Naïve Bayes classifiers
Types of classifiers

- We can divide the large variety of classification approaches into three major types

1. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

2. Generative:
   - build a generative statistical model
   - e.g., Bayesian networks

3. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., decision tree
Bayes decision rule

- If we know the conditional probability \( P(X \mid Y) \) we can determine the appropriate class by using Bayes rule:

\[
P(y = i \mid x) = \frac{P(x \mid y = i)P(y = i)}{P(x)} \overset{\text{def}}{=} q_i(x)
\]

But how do we determine \( p(X \mid Y) \)?
Consider a dataset with 16 attributes (we assume they are all binary). How many values do we need to know to fully determine \( p(x|y) \)?

<table>
<thead>
<tr>
<th>age</th>
<th>employment</th>
<th>education</th>
<th>edunum</th>
<th>marital</th>
<th>job</th>
<th>relation</th>
<th>race</th>
<th>gender</th>
<th>hours_worked</th>
<th>country</th>
<th>wealth</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>State_gov</td>
<td>Bachelors</td>
<td>13</td>
<td>Never_married</td>
<td>Adm_cler</td>
<td>Not_in_family</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>51</td>
<td>Self_emp</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>Exec_mananger</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>13</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>39</td>
<td>Private</td>
<td>HS_grad</td>
<td>9</td>
<td>Divorced</td>
<td>Handlers</td>
<td>Not_in_family</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>54</td>
<td>Private</td>
<td>11th</td>
<td>7</td>
<td>Married</td>
<td>Handlers</td>
<td>(Husband</td>
<td>Black</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>28</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>Prof_special</td>
<td>Wife</td>
<td>Black</td>
<td>Female</td>
<td>40</td>
<td>Cuba</td>
<td>poor</td>
</tr>
<tr>
<td>38</td>
<td>Private</td>
<td>Masters</td>
<td>14</td>
<td>Married</td>
<td>Exec_mananger</td>
<td>Wife</td>
<td>White</td>
<td>Female</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>50</td>
<td>Private</td>
<td>9th</td>
<td>5</td>
<td>Married_spouse</td>
<td>Other_service</td>
<td>Not_in_family</td>
<td>White</td>
<td>Female</td>
<td>16</td>
<td>Jamaica</td>
<td>poor</td>
</tr>
<tr>
<td>52</td>
<td>Private</td>
<td>HS_grad</td>
<td>9</td>
<td>Married</td>
<td>Exec_mananger</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>45</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>31</td>
<td>Private</td>
<td>Masters</td>
<td>14</td>
<td>Never_married</td>
<td>Prof_special</td>
<td>Not_in_family</td>
<td>White</td>
<td>Female</td>
<td>50</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>42</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>Exec_mananger</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>37</td>
<td>Private</td>
<td>Some_coll</td>
<td>10</td>
<td>Married</td>
<td>Exec_mananger</td>
<td>Husband</td>
<td>Black</td>
<td>Male</td>
<td>80</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>30</td>
<td>State_gov</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>Prof_special</td>
<td>Husband</td>
<td>Asian</td>
<td>Male</td>
<td>40</td>
<td>India</td>
<td>rich</td>
</tr>
<tr>
<td>24</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Never_married</td>
<td>Adm_cler</td>
<td>Own_child</td>
<td>White</td>
<td>Female</td>
<td>30</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>33</td>
<td>Private</td>
<td>Assoc_acc</td>
<td>12</td>
<td>Never_married</td>
<td>Sales</td>
<td>Not_in_family</td>
<td>Black</td>
<td>Male</td>
<td>50</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>41</td>
<td>Private</td>
<td>Assoc_voc</td>
<td>11</td>
<td>Married</td>
<td>Craft_repair</td>
<td>Husband</td>
<td>Asian</td>
<td>Male</td>
<td>40</td>
<td><em>MissingValue</em></td>
<td>rich</td>
</tr>
<tr>
<td>34</td>
<td>Private</td>
<td>7th_8th</td>
<td>4</td>
<td>Married</td>
<td>Transport</td>
<td>Husband</td>
<td>Amer_Indian</td>
<td>Male</td>
<td>45</td>
<td>Mexico</td>
<td>poor</td>
</tr>
<tr>
<td>26</td>
<td>Self_emp</td>
<td>HS_grad</td>
<td>9</td>
<td>Never_married</td>
<td>Farming</td>
<td>Own_child</td>
<td>White</td>
<td>Male</td>
<td>35</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>33</td>
<td>Private</td>
<td>HS_grad</td>
<td>9</td>
<td>Never_married</td>
<td>Machine</td>
<td>Unmarried</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>38</td>
<td>Private</td>
<td>11th</td>
<td>7</td>
<td>Married</td>
<td>Sales</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>50</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>44</td>
<td>Self_emp</td>
<td>Masters</td>
<td>14</td>
<td>Divorced</td>
<td>Exec_mananger</td>
<td>Unmarried</td>
<td>White</td>
<td>Female</td>
<td>45</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>41</td>
<td>Private</td>
<td>Doctorate</td>
<td>16</td>
<td>Married</td>
<td>Prof_special</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>60</td>
<td>United_States</td>
<td>rich</td>
</tr>
</tbody>
</table>

