Classification

UE 141 Spring 2013

Jing Gao
SUNY Buffalo
### Classification

#### Features

<table>
<thead>
<tr>
<th>Patient</th>
<th>Temp.</th>
<th>Blood Pres.</th>
<th>Heart Rate</th>
<th>Sick?</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Patient 1]</td>
<td>99</td>
<td>110</td>
<td>90</td>
<td>Yes</td>
</tr>
<tr>
<td>![Patient 2]</td>
<td>100</td>
<td>120</td>
<td>100</td>
<td>Yes</td>
</tr>
<tr>
<td>![Patient 3]</td>
<td>96</td>
<td>130</td>
<td>65</td>
<td>No</td>
</tr>
</tbody>
</table>

#### Class Labels

- *a model: f(x)=y*: features $\rightarrow$ class labels

<table>
<thead>
<tr>
<th>Patient</th>
<th>Temp.</th>
<th>Blood Pres.</th>
<th>Heart Rate</th>
<th>Sick?</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Patient 1]</td>
<td>98</td>
<td>130</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>![Patient 2]</td>
<td>115</td>
<td>110</td>
<td>95</td>
<td></td>
</tr>
</tbody>
</table>

#### Diagram

- Labeled data
- Unlabeled data
- Training data
- Test data
Ensemble Learning

Training set

Test example $x$

$f^1(x)$

$f^2(x)$

$f^k(x)$

$f(x)$
Ensemble Learning

• **Problem**

  – Given a data set \( D=\{x_1, x_2, \ldots, x_n\} \) and their corresponding labels \( L=\{l_1, l_2, \ldots, l_n\} \)

  – An ensemble approach computes:
    
    • A set of classifiers \( \{f_1, f_2, \ldots, f_k\} \), each of which maps data to a class label: \( f_j(x)=l \)
    
    • A combination of classifiers \( f^* \) based on \( \{f_1, f_2, \ldots, f_k\} \)
Why Ensemble Works? (1)

• **Intuition**
  – combining diverse, independent opinions in human decision-making as a protective mechanism (e.g. stock portfolio)

• **Stock investment**
  – Invest all the money on one stock is very risky
  – Distribute your money across multiple stocks is the best way to guarantee stable return
Why Ensemble Works? (2)

• **Uncorrelated error reduction**
  – Suppose we have 5 completely independent classifiers for majority voting
  – If accuracy is 70% for each
    • $10 \cdot (.7^3)(.3^2)+5(.7^4)(.3)+(.7^5)$
    • 83.7% majority vote accuracy
  – 101 such classifiers
    • 99.9% majority vote accuracy
Why Ensemble Works? (3)

- Overcome limitations of single hypothesis
  - The target function may not be implementable with individual classifiers, but may be approximated by model averaging.
Generating Base Classifiers

- **Sampling training examples**
  - Train k classifiers on k subsets drawn from the training set

- **Using different learning models**
  - Use all the training examples, but apply different learning algorithms

- **Sampling features**
  - Train k classifiers on k subsets of features drawn from the feature space

- **Learning “randomly”**
  - Introduce randomness into learning procedures
Bagging

• **Training set**
  – Sampling with replacement
  – Sample a subset from the training set

• **Ensemble learning**
  – Train a classifier on each sample
  – Use majority voting to determine the class label of ensemble classifier
Bagging

Original Data:

<table>
<thead>
<tr>
<th>x</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Samples and classifiers:

<table>
<thead>
<tr>
<th>x</th>
<th>0.1</th>
<th>0.2</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.9</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.5</th>
<th>0.9</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.4</th>
<th>0.5</th>
<th>0.7</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>0.1</th>
<th>0.2</th>
<th>0.5</th>
<th>0.5</th>
<th>0.5</th>
<th>0.7</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Combine predictions by majority voting
**Boosting**

**Principles**
- Boost a set of weak learners to a strong learner
- Make records currently misclassified more important

**Example**
- Record 4 is hard to classify
- Its weight is increased, therefore it is more likely to be chosen again in subsequent rounds

<table>
<thead>
<tr>
<th>Original Data</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boosting (Round 1)</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Boosting (Round 2)</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Boosting (Round 3)</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Classifications (colors) and Weights (size) after 1 iteration of Boosting
Random Forests

...... A lot more ways to build a decision tree from the data

Instead of selecting one best tree among all the trees, let’s combine them!