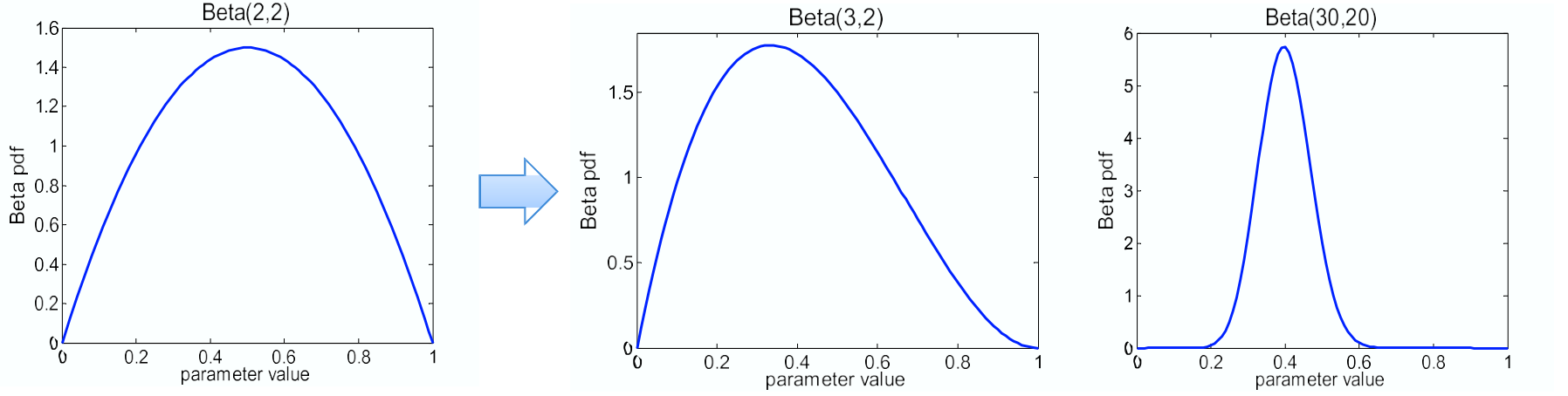
Beta distribution



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"

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

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,

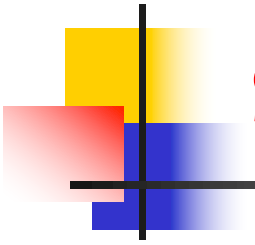
$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

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.

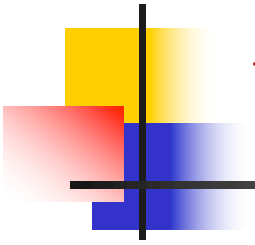
http://en.wikipedia.org/wiki/Dirichlet_distribution



Some terminology

- Likelihood function: $P(\text{data} \mid \theta)$
- Prior: $P(\theta)$
- Posterior: $P(\theta \mid \text{data})$

- Conjugate prior: $P(\theta)$ is the conjugate prior for likelihood function $P(\text{data} \mid \theta)$ if the forms of $P(\theta)$ and $P(\theta \mid \text{data})$ are the same.



You should know

- Probability basics
  - random variables, events, sample space, conditional probs, …
  - independence of random variables
  - Bayes rule
  - Joint probability distributions
  - calculating probabilities from the joint distribution
- Point estimation
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – binomial, Beta, Dirichlet, …

Chain Rule & Bayes Rule

Chain rule:

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

Bayes rule:

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

Which is shorthand for:

$$(\forall i,j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

Equivalently:

$$(\forall i,j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}$$

Bayesian Learning

$\mathcal{D}$ is the measured data.

Our goal is to estimate parameter θ.

- Use Bayes rule:

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

- Or equivalently:

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

posterior likelihood prior

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:

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

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

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

Only 9%!...

Improving the diagnosis

Use a weaker follow-up test!

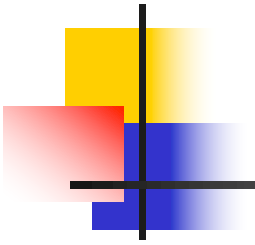
- Approximately 0.1% are infected
- Test 2 reports positive for 90% infections
- Test 2 reports positive for 5% healthy people

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

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

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

64%!...