Learning the values for the full conditional probability table would require enormous amounts of data.
Naïve Bayes Classifiers assume that given the class label (Y) the attributes are conditionally independent of each other:

\[
p(x | y) = \prod_{i} p_i(x_i | y)
\]

Specific model for attribute \(i\)

Product of probability terms

Using this idea the full classification rule becomes:

\[
\hat{y} = \arg \max_v p(y = v | x)
\]

\[
= \arg \max_v \frac{p(x | y = v)p(y = v)}{p(x)}
\]

\[
= \arg \max_v \prod_{i} p_i(x_i | y = v)p(y = v)
\]
Data likelihood

• The biggest advantage of the Naïve Bayes classifier is that it requires relatively little data for estimating the parameters.
• The conditional independence assumption means that we can learn the parameters independently for each feature / attribute.
• The global likelihood of the data can be expressed as:

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

• Since the two parts of this product do not share parameters we can maximize them separately.

Recall that in supervised learning we receive both the features and the labels of the examples.
The global likelihood of the data can be expressed as:

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

Since the two parts of this product do not share parameters we can maximize them separately.

For binary attributes, assume we observe $n_0$ instances of $x_i=0$ and $n_1$ instances of $x_i=1$ for class 1 then:

$$L(X_i | y = 1) = \prod_{j} p_i(x_i^j | y = 1)$$

$$= \theta_{i|y=1}^{n_0} (1 - \theta_{i|y=1})^{n_1}$$

where $\theta_{i|y=1}$ is the conditional probability that $x_i = 0$ given the class $y=1$
Learning the parameters

\[ L(X \mid y = 1) = \prod \theta_{i|y=1}^{n_0} (1 - \theta_{i|y=1})^{n_1} \]

• To learn the parameters for this model we look for the value of \( \theta_{i|y=1} \) that maximizes the likelihood of the data given the model.
• In turns out (see the density estimation class) that the MLE estimator is:
  \[ \theta_{i|y=1} = \frac{n_0}{n_0 + n_1} \]
• We repeat this for each attribute in each of the two classes and obtain the full set of conditional probabilities
• How do we compute \( p(y=1) \)?
Example: Text classification

- What is the major topic of this article?

End of Battle Centers on Turf Bush Carried

In Denver on Sunday, Senator Barack Obama appeared at a rally at Civic Center Park that drew tens of thousands of people.

