Lecture 8:

- Naïve Bayes Classifier
Today

• Bayes rule
  - Naïve Bayes Classifier

• Application
  - Text classification
  - “Mind reading” = fMRI data processing
Recap: 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
  \[ \hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta|D) \]
  \[ = \arg \max_{\theta} P(D|\theta)P(\theta) \]
Recap: 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 $\theta$, we obtain a distribution over possible values of $\theta$.
Recap: Chain Rule & Bayes Rule

Chain rule:

\[ P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X) \]

Bayes rule:

\[ P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \]
Recap: Bayesian Learning

D is the measured data  
Our goal is to estimate parameter $\theta$

• Use Bayes rule:

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

• Or equivalently:

$$P(\theta \mid D) \propto P(D \mid \theta)P(\theta)$$

posterior  likelihood  prior
Recap: MAP estimation for Binomial distribution

In the 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: \[ P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T) \]

then the posterior is Beta distribution
Recap: Beta conjugate prior

\[ P(\theta) \sim \text{Beta}(\beta_H, \beta_T) \]

\[ P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

As the number of samples \( n = \alpha_H + \alpha_T \) increases,

As we get more samples, effect of prior is “washed out”
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%!...
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%!...
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
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:

$$P(X_1...X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$
Naïve Bayes Assumption, Example

**Task:** Predict whether or not a picnic spot is enjoyable

**Training Data:** \( X = (X_1, X_2, X_3, \ldots, X_d) \)

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

\( n \) rows
Naïve Bayes Assumption, Example

**Task:** Predict whether or not a picnic spot is enjoyable

**Training Data:**

\[ X = (X_1, X_2, X_3, \ldots, X_d) \]

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

\[ P(X_1 \ldots X_d | Y) = \prod_{i=1}^{d} P(X_i | Y) \]
Task: Predict whether or not a picnic spot is enjoyable

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

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Naïve Bayes assumption:

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

How many parameters to estimate?
(X is composed of $d$ binary features, $Y$ has $K$ possible class labels)

slide by Barnabás Póczos & Aarti Singh
Naïve Bayes Assumption, Example

**Task:** Predict whether or not a picnic spot is enjoyable

**Training Data:**

\[
X = (X_1, X_2, X_3, \ldots, X_d) \quad Y
\]

- **Sky**
- **Temp**
- **Humid**
- **Wind**
- **Water**
- **Forecast**
- **EnjoySpt**

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Naïve Bayes assumption:**

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

**How many parameters to estimate?**

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

\((2^d-1)K \text{ vs } (2-1)dK\)
Naïve Bayes Classifier

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

Naïve Bayes Decision rule:

$$f_{NB}(x) = \arg \max_y P(x_1, \ldots, x_d | y)P(y)$$

$$= \arg \max_y \prod_{i=1}^d P(x_i | y)P(y)$$
Naïve Bayes Algorithm for discrete features

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

\[ X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)}) \]

\[ n \ d \text{-dimensional discrete features} + K \text{ class labels} \]

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

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

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

**NB Prediction for test data:**

\[ X = (x_1, \ldots, x_d) \]

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

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

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

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

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

What now???
Naïve Bayes Alg — Discrete features

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

Use your expert knowledge & apply prior distributions:

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

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

MAP Estimate:

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

called Laplace smoothing
Case Study: Text Classification
Is this spam?

BBM 406 on Piazza

[Instr Note] A suggestion: The Cornell Note-taking System

To: Aykut Erdem,
Reply-To: reply@piazza.com

-- Reply directly to this email above this line to create a new follow up. Or Click here to view.--
Instructor Aykut Erdem posted a new Note. Your instructor selected to notify everyone in real time of this post, bypassing user email preferences.

A suggestion: The Cornell Note-taking System

Dear all,

To get the most out of the lectures, I suggest to you to use Cornell Notes during and after the lectures. There are many resources about this note-taking system on the web but here is one from the Cornell University:

http://lsc.cornell.edu/study-skills/cornell-%20note-taking-system/

Best,
Aykut

Search or link to this question with @10.

Sign up for more classes at http://piazza.com/hacettepe.edu.tr.

Tell a colleague about Piazza. It's free, after all.

Thanks,
The Piazza Team

--
Contact us at team@piazza.com
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 \( C = \{c_1, c_2, \ldots, c_J\} \)

• 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
  - \( Y = \{ \text{Spam, NotSpam} \} \)

• Classify news articles
  - \( Y = \{ \text{what is the topic of the article?} \} \)

What are the features \( X \)?

The text!

Let \( X_i \) represent \( i^{\text{th}} \) 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
A problem: The support of $P(X|Y)$ is huge!

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

$$X_i \in \{1,\ldots,50000\} \Rightarrow K(1000^{50000} - 1)$$
parameters to estimate without the NB assumption....

