Late homework submissions:
- Strictly by 11:59PM on the due date (already N weeks + 1 extra day)
- No points will be given after that
- I may or may not send you reminders

Suggestions:
- Set reminders on your google calendar, remind each other, etc.
- Submit one version before the deadline, and update it later.

STDLINUX account?
- If not, contact help desk at CSE: help@cse.ohio-state.edu
Classification: Basic Concepts

- Classification: Basic Concepts
- Decision Tree Induction
- Model Evaluation and Selection
- Practical Issues of Classification
- Bayes Classification Methods
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Summary

This class
Decision Tree Induction: An Example

- Training data set: Buys_computer
- The data set follows an example of Quinlan’s ID3 (Playing Tennis)
- Resulting tree:

```
<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31...40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31...40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31...40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31...40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>
```
Algorithm for Decision Tree Induction

- **Basic algorithm (a greedy algorithm)**
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

- **Conditions for stopping partitioning**
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning—majority voting is employed for classifying the leaf
  - There are no samples left
Algorithm Outline

- Split (node, {data tuples})
  - A <= the best attribute for splitting the {data tuples}
  - Decision attribute for this node <= A
  - For each value of A, create new child node
  - For each child node / subset:
    - If one of the stopping conditions is satisfied: STOP
    - Else: Split (child_node, {subset})

ID3 algorithm: how it works
https://www.youtube.com/watch?v= XhOdSLIE5c
Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let $p_i$ be the probability that an arbitrary tuple in $D$ belongs to class $C_i$, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in $D$:
  \[
  Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)
  \]
- Information needed (after using $A$ to split $D$ into $v$ partitions) to classify $D$:
  \[
  Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)
  \]
- Information gained by branching on attribute $A$
  \[
  Gain(A) = Info(D) - Info_A(D)
  \]
**Attribute Selection: Information Gain**

- **Class P:** buys_computer = “yes”
- **Class N:** buys_computer = “no”

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31…40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31…40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31…40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31…40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>

\[
\text{Info}(D) = I(9,5) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940
\]

\[
\text{Info}_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694
\]

\[
\text{Gain}(age) = \text{Info}(D) - \text{Info}_{age}(D) = 0.246
\]

Similarly,
\[
\text{Gain}(income) = 0.029
\]
\[
\text{Gain}(student) = 0.151
\]
\[
\text{Gain}(credit\_rating) = 0.048
\]
Gain Ratio for Attribute Selection (C4.5)

- Information gain measure is biased towards attributes with a large number of values.
- C4.5 (a successor of ID3) uses gain ratio to overcome the problem (normalization to information gain).

\[
SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right)
\]

- The entropy of the partitioning, or the potential information generated by splitting \( D \) into \( v \) partitions.

- \( \text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)} \) (normalizing Information Gain)
Splitting Based on Nominal Attributes

- **Multi-way split:** Use as many partitions as distinct values.

- **Binary split:** Divides values into two subsets. Need to find optimal partitioning.
Measures of Node Impurity

- **Entropy:**
  \[ H(Y) = -\sum_{i=1}^{m} p_i \log(p_i) \text{ where } p_i = P(Y = y_i) \]
  - Higher entropy $\Rightarrow$ higher uncertainty, higher node impurity
  - Why entropy is used in information gain

- **Gini Index**

- **Misclassification error**
Gini Index (CART, IBM IntelligentMiner)

- If a data set $D$ contains examples from $n$ classes, gini index, $gini(D)$ is defined as
  \[
gini(D) = 1 - \sum_{j=1}^{n} p_j^2,
\]
  where $p_j$ is the relative frequency of class $j$ in $D$.

- If a data set $D$ is split on $A$ into two subsets $D_1$ and $D_2$, the gini index after the split is defined as
  \[
gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2).
\]

- Reduction in impurity:
  \[
  \Delta gini(A) = gini(D) - gini_A(D).
\]

- The attribute provides the smallest $gini_A(D)$ (or, the largest reduction in impurity) is chosen to split the node.
Binary Attributes: Computing Gini Index

- Splits into two partitions
- Effect of weighing partitions:
  - Prefer Larger and Purer Partitions.