By ADAM NAUCKEY and JEFF ZELENY
Published: October 25, 2008

Senator John McCain and Senator Barack Obama are heading into the final week of the presidential campaign, planning to spend nearly all their time in states that President Bush won last time, testimony to the increasingly dire position of Mr. McCain and his party as Election
Example: Text classification

- Text classification is all around us
Feature transformation

• How do we encode the set of features (words) in the document?
• What type of information do we wish to represent? What can we ignore?
• Most common encoding: ‘Bag of Words’
• Treat document as a collection of words and encode each document as a vector based on some dictionary
• The vector can either be binary (present / absent information for each word) or discrete (number of appearances)

• Google is a good example
• Other applications include job search adds, spam filtering and many more.
Feature transformation: Bag of Words

• In this example we will use a binary vector
• For document $x_j$ we will use a vector of $m^*$ indicator features $\{\phi(x_j)\}$ for whether a word appears in the document
  - $\phi(x_j) = 1$, if word $i$ appears in document $x_j$;
  - $\phi(x_j) = 0$ if it does not appear in the document
• $\Phi(x_j) = [\phi^1(x_j) \ldots \phi^m(x_j)]^T$ is the resulting feature vector for the entire dictionary
• For notational simplicity we will replace each document $x_j$ with a fixed length vector $\Phi_j = [\phi^1 \ldots \phi^m]^T$, where $\phi = \phi(x_j)$.

*The size of the vector for English is usually ~10000 words*
Example

Dictionary
- Washington
- Congress
...
54. McCain
55. Obama
56. Nader

\[ \phi^{54} = \phi^{54}(x_j) = 1 \]
\[ \phi^{55} = \phi^{55}(x_j) = 1 \]
\[ \phi^{56} = \phi^{56}(x_j) = 0 \]

Assume we would like to classify documents as election related or not.

End of Battle Centers on Turf Bush Carried

In Denver on Sunday, Senator Barack Obama appeared at a rally at Civic Center Park that drew tens of thousands of people.

By ADAM NAGOURNEY and JEFF ZELENY
Published: October 20, 2008

Senator John McCain and Senator Barack Obama are heading into the final week of the presidential campaign planning to spend nearly all their time in states that President Bush won last time, testimony to the increasingly dire position of Mr. McCain and his party as Election...
Example: cont.

We would like to classify documents as election related or not.

- Given a collection of documents with their labels (usually termed ‘training data’) we learn the parameters for our model.
- For example, if we see the word ‘Bush’ in $n_1$ out of the $n$ documents labeled as ‘election’ we set $p(\text{‘bush’} | \text{‘election’}) = n_1/n$
- Similarly we compute the priors ($p(\text{‘election’})$) based on the proportion of the documents from both classes.
Example: Classifying Election (E) or Sports (S)

Assume we learned the following model

\[
P(\phi_{\text{bush}} = 1 \mid E) = 0.8, \quad P(\phi_{\text{bush}} = 1 \mid S) = 0.1 \quad P(S) = 0.5
\]
\[
P(\phi_{\text{obama}} = 1 \mid E) = 0.9, \quad P(\phi_{\text{obama}} = 1 \mid S) = 0.05 \quad P(E) = 0.5
\]
\[
P(\phi_{\text{mccain}} = 1 \mid E) = 0.9, \quad P(\phi_{\text{mccain}} = 1 \mid S) = 0.05
\]
\[
P(\phi_{\text{football}} = 1 \mid E) = 0.1, \quad P(\phi_{\text{football}} = 1 \mid S) = 0.7
\]

For a specific document we have the following feature vector

\[
\phi_{\text{bush}} = 1, \quad \phi_{\text{obama}} = 1, \quad \phi_{\text{mccain}} = 1, \quad \phi_{\text{football}} = 0
\]

\[
P(Y = E \mid 1,1,1,0) \propto 0.8*0.9*0.9*0.9*0.5 \quad = 0.5832
\]
\[
P(Y = S \mid 1,1,1,0) \propto 0.1*0.05*0.05*0.3*0.5 \quad = 0.000075
\]

So the document is classified as ‘Election’
Naïve Bayes classifiers for continuous values

• So far we assumed a binomial or discrete distribution for the data given the model \( p(x_i|y) \)
• However, in many cases the data contains continuous features:
  - Height, weight
  - Levels of genes in cells
  - Brain activity
• For these types of data we often use a Gaussian model
• In this model we assume that the observed input vector \( X \) is generated from the following distribution
  \[ X \sim N(\mu, \Sigma) \]
Gaussian Bayes Classification

