Today:
• Bayes Classifiers
• Naïve Bayes
• Gaussian Naïve Bayes

Readings:
Mitchell:
“Naïve Bayes and Logistic Regression”
(available on class website)

Estimating Parameters

• Maximum Likelihood Estimate (MLE): choose \( \theta \) that maximizes probability of observed data \( \mathcal{D} \)

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

• Maximum a Posteriori (MAP) estimate: choose \( \theta \) 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

- \( P(\theta) \) and \( P(\theta | D) \) have the same form

**Eg. 1** Coin flip problem

Likelihood is \( \sim \) Binomial

\[
P(D | \theta) = \theta^\alpha T (1 - \theta)^{\alpha T}
\]

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

- $P(0)$ and $P(0 \mid D)$ have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is $\sim$ Multinomial($0 = \{0_1, 0_2, \ldots, 0_6\}$

$$P(D \mid \theta) = \theta_1^{0_1} \theta_2^{0_2} \ldots \theta_6^{0_6}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1-1} \theta_2^{\beta_2-1} \ldots \theta_6^{\beta_6-1}}{B(\beta_1, \beta_2, \ldots \beta_6)} \sim \text{Dirichlet}$$

Then posterior is Dirichlet distribution

$$P(\theta \mid D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_6 + \alpha_6)$$

For Multinomial, conjugate prior is Dirichlet distribution.

Let’s learn classifiers by learning $P(Y \mid X)$

Consider $Y=$ Wealth, $X=$<Gender, HoursWorked>

| gender | hours_worked | wealth | P(wealth | $G,H,W$) |
|--------|-------------|--------|------------|
| Female | v0:40.5-    | poor   | 0.253122   |
|        |             | rich   | 0.0245895  |
|        | v1:40.5+    | poor   | 0.0421768  |
|        |             | rich   | 0.0116293  |
| Male   | v0:40.5-    | poor   | 0.331313   |
|        |             | rich   | 0.0971295  |
|        | v1:40.5+    | poor   | 0.134106   |
|        |             | rich   | 0.165933   |

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>&lt;40.5</td>
<td>.09</td>
<td>.91</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>&gt;40.5</td>
<td>.21</td>
<td>.79</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>&lt;40.5</td>
<td>.23</td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>&gt;40.5</td>
<td>.38</td>
<td>.62</td>
<td></td>
</tr>
</tbody>
</table>
How many parameters must we estimate?

Suppose \( X = \langle X_1, \ldots, X_n \rangle \)

where \( X_i \) and \( Y \) are boolean RV’s

To estimate \( P(Y \mid X_1, X_2, \ldots, X_n) \)

If we have 30 boolean \( X_i \)’s: \( P(Y \mid X_1, X_2, \ldots, X_{30}) \)

Bayes Rule

\[
P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}
\]

Which is shorthand for:

\[
(\forall i, j) P(Y = y_i \mid X = x_j) = \frac{P(X = x_j \mid Y = y_i)P(Y = y_i)}{P(X = x_j)}
\]

Equivalently:

\[
(\forall i, j) P(Y = y_i \mid X = x_j) = \frac{P(X = x_j \mid Y = y_i)P(Y = y_i)}{\sum_k P(X = x_j \mid Y = y_k)P(Y = y_k)}
\]
Can we reduce params using Bayes Rule?

Suppose \( X = \langle X_1, \ldots, X_n \rangle \)
where \( X_i \) and \( Y \) are boolean RV’s

\[
P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}
\]

Naïve Bayes

Naïve Bayes assumes

\[
P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)
\]

i.e., that \( X_i \) and \( X_j \) are conditionally independent given \( Y \), for all \( i \neq j \)
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.,

\[P(\text{Thunder}|\text{Rain, Lightning}) = P(\text{Thunder}|\text{Lightning})\]

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: \[P(X_1...X_n|Y) = \prod_i P(X_i|Y)\]

How many parameters to describe \(P(X_1...X_n|Y)\)? \(P(Y)\)?
- Without conditional indep assumption?
- With conditional indep assumption?
Naïve Bayes in a Nutshell

Bayes rule:
\[ P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)P(X_1 \ldots X_n|Y = y_k)}{\sum_j P(Y = y_j)P(X_1 \ldots X_n|Y = y_j)} \]

Assuming conditional independence among \(X_i\)'s:
\[ P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)} \]

So, to pick most probable \(Y\) for \(X_{\text{new}} = <X_1, \ldots, X_n>\)
\[ Y_{\text{new}} \leftarrow \arg \max_{y_k} P(Y = y_k)\prod_i P(X_i^{\text{new}}|Y = y_k) \]

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 \((X_{\text{new}})\)
\[ Y_{\text{new}} \leftarrow \arg \max_{y_k} P(Y = y_k)\prod_i P(X_i^{\text{new}}|Y = y_k) \]
\[ Y_{\text{new}} \leftarrow \arg \max_{y_k} \pi_k\prod_i \theta_{ijk} \]