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

<table>
<thead>
<tr>
<th>Parent</th>
<th>N1</th>
<th>N2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>C2</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Gini(N1) = 1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2 = 0.408

Gini(N2) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.320

Gini(Children) = \frac{7}{12} \times 0.408 + \frac{5}{12} \times 0.320 = 0.371

Node N1

Node N2

B?

Yes

No

Some errors in last lectures

weighting
Categorical Attributes: Computing Gini Index

- For each distinct value, gather counts for each class in the dataset
- Use the count matrix to make decisions

<table>
<thead>
<tr>
<th>CarType</th>
<th>Family</th>
<th>Sports</th>
<th>Luxury</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.393</td>
</tr>
<tr>
<td>C2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Multi-way split

<table>
<thead>
<tr>
<th>CarType</th>
<th>{Sports, Luxury}</th>
<th>{Family}</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>3</td>
<td>1</td>
<td>0.400</td>
</tr>
<tr>
<td>C2</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Two-way split (find best partition of values)

<table>
<thead>
<tr>
<th>CarType</th>
<th>{Sports}</th>
<th>{Family, Luxury}</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>2</td>
<td>2</td>
<td>0.419</td>
</tr>
<tr>
<td>C2</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Continuous Attributes: Computing Gini Index or Information Gain

- To discretize the attribute values
  - Use Binary Decisions based on one splitting value

- Several Choices for the splitting value
  - Number of possible splitting values = Number of distinct values - 1
  - Typically, the midpoint between each pair of adjacent values is considered as a possible split point
    - \((a_i + a_{i+1})/2\) is the midpoint between the values of \(a_i\) and \(a_{i+1}\)

- Each splitting value has a count matrix associated with it
  - Class counts in each of the partitions, \(A < v\) and \(A \geq v\)

- Simple method to choose best \(v\)
  - For each \(v\), scan the database to gather count matrix and compute its Gini index
  - Computationally Inefficient! Repetition of work.

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Continuous Attributes:
Computing Gini Index or expected information requirement

First decide the splitting value to discretize the attribute:

- For efficient computation: for each attribute,
  1. Sort the attribute on values
  2. Linearly scan these values, each time updating the count matrix
  3. **Computing Gini index and choose the split position that has the least Gini index**

<table>
<thead>
<tr>
<th>Cheat</th>
<th>No</th>
<th>No</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
<th>No</th>
<th>No</th>
<th>No</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Taxable Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
</tr>
</tbody>
</table>

**Step 1:**
Possible Splitting Values
Sorted Values

<table>
<thead>
<tr>
<th>&lt;=</th>
<th>&gt;</th>
<th>&lt;=</th>
<th>&gt;</th>
<th>&lt;=</th>
<th>&gt;</th>
<th>&lt;=</th>
<th>&gt;</th>
<th>&lt;=</th>
<th>&gt;</th>
<th>&lt;=</th>
<th>&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>72</td>
<td>80</td>
<td>87</td>
<td>92</td>
<td>97</td>
<td>110</td>
<td>122</td>
<td>172</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 2:**

<table>
<thead>
<tr>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Step 3:**

<table>
<thead>
<tr>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.400</td>
</tr>
</tbody>
</table>

At each level of the decision tree, for attribute selection, (1) First, discretize a continuous attribute by deciding the splitting value; (2) Then, compare the discretized attribute with other attributes in terms of Gini Index reduction or Information Gain.
Classification: Basic Concepts

- Classification: Basic Concepts
- Decision Tree Induction
- Model Evaluation and Selection
- Practical Issues of Classification
- Bayes Classification Methods
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Summary
**Classifier Evaluation Metrics: Confusion Matrix**

<table>
<thead>
<tr>
<th>Actual class\Predicted class</th>
<th>C₁</th>
<th>¬ C₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>True Positives (TP)</td>
<td>False Negatives (FN)</td>
</tr>
<tr>
<td>¬ C₁</td>
<td>False Positives (FP)</td>
<td>True Negatives (TN)</td>
</tr>
</tbody>
</table>