• To determine the class when using the Gaussian assumption we need to compute $p(x|y)$:

\[
P(x \mid y) = \frac{1}{(2\pi)^{1/2} |\Sigma|^{1/2}} \exp \left[ (X - \mu)^T \Sigma^{-1} (X - \mu) \right]
\]

Once again, we need lots of data to compute the values of the covariance matrix $\Sigma$
Gaussian Bayes Classification

• Once again we use the Naïve Bayes assumption: Attributes are independent given the class label
• In the Gaussian model this means that the covariance matrix becomes a diagonal matrix with zeros everywhere except for the diagonal
• Thus, we only need to learn the values for the variance term for each attribute: \( x_i \sim N(\mu_i, \sigma_i^2) \).

\[
P(X | y = v) = \prod_i \frac{1}{\sqrt{2\pi} \sigma_{i,v}} \exp \left[ -\frac{(x_i - \mu_{i,v})^2}{2\sigma_{i,v}^2} \right]
\]

Separate means and variance for each class
MLE for Gaussian Naïve Bayes Classifier

- For each class we need to estimate one global value (prior) and two values for each feature (mean and variance).
- The prior is computed in the same way we did before (counting) which is the MLE estimate for each feature.
- The MLE for mean and variance is computed by setting:

\[
\mu_{i,0} = \frac{1}{n_0} \sum_{j=1}^{n} x_{i,j} \\
\sigma^2_{i,0} = \frac{1}{n_0} \sum_{j=1}^{n} (x_{i,j} - \mu_{i,0})^2
\]
Example: Classifying gene expression data

• Measures the levels (up or down) of genes in our cells
• Differs between healthy and sick people and between different disease types
• Given measurement of patients with two different types of cancer we would like to generate a classifier to distinguish between them
Classifying cancer types

• We select a subset of the genes (more in our ‘feature selection’ class later in the course).

• We compute the mean for each of the genes in each of the classes

\[
\mu = 1.8 \\
\sigma^2 = 1.1 \\
\mu = -0.6 \\
\sigma^2 = 0.4
\]
Classifying cancer types

- We compute the mean for each of the genes in each of the classes
  \[ \mu = 1.8 \quad \sigma^2 = 1.1 \]
- We determine the prior for each class
  \[ \mu = -0.6 \quad \sigma^2 = 0.4 \]
- We can compute our decision for each of the samples we have labels for
Classification accuracy

- The figure shows the value of the discriminate function

\[ f(x) = \log \frac{p(Y = 1 \mid X)}{p(Y = 0 \mid X)} \]

across the test examples

- The only test error is also the decision with the lowest confidence
FDA Approves Gene-Based Breast Cancer Test*

“MammaPrint is a DNA microarray-based test that measures the activity of 70 genes... The test measures each of these genes in a sample of a woman's breast-cancer tumor and then uses a specific formula to determine whether the patient is deemed low risk or high risk for the spread of the cancer to another site.”

*Washington Post, 2/06/2007
Possible problems with Naïve Bayes classifiers: Assumptions

• In most cases, the assumption of conditional independence given the class label is violated
  - much more likely to find the word ‘Barack’ if we saw the word ‘Obama’ regardless of the class

• This is, unfortunately, a major shortcoming which makes these classifiers inferior in many real world applications (though not always)

• There are models that can improve upon this assumption without using the full conditional model (one such model are Bayesian networks which we will discuss later in this class).
Possible problems with Naïve Bayes classifiers: Parameter estimation

- Even though we need far less data than the full Bayes model, there may be cases when the data we have is not enough.
- For example, what is $p(S=1,N=1|E=2)$?
- This can get worst. Assume we have 20 variables, almost all pointing in the direction of the same class except for one for which we have no record for this class.
- Solutions?

<table>
<thead>
<tr>
<th>Summer?</th>
<th>Num &gt; 20</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Important points

• Problems with estimating full joints
• Advantages of Naïve Bayes assumptions
• Applications to discrete and continuous cases
• Problems with Naïve Bayes classifiers