Improving the diagnosis

Why can't we use Test 1 twice?

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

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

The Naïve Bayes Classifier

24

How many parameters must we estimate?

Suppose $X = \langle X_1, \ldots X_n \rangle$
where X_i and Y are boolean RV's

Y

d 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

To estimate $P(Y| X_1, X_2, \ldots X_n)$ **$(2^n-1)\ 2$**

If we have 30 X_i's instead of 2: $P(Y| X_1, X_2, \ldots X_{30})$ **$2^{30} \cong$ 1 Billion**

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:

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

Conditional Independence

Definition: 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

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

Which we often write

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

E.g.,

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

Naïve Bayes Assumption

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y.

Given this assumption, then:

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

in general: $P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$

How many parameters to describe $P(X_1 \ldots X_n|Y)$? $P(Y)$?

Without conditional indep assumption? $2(2^n - 1) + 1$

With conditional indep assumption? $2n + 1$

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \dots X_n | Y = y_j)}$$

Assuming conditional independence among X_i's:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, classification rule for $X^{new} = < X_1, \dots, X_n >$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples) for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 of these...

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates (MLE's): **(Relative Frequencies)**

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in dataset D for which $Y=y_k$

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.

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

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

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??? What can be done to avoid this?

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\hat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

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

Conjugate priors

- P(θ) and P(θ|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

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

For Binomial, conjugate prior is Beta distribution.

[A. Singh]

Conjugate priors

- P(θ) and P(θ|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.

Estimating Parameters: Y, X_i discrete-valued

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

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

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + \alpha_k}{|D| + \sum_m \alpha_m}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + \alpha'_k}{\#D\{Y = y_k\} + \sum_m \alpha'_m}$$

Only difference: "imaginary" examples

Case Study: Text Classification

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

What are the features **X**?

The text!

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

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
uIbbdna4IUDXj6UFe16SpLDCkptd8OZ3gr7+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

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(**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 the entire vocabulary, e.g., Webster Dictionary (or more).

X_i 2 {1,…,50000}) K(1000^{50000} -1) parameters to estimate without the NB assumption….

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

NB for Text Classification

X_i 2 {1,…,50000}) 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

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

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

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

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 $x_1,x_2,\ldots$

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

Bag of words model

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

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

Bag of words approach

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

Learning to classify document: P(Y | X) the "Bag of Words" model

- Y discrete valued. e.g., Spam or not
- $X = \langle X_1, X_2, \ldots X_n \rangle$ = document
- X_i is a random variable describing the word at position i in the document
- possible values for Xi : any word wk in English
- Document = bag of words: the vector of counts for all w_k's
- This vector of counts follows a ?? Distribution

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes

  (examples) for each value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

  prob that word x_{ij} appears in position i, given $Y=y_k$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* Additional assumption: word probabilities are position independent

$$\theta_{ijk} = \theta_{mjk} \quad \text{for} \quad i \neq m$$

MAP estimates for bag of words

Map estimate for multinomial

$$\theta_i = \frac{\alpha_i + \beta_i - 1}{\sum_{m=1}^{k} \alpha_m + \sum_{m=1}^{k} (\beta_m - 1)}$$

$$\theta_{aardvark} = P(X_i = \text{aardvark}) = \frac{\#\text{ observed 'aardvark'} + \#\text{ hallucinated 'aardvark'} - 1}{\#\text{ observed words} + \#\text{ hallucinated words} - k}$$

What β's should we choose?