**Example of Confusion Matrix:**

<table>
<thead>
<tr>
<th>Actual class\Predicted class</th>
<th>buy_computer = yes</th>
<th>buy_computer = no</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>buy_computer = yes</td>
<td>6954</td>
<td>46</td>
<td>7000</td>
</tr>
<tr>
<td>buy_computer = no</td>
<td>412</td>
<td>2588</td>
<td>3000</td>
</tr>
<tr>
<td>Total</td>
<td>7366</td>
<td>2634</td>
<td>10000</td>
</tr>
</tbody>
</table>

- Given $m$ classes, an entry, $CM_{ij}$, in a **confusion matrix** indicates # of tuples in class $i$ that were labeled by the classifier as class $j$
  - May have extra rows/columns to provide totals
Classifier Evaluation Metrics: Accuracy, Error Rate

<table>
<thead>
<tr>
<th>A \ P</th>
<th>C</th>
<th>¬C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>¬C</td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td></td>
<td>P'</td>
<td>N'</td>
</tr>
</tbody>
</table>

- **Classifier Accuracy**, or recognition rate: percentage of test set tuples that are correctly classified

  \[ \text{Accuracy} = \frac{(TP + TN)}{\text{All}} \]

- **Error rate**: \(1 - \text{accuracy}, \text{or} \)

  \[ \text{Error rate} = \frac{(FP + FN)}{\text{All}} \]
Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If a model predicts everything to be class 0,
  Accuracy is $9990/10000 = 99.9\%$

- Accuracy is misleading because model does not detect any class 1 example
Cost-Sensitive Measures

\[
\text{Precision (p)} = \frac{a}{a + c}
\]

\[
\text{Recall (r)} = \frac{a}{a + b}
\]

\[
\text{F-measure (F)} = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}
\]

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

Weighted Accuracy = \[
\frac{w_1a + w_4d}{w_1a + w_2b + w_3c + w_4d}
\]

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class=Yes</td>
<td>a (TP)</td>
</tr>
<tr>
<td>Class=No</td>
<td>b (FN)</td>
</tr>
<tr>
<td>Class=Yes</td>
<td>c (FP)</td>
</tr>
<tr>
<td>Class=No</td>
<td>d (TN)</td>
</tr>
</tbody>
</table>
Evaluating Classifier Accuracy: Holdout & Cross-Validation Methods

- **Holdout method**
  - Given data is randomly partitioned into two independent sets
    - Training set (e.g., 2/3) for model construction
    - Test set (e.g., 1/3) for accuracy estimation
  - **Random sampling**: a variation of holdout
    - Repeat holdout k times, accuracy = avg. of the accuracies obtained

- **Cross-validation** (k-fold, where k = 10 is most popular)
  - Randomly partition the data into k mutually exclusive subsets, each approximately equal size
  - At i-th iteration, use D_i as test set and others as training set
  - **Leave-one-out**: k folds where k = # of tuples, for small sized data
  - **Stratified cross-validation**: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data
Using ROC for Classification Model Comparison

- **ROC (Receiver Operating Characteristics)** curves: for visual comparison of classification models
- Originated from signal detection theory
- Shows the trade-off between the true positive rate and the false positive rate
- The area under the ROC curve is a measure of the accuracy of the model
- Rank the test tuples in decreasing order: the one that is most likely to belong to the positive class appears at the top of the list
- The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model
Using ROC for Classification Model Comparison

- No model consistently outperform the other
  - $M_1$ is better for small FPR
  - $M_2$ is better for large FPR

- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess (diagonal line):
    - Area = 0.5
Classification: Basic Concepts

- Classification: Basic Concepts
- Decision Tree Induction
- Model Evaluation and Selection
- Practical Issues of Classification
- Bayes Classification Methods
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Summary
Practical Issues of Classification

- Underfitting and Overfitting
- Missing Values
- Costs of Classification
Underfitting and Overfitting

Underfitting: when model is too simple, both training and test errors are large
Overfitting due to Noise

Decision boundary is distorted by noise point
Overfitting due to Insufficient Examples

Lack of data points in the lower half of the diagram makes it difficult to predict correctly the class labels of that region.

- Insufficient number of training records in the region causes the decision tree to predict the test examples using other training records that are irrelevant to the classification task.
Notes on Overfitting

- Overfitting results in decision trees that are more complex than necessary.