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

\[ X_i \in \{1, \ldots, 50000\} \Rightarrow K(1000^{50000} - 1) \text{ parameters to estimate} \ldots \]

**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 \( y \)

\[ \Rightarrow 1000K(50000 - 1) \text{ parameters to estimate with NB assumption} \]

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

\[
\hat{h}_{NB}(x) = \arg \max_y P(y) \ \prod_{i=1}^{LengthDoc} P(X_i = x_i | y)
\]
Bag of words model

Typical additional assumption:

**Position in document doesn’t matter:**

\[ P(X_i=x_i \mid Y=y) = P(X_k=x_i \mid Y=y) \]

- “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!

\[ \Rightarrow K(50000-1) \text{ parameters to estimate} \]

The probability of a document with words \( x_1, x_2, \ldots \)

\[
\frac{\text{Length}}{\text{Doc}} \prod_{i=1}^{\text{Length}} P(x_i \mid y) = \prod_{w=1}^{W} P(w \mid y)^{\text{count}_w}
\]
The bag of words representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.
The bag of words representation: using a subset of words

\[ \gamma(x) = C \]

- love
- sweet
- satirical
- great
- fun
- whimsical
- romantic
- laughing
- recommend
- several
- happy
- again
The bag of words representation

\[ \gamma(\cdot) = \mathcal{C} \]

- great: 2
- love: 2
- recommend: 1
- laugh: 1
- happy: 1
- ...: ...
\[
\hat{P}(c) = \frac{N_c}{N}
\]

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

<table>
<thead>
<tr>
<th>Doc</th>
<th>Words</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1  Chinese Beijing Chinese</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>2  Chinese Chinese Shanghai</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>3  Chinese Macao</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>4  Tokyo Japan Chinese</td>
<td>j</td>
</tr>
<tr>
<td>Test</td>
<td>5  Chinese Chinese Chinese Tokyo Japan</td>
<td>?</td>
</tr>
</tbody>
</table>
$\hat{P}(c) = \frac{N_c}{N}$

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

**Priors:**

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

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

<table>
<thead>
<tr>
<th>Doc</th>
<th>Words</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training 1</td>
<td>Chinese Beijing Chinese</td>
<td>c</td>
</tr>
<tr>
<td>2</td>
<td>Chinese Chinese Shanghai</td>
<td>c</td>
</tr>
<tr>
<td>3</td>
<td>Chinese Macao</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>Tokyo Japan Chinese</td>
<td>j</td>
</tr>
<tr>
<td>Test 5</td>
<td>Chinese Chinese Chinese Tokyo Japan</td>
<td>?</td>
</tr>
</tbody>
</table>
\[ \hat{P}(c) = \frac{N_c}{N} \]
\[ \hat{P}(w \mid c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|} \]

**Priors:**
\[ P(c) = \frac{3}{4} \]
\[ P(j) = \frac{1}{4} \]

**Conditional Probabilities:**
\[ P(\text{Chinese} \mid c) = \frac{(5+1)}{(8+6)} = \frac{6}{14} = \frac{3}{7} \]
\[ P(\text{Tokyo} \mid c) = \frac{(0+1)}{(8+6)} = \frac{1}{14} \]
\[ P(\text{Japan} \mid c) = \frac{(0+1)}{(8+6)} = \frac{1}{14} \]
\[ P(\text{Chinese} \mid j) = \frac{(1+1)}{(3+6)} = \frac{2}{9} \]
\[ P(\text{Tokyo} \mid j) = \frac{(1+1)}{(3+6)} = \frac{2}{9} \]
\[ P(\text{Japan} \mid j) = \frac{(1+1)}{(3+6)} = \frac{2}{9} \]
Choosing a class:

\[ P(c|d) = \frac{3}{4} \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003 \]

\[ P(j|d) = 1/4 \times (2/9)^3 \times 2/9 \times 2/9 \approx 0.0001 \]
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

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):

\[
P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik} \sqrt{2\pi}} \exp\left(-\frac{(x - \mu_{ik})^2}{2\sigma_{ik}^2}\right)
\]

Different mean and variance for each class \( k \) and each pixel \( i \).

Sometimes assume variance

- is independent of \( Y \) (i.e., \( \sigma_i \)),
- or independent of \( X_i \) (i.e., \( \sigma_k \))
- or both (i.e., \( \sigma \))
Estimating parameters: $Y$ discrete, $X_i$ continuous

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

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

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

\[
\hat{\sigma}_{unbiased}^2 = \frac{1}{N - 1} \sum_{j=1}^{N} (x_j - \hat{\mu})^2
\]
Estimating parameters: 
Y discrete, X \(_i \) continuous

Maximum likelihood estimates:

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

\[
\hat{\mu}_{ik} = \frac{1}{\sum_{j} \delta(Y^j = y_k)} \sum_{j} X^j_i \delta(Y^j = y_k)
\]

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

\[
\hat{\sigma}^2_{ik} = \frac{1}{\sum_{j} \delta(Y^j = y_k) - 1} \sum_{j}(X^j_i - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)
\]
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 $P$(BrainActivity $|$ WordCategory)

Pairwise classification accuracy: 78-99%, 12 participants

[Mitchell et al.]
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