Today:
• Bayes Classifiers
• Conditional Independence
• Naïve Bayes

Readings:
Mitchell:
“Naïve Bayes and Logistic Regression”
(available on class website)
Two Principles for Estimating Parameters

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

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

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

$$
\hat{\theta} = \arg \max_{\theta} P(\theta | \mathcal{D}) = \arg \max_{\theta} \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}
$$
Maximum Likelihood Estimate

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

  $\hat{\theta}^{MLE} = \arg \max_\theta P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$
Maximum A Posteriori (MAP) Estimate

- Data set $D$ of independent, identically distributed (iid) flips produces $\alpha_1$ ones, $\alpha_0$ zeros

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

- Assume prior $P(\theta) = \text{Beta}(\beta_1, \beta_0) = \frac{1}{B(\beta_1, \beta_0)} \theta^{\beta_1-1}(1 - \theta)^{\beta_0-1}$

- Then

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \frac{\alpha_1 + \beta_1 - 1}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)}$$

(like MLE, but hallucinating $\beta_1 - 1$ additional heads, $\beta_0 - 1$ additional tails)
Let’s learn classifiers by learning $P(Y|X)$

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

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
<th>$P(rich \mid G, HW)$</th>
<th>$P(poor \mid G, HW)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.253122</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0245895</td>
<td></td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.0421768</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0116293</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.331313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0971295</td>
<td></td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.134106</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.105933</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>HrsWorked</th>
<th>$P(rich \mid G, HW)$</th>
<th>$P(poor \mid G, HW)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>&lt;40.5</td>
<td>.09</td>
<td>.91</td>
</tr>
<tr>
<td>F</td>
<td>&gt;40.5</td>
<td>.21</td>
<td>.79</td>
</tr>
<tr>
<td>M</td>
<td>&lt;40.5</td>
<td>.23</td>
<td>.77</td>
</tr>
<tr>
<td>M</td>
<td>&gt;40.5</td>
<td>.38</td>
<td>.62</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: $P(Y \mid X_1, X_2, \ldots, X_{30})$
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 \( P(Y|X_1, X_2, \ldots, X_n) \),

\[ 2^n \]

If we have 30 \( X_i \)’s instead of 2,

\[ 2^{30} \approx 1 \text{ Billion} \]
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:

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

How many parameters to define $P(X_1, \ldots, X_n \mid Y)$?

How many parameters to define $P(Y)$?
Can we reduce params using Bayes Rule?

Suppose $X = <X_1, \ldots, X_n>$
where $X_i$ and $Y$ are boolean RV’s

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

how many params for $P(X_1 \ldots X_n | Y)$ $(2^n - 1) \cdot 2$

how many for $P(Y) = 1$
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.,

\[P(\text{Thunder}|\text{Rain}, \text{Lightning}) = P(\text{Thunder}|\text{Lightning})\]
Naïve Bayes uses assumption that the $X_i$ are conditionally independent, given $Y$. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$P(X_1, X_2|Y) =$$
Naïve Bayes uses assumption that the $X_i$ are conditionally independent, given $Y$. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

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: $P(X_1...X_n|Y) = \prod_{i} P(X_i|Y)$
Naïve Bayes uses assumption that the $X_i$ are conditionally independent, given $Y$. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

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 uses assumption that the $X_i$ are conditionally independent, given $Y$

Given this assumption, then:

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

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

Chain rule

Cond. Indep.

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? $2(2^n - 1) + 1$
- With conditional indep assumption? $2n + 1$
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 \(X^{new} = < X_1, \ldots, X_n >\)

\[
Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)
\]
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 $\theta_{ijk} \equiv P(X_i = x_{ij}\mid Y = y_k)$

- Classify ($X^{new}$)
  $Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new}\mid Y = y_k)$
  $Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$

* 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):

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

\[
\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\# D\{X_i = x_{ij} \land Y = y_k\}}{\# D\{Y = y_k\}}
\]
Example: Live in Sq Hill? $P(S|G,D,M)$

- $S=1$ iff live in Squirrel Hill
- $G=1$ iff shop at SH Giant Eagle
- $D=1$ iff Drive to CMU
- $M=1$ iff Rachel Maddow fan

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

- $S=1$ iff live in Squirrel Hill
- $G=1$ iff shop at SH Giant Eagle
- $D=1$ iff Drive to CMU
- $M=1$ iff Rachel Maddow fan

\[
P(S=1) : \; 0.4 \quad \frac{32}{80} = 0.4
\]

\[
P(S=0) : \; 0.6
\]

\[
P(D=1 | S=1) : \; \frac{6}{32} = 0.188
\]

\[
P(D=0 | S=1) : \; \frac{26}{80} = 0.325
\]

\[
P(D=1 | S=0) : \; \frac{18}{48} = 0.375
\]

\[
P(D=0 | S=0) : \; \frac{30}{80} = 0.375
\]

\[
P(G=1 | S=1) : \; \frac{29}{32} = 0.91
\]

\[
P(G=0 | S=1) : \; \frac{3}{32} = 0.09
\]

\[
P(G=1 | S=0) : \; \frac{17}{48} = 0.35
\]

\[
P(G=0 | S=0) : \; \frac{31}{48} = 0.65
\]

\[
P(M=1 | S=1) : \; \frac{2}{32} = 0.06
\]

\[
P(M=0 | S=1) : \; \frac{30}{32} = 0.94
\]

\[
P(M=1 | S=0) : \; \frac{29}{48} = 0.61
\]

\[
P(M=0 | S=0) : \; \frac{19}{48} = 0.39
\]

\[
P(S = 1 | G = 1, M = 0, D = 1) \Rightarrow P(S = 1) \times P(G = 1 | S = 1) \times P(D = 1 | S = 1) \times P(M = 0 | S = 1)
\]