- Training error no longer provides a good estimate of how well the tree will perform on previously unseen records.

- Need new ways for estimating errors.
Estimating Generalization Errors

- **Re-substitution errors**: error on training ($\sum e(t)$)
- **Generalization errors**: error on testing ($\sum e'(t)$)
Estimating Generalization Errors

- **Re-substitution errors**: error on training ($\sum e(t)$)
- **Generalization errors**: error on testing ($\sum e'(t)$)

- **Methods for estimating generalization errors**:
  - Optimistic approach: $e'(t) = e(t)$
Estimating Generalization Errors

- **Re-substitution errors**: error on training ($\sum e(t)$)
- **Generalization errors**: error on testing ($\sum e'(t)$)

**Methods for estimating generalization errors:**

- **Optimistic approach**: $e'(t) = e(t)$
- **Pessimistic approach**:
  - For each leaf node: $e'(t) = (e(t)+0.5)$
  - Total errors: $e'(T) = e(T) + N \times 0.5$ (N: number of leaf nodes)
  - For a tree with 30 leaf nodes and 10 errors on training (out of 1000 instances):

    - Training error = $10/1000 = 1\%$
    - Generalization error = $(10 + 30 \times 0.5)/1000 = 2.5\%$
Estimating Generalization Errors

- **Re-substitution errors**: error on training ($\sum e(t)$)
- **Generalization errors**: error on testing ($\sum e'(t)$)

**Methods for estimating generalization errors**:
- **Optimistic approach**: $e'(t) = e(t)$
- **Pessimistic approach**:
  - For each leaf node: $e'(t) = (e(t) + 0.5)$
  - Total errors: $e'(T) = e(T) + N \times 0.5$ ($N$: number of leaf nodes)
  - For a tree with 30 leaf nodes and 10 errors on training (out of 1000 instances):
    - Training error = $10/1000 = 1\%$
    - Generalization error = $(10 + 30 \times 0.5)/1000 = 2.5\%$

- **Reduced error pruning (REP)**:
  - uses validation data set to estimate generalization error
How to Address Overfitting

- Pre-Pruning (Early Stopping Rule)
  - Stop the algorithm before it becomes a fully-grown tree
  - Typical stopping conditions for a node:
    - Stop if all instances belong to the same class
    - Stop if all the attribute values are the same
  - More restrictive conditions:
    - Stop if number of instances is less than some user-specified threshold
    - Stop if class distribution of instances are independent of the available features (e.g., using $\chi^2$ test)
    - Stop if expanding the current node does not improve impurity measures (e.g., Gini or information gain).
How to Address Overfitting…

- **Post-pruning**
  - Grow decision tree to its entirety
  - Trim the nodes of the decision tree in a bottom-up fashion
  - If generalization error improves after trimming, replace sub-tree by a leaf node.
  - Class label of leaf node is determined from majority class of instances in the sub-tree
  - Can use MDL for post-pruning
Example of Post-Pruning

| Class = Yes | 20 |
| Class = No | 10 |
| **Error** = 10/30 |

Training Error (Before splitting) = 10/30

Pessimistic error = (10 + 0.5)/30 = 10.5/30

Training Error (After splitting) = 9/30

Pessimistic error (After splitting) = (9 + 4 \times 0.5)/30 = 11/30

PRUNE!
Examples of Post-pruning

- **Optimistic error?**
  Don’t prune for both cases

- **Pessimistic error?**
  Don’t prune case 1, prune case 2

- **Reduced error pruning?**
  Depends on validation set

Case 1:
- C0: 11
- C1: 3
- C0: 2
- C1: 4

Case 2:
- C0: 14
- C1: 3
- C0: 2
- C1: 2
Occam’s Razor

- Given two models of similar generalization errors, one should prefer the simpler model over the more complex model.

- For complex models, there is a greater chance that it was fitted accidentally by errors in data.

- Therefore, one should include model complexity when evaluating a model.
Handling Missing Attribute Values

- Missing values affect decision tree construction in three different ways:
  - Affects how impurity measures are computed
  - Affects how to distribute instance with missing value to child nodes
  - Affects how a test instance with missing value is classified

- While the book describes a few ways it can be handled as part of the process – it is often best to handle this using standard statistical methods
  - EM-based estimation (Covered in the second lecture, 08/24/2017)
Bayesian Classification: Why?