Twenty news groups results

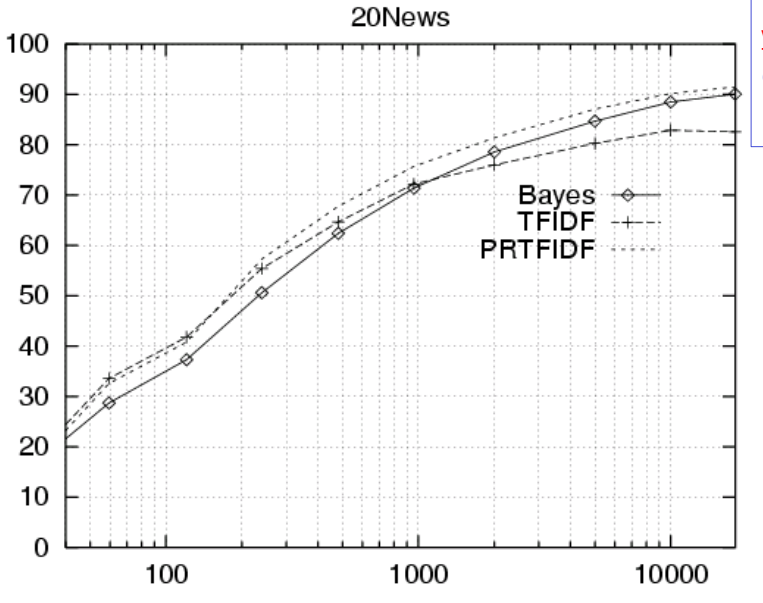
Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

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

Twenty news groups results

Learning Curve for 20 Newsgroups



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

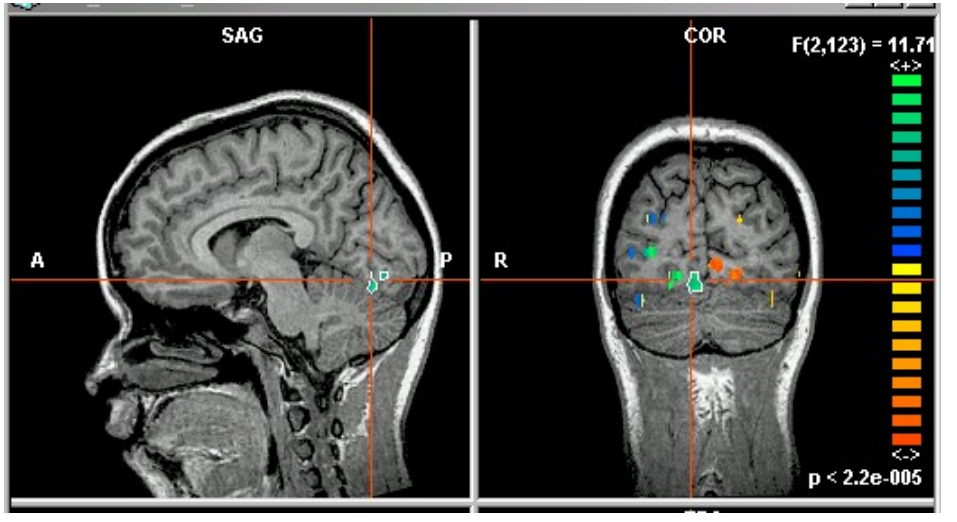
For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"

What if we have continuous X_i?

Eg., image classification: X_i is ith pixel

What if we have continuous X_i?

image classification: X_i is i[th] pixel, Y = mental state



Still have:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Just need to decide how to represent $P(X_i \mid Y)$

What if features are continuous?

Eg., image classification: X_i is ith pixel

Gaussian Naïve Bayes (GNB): assume

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

Sometimes assume σ_{ik}

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

Gaussian Naïve Bayes Algorithm – continuous X_i

(but still discrete Y)

- Train Naïve Bayes

  (examples) for each value y_k

  estimate* $\pi_k \equiv P(Y = y_k)$

  for each attribute X_i estimate

  class conditional mean μ_{ik} , variance σ_{ik}

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating parameters: Y discrete, Xi continuous

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_i P(X_i = x_i | y)$$

$$\approx \arg\max_k \hat{P}(Y = k) \prod_i \mathcal{N}(\hat{\mu}_{ik}, \hat{\sigma}_{ik})$$

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

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

Estimating parameters: Y discrete, Xi continuous

Maximum likelihood estimates:

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

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

k^{th} class

j^{th} training image

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

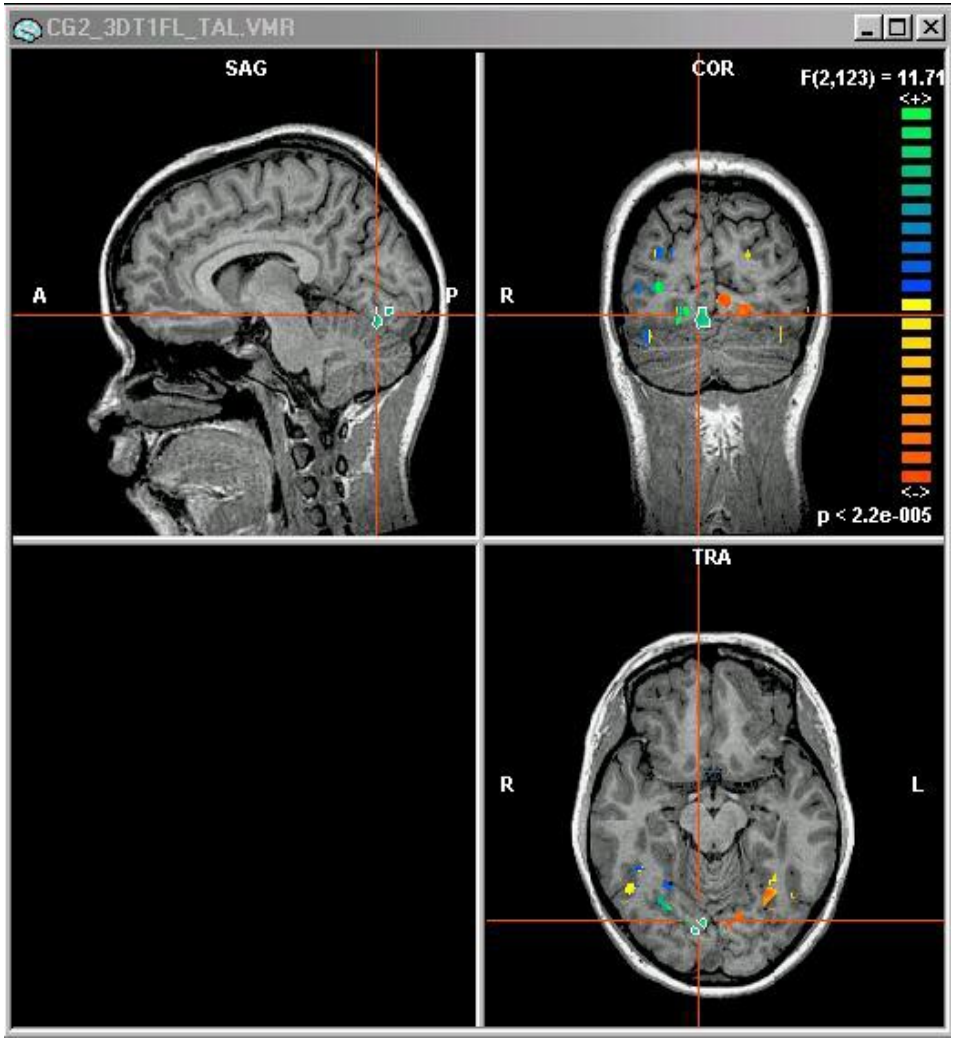
δ(z)=1 if z true, else 0

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

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

Example: GNB for classifying mental states

Classify a person's cognitive activity, based on brain image



~1 mm resolution

~2 images per sec.