Test: $S=1, G=1, D=1, M=0$ 

\[
P(S = 0 | X_{test}) \propto 0.6 \times 0.35 \times 0.38 = 0.74
\]

\[
P(S = 1 | X_{test}) \propto 0.4 \times 0.91 \times 0.19 = 0.64
\]

\[
P(S = 0 | X_{test}) = 0.54
\]

\[
P(S = 1 | X_{test}) = 0.46
\]
Example: Live in Sq Hill? \( P(S|G,D,B) \)

- \( S=1 \) iff live in Squirrel Hill
- \( G=1 \) iff shop at SH Giant Eagle
- \( D=1 \) iff Drive or carpool to CMU
- \( B=1 \) iff Birthday is before July 1

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
- $B=1$ iff Birthday is before July 1

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

$P(S=0)$ :
$P(D=0|S=1)$ :
$P(D=0|S=0)$ :
$P(G=0|S=1)$ :
$P(G=0|S=0)$ :
$P(B=0|S=1)$ :
$P(B=0|S=0)$ :
Naïve Bayes: Subtlety #1

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 $P(Y|X)$?
  – Extreme case: what if we add two copies: $X_i = X_k$
Extreme case: what if we add two copies: $X_i = X_k$
Extreme case: what if we add two copies: $X_i = X_k$

$$P(Y=y | x) \propto P(Y=y) \prod_i P(X_i=x | Y=y) \frac{P(x_1, \ldots, x_n | Y=y)}{P(x_1, \ldots, x_n)}$$
If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero. (for example, $X_i = \text{birthdate. } X_i = \text{Jan}_25\_1992$)

• Why worry about just one parameter out of many?

• What can be done to address this?
Naïve Bayes: Subtlety #2

If unlucky, our MLE estimate for \( P(X_i \mid Y) \) might be zero. (e.g., \( X_i = \text{Birthday \_Is \_January \_30 \_1992} \))

- Why worry about just one parameter out of many?

\[
P(Y \mid x) \propto P(Y) \prod_i P(x_i = x_{\text{new}} \mid Y)
\]

- What can be done to address this?
Estimating Parameters

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

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

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

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

$$= \arg \max_{\theta} \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}$$
Maximum likelihood estimates:

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

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

MAP estimates (Beta, Dirichlet priors):

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

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

Only difference: “imaginary” examples
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

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.
Learning to classify document: \( P(Y|X) \) the “Bag of Words” model

- \( Y \) discrete valued. e.g., Spam or not
- \( X = <X_1, X_2, \ldots, X_n> \) = 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 many more than 2 values
  - assume word probabilities are position independent
    (i.i.d. rolls of a 50,000-sided die)
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 ($X^{new}$)

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

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

* Additional assumption: word probabilities are position independent

  $\theta_{ijk} = \theta_{mjk}$ for all $i, m$
MAP estimates for bag of words

Map estimate for multinomial

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

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     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

Naive Bayes: 89% classification accuracy
Learning Curve for 20 Newsgroups

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

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

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

• 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 $X_i$?
What if we have continuous $X_i$?

Eg., image classification: $X_i$ is $i^{th}$ 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 $P(X_i | Y)$
What if we have continuous $X_i$?

Eg., image classification: $X_i$ is $i^{th}$ 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 $\sigma_{ik}$
- is independent of $Y$ (i.e., $\sigma_i$),
- or independent of $X_i$ (i.e., $\sigma_k$)
- or both (i.e., $\sigma$)
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 $\mu_{ik}$, variance $\sigma_{ik}$

- 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, $X_i$ continuous

Maximum likelihood estimates:

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

$$\hat{\sigma}^2_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X^j_i - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$
GNB Example: Classify a person’s cognitive activity, based on brain image

• are they reading a sentence or viewing a picture?

• reading the word “Hammer” or “Apartment”

• viewing a vertical or horizontal line?

• answering the question, or getting confused?
Stimuli for our study:

60 distinct exemplars, presented 6 times each
fMRI voxel means for “bottle”: means defining $P(X_i \mid Y=\text{“bottle”})$

Mean fMRI activation over all stimuli:

“bottle” minus mean activation:
Rank Accuracy Distinguishing among 60 words

![Graph showing identification accuracy for participants rank-ordered by within-participant accuracy. The graph compares within participants and across participants.]
Tools vs Buildings: where does brain encode their word meanings?

Accuracies of cubical 27-voxel Naïve Bayes classifiers centered at each voxel [0.7-0.8]
Expected values

Given discrete random variable $X$, the expected value of $X$, written $E[X]$ is

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

We also can talk about the expected value of functions of $X$

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x) P(X = x)$$
Covariance

Given two random vars $X$ and $Y$, we define the covariance of $X$ and $Y$ as

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

e.g., $X$=gender, $Y$=playsFootball
or $X$=gender, $Y$=leftHanded

Remember: $$E[X] = \sum_{x \in \mathcal{X}} xP(X = x)$$