* 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):

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

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

Example: Live in Sq Hill? $P(S|G,D,E)$

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- E=1 iff even # of letters in last name

What probability parameters must we estimate?
Example: Live in Sq Hill? $P(S|G,D,E)$

- $S=1$ iff live in Squirrel Hill
- $G=1$ iff shop at SH Giant Eagle
- $D=1$ iff Drive or Carpool to CMU
- $E=1$ iff Even # letters last name

$P(S=1):$ $P(S=0):$
$P(D=1 | S=1):$ $P(D=0 | S=1):$
$P(D=1 | S=0):$ $P(D=0 | S=0):$
$P(G=1 | S=1):$ $P(G=0 | S=1):$
$P(G=1 | S=0):$ $P(G=0 | S=0):$
$P(E=1 | S=1):$ $P(E=0 | S=1):$
$P(E=1 | S=0):$ $P(E=0 | S=0):$

Naïve Bayes: Subtlety #1

If unlucky, our MLE estimate for $P(X_i | Y)$ might be zero. (e.g., $X_i$ = Birthday_Is_January_30_1990)

- Why worry about just one parameter out of many?

- What can be done to avoid this?
Estimating Parameters

• Maximum Likelihood Estimate (MLE): choose \( \theta \) that maximizes probability of observed data \( \mathcal{D} \)

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

• Maximum a Posteriori (MAP) estimate: choose \( \theta \) that is most probable given prior probability and the data

\[
\hat{\theta} = \arg \max_\theta P(\theta \mid \mathcal{D}) = \arg \max_\theta \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}
\]

Estimating Parameters: \( Y, X_i \) discrete-valued

Maximum likelihood estimates:

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

\[
\hat{\theta}_{ijk} = \hat{P}(X_i = x_j \mid Y = y_k) = \frac{\#D\{X_i = x_j \land Y = y_k\}}{\#D\{Y = y_k\}}
\]

MAP estimates (Beta, Dirichlet priors):

\[
\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_{m} (\beta_m - 1)}
\]

\[
\hat{\theta}_{ijk} = \hat{P}(X_i = x_j \mid Y = y_k) = \frac{\#D\{X_i = x_j \land Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_{m} (\beta_m - 1)}
\]

Only difference: “imaginary” examples
Naïve Bayes: Subtlety #2

Often the $X_i$ are not really conditionally independent

- We use Naïve Bayes in many cases anyway, and it often works pretty well
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

- What is effect on estimated $P(Y|X)$?
  - Special case: what if we add two copies: $X_i = X_k$
Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?

Baseline: Bag of Words Approach

```
<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>0</td>
</tr>
<tr>
<td>about</td>
<td>2</td>
</tr>
<tr>
<td>all</td>
<td>2</td>
</tr>
<tr>
<td>Africa</td>
<td>1</td>
</tr>
<tr>
<td>apple</td>
<td>0</td>
</tr>
<tr>
<td>anxious</td>
<td>0</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>gas</td>
<td>1</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>oil</td>
<td>1</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>Zaire</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Learning to classify document: \( P(Y|X) \) 
the “Bag of Words” model

- \( Y \) discrete valued. e.g., Spam or not
- \( X = <X_1, X_2, ..., X_n> \) = document

- \( X_i \) is a random variable describing the word at position \( i \) in the document
- possible values for \( X_i \): any word \( w_k \) in English

- Document = bag of words: the vector of counts for all \( w_k \)’s
  - (like #heads, #tails, but we have more than 2 values)

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 \( x_j \) of each attribute \( X_i \)
  estimate \( \theta_{ijk} \equiv P(X_i = x_j | Y = y_k) \)

- Classify \( (X^{new}) \)

\[
Y^{new} \leftarrow \arg \max_{y_k} \quad P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)
\]

\[
Y^{new} \leftarrow \arg \max_{y_k} \quad \pi_k \prod_i \theta_{ijk}
\]

* Additional assumption: word probabilities are position independent
  \( \theta_{ijk} = \theta_{mjk} \) for all \( i, m \)
MAP estimates for bag of words

Map estimate for multinomial

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

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

What \( \beta \)'s should we choose?

Twenty NewsGroups

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

- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- alt.atheism
- soc.religion.christian
- talk.religion.misc
- talk.politics.mideast
- talk.politics.misc
- talk.politics.guns
- misc.forsale
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.space
- sci.crypt
- sci.electronics
- sci.med

Naive Bayes: 89% classification accuracy
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)
Questions:

• What error will the classifier achieve if Naïve Bayes assumption is satisfied and we have infinite training data?

• Can you use Naïve Bayes for a combination of discrete and real-valued $X_i$?

• How can we extend Naïve Bayes if just 2 of the $n X_i$ are dependent?

• What does the decision surface of a Naïve Bayes classifier look like?