15,000 voxels/image

non-invasive, safe

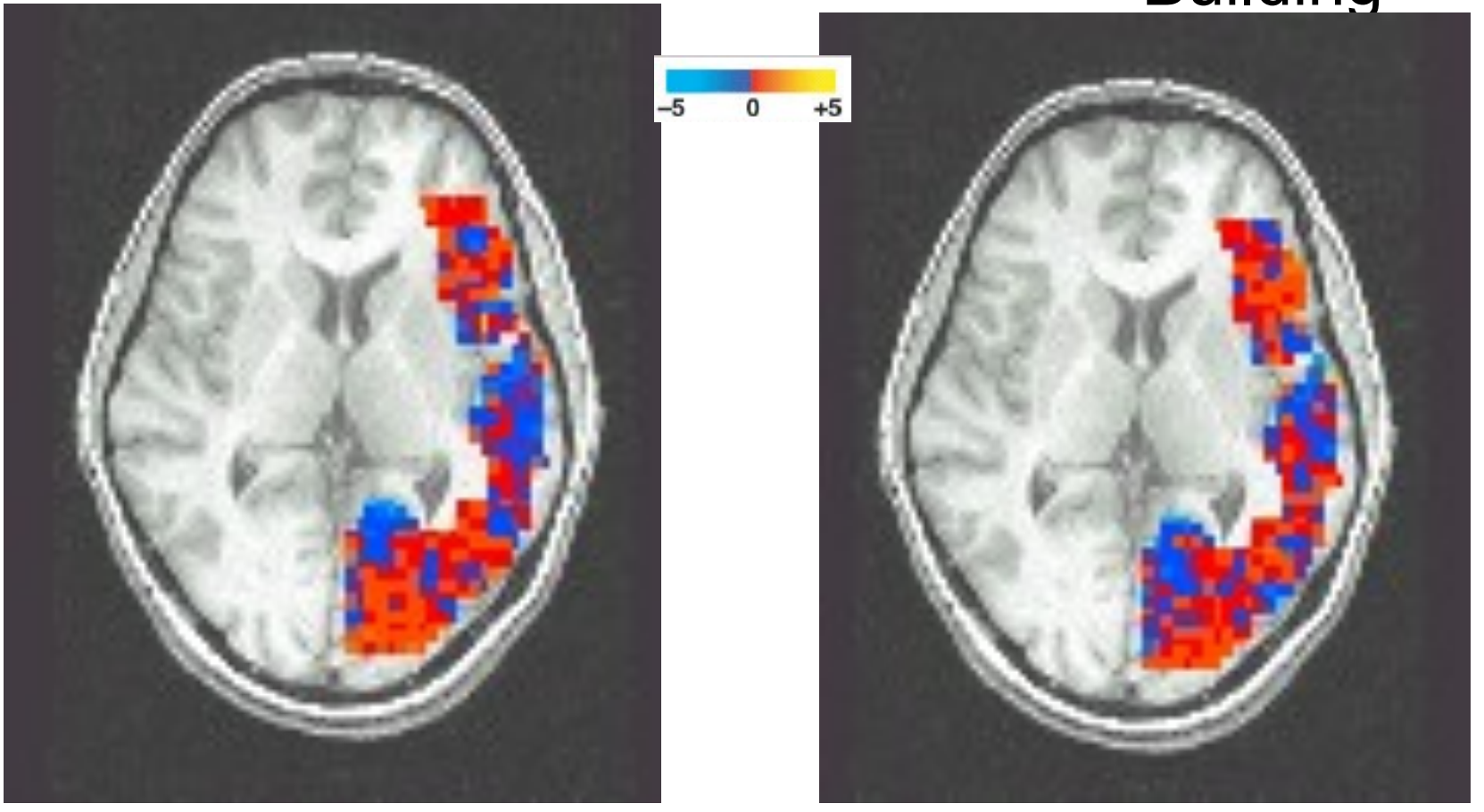
measures Blood Oxygen Level Dependent (BOLD) response

[Mitchell et al.]

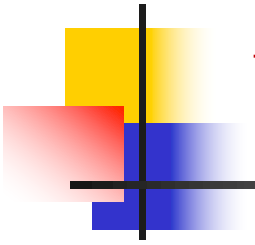
Learned Naïve Bayes Models – Means for P(BrainActivity | WordCategory)

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

Tool words



Building



What you should know…

- Training and using classifiers based on Bayes rule

- Conditional independence
  - What it is
  - Why it's important

- Naïve Bayes
  - What it is
  - Why we use it so much
  - Training using MLE, MAP estimates
  - Discrete variables and continuous (Gaussian)