- **A statistical classifier:** performs *probabilistic prediction*, i.e., predicts class membership probabilities.

- **Foundation:** Based on Bayes’ Theorem.

- **Performance:** A simple Bayesian classifier, naïve Bayesian classifier, has comparable performance with decision tree and selected neural network classifiers.

- **Incremental:** Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data.

- **Standard:** Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured.
Bayes’ Theorem: Basics

- Bayes’ Theorem:
  - Let $X$ be a data sample ("evidence"): class label is unknown
  - Let $H$ be a hypothesis that $X$ belongs to class $C$
  - Classification is to determine $P(H|X)$, (i.e., posteriori probability): the probability that the hypothesis holds given the observed data sample $X$

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} = P(X|H) \times P(H)/P(X)$$
Bayes’ Theorem: Basics

- Bayes’ Theorem:
  - Let \( X \) be a data sample (“evidence”): class label is unknown
  - Let \( H \) be a hypothesis that \( X \) belongs to class \( C \)
  - Classification is to determine \( P(H \mid X) \), (i.e., posteriori probability): the probability that the hypothesis holds given the observed data sample \( X \)
    
    \[
P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} = P(X \mid H) \times P(H) / P(X)
    \]
  - \( P(H) \) (prior probability): the initial probability
    - E.g., \( X \) will buy computer, regardless of age, income, ...
Bayes’ Theorem: Basics

- **Bayes’ Theorem:**
  - Let $X$ be a data sample ("evidence"): class label is unknown
  - Let $H$ be a hypothesis that $X$ belongs to class $C$
  - Classification is to determine $P(H|X)$, (i.e., *posteriori probability*): the probability that the hypothesis holds given the observed data sample $X$
  
  $$P(H|X) = \frac{P(X|H)P(H)}{P(X)} = \frac{P(X|H) \times P(H)}{P(X)}$$

- $P(H)$ (*prior probability*): the initial probability
  - E.g., $X$ will buy computer, regardless of age, income, …
- $P(X)$: probability that sample data is observed
Bayes’ Theorem: Basics

- **Bayes’ Theorem:**
  - Let $X$ be a data sample (“evidence”): class label is unknown
  - Let $H$ be a hypothesis that $X$ belongs to class $C$
  - Classification is to determine $P(H|X)$, (i.e., *posterior probability*): the probability that the hypothesis holds given the observed data sample $X$

\[
P(H|X) = \frac{P(X|H)P(H)}{P(X)} = P(X|H) \times P(H) / P(X)
\]

- $P(H)$ (*prior probability*): the initial probability
  - E.g., $X$ will buy computer, regardless of age, income, …
- $P(X)$: probability that sample data is observed
- $P(X|H)$ (likelihood): the probability of observing the sample $X$, given that the hypothesis holds
  - E.g., Given that $X$ will buy computer, the prob. that $X$ is 31..40, medium income
Prediction Based on Bayes’ Theorem

- Given training data \( X \), posteriori probability of a hypothesis \( H \), \( P(H \mid X) \), follows the Bayes’ theorem

\[
P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} = P(X \mid H) \times P(H) / P(X)
\]

- Informally, this can be viewed as

  \[
  \text{posteriori} = \text{likelihood} \times \text{prior/evidence}
  \]

- Predicts \( X \) belongs to \( C_i \) iff the probability \( P(C_i \mid X) \) is the highest among all the \( P(C_k \mid X) \) for all the \( k \) classes

- Practical difficulty: It requires initial knowledge of many probabilities, involving significant computational cost
Classification Is to Derive the Maximum Posteriori

- Let $D$ be a training set of tuples and their associated class labels, and each tuple is represented by an $n$-dimensional attribute vector $X = (x_1, x_2, ..., x_n)$.
- Suppose there are $m$ classes $C_1, C_2, ..., C_m$.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i | X)$. 
Classification Is to Derive the Maximum Posteriori

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-dimensional attribute vector \( X = (x_1, x_2, \ldots, x_n) \).
- Suppose there are \( m \) classes \( C_1, C_2, \ldots, C_m \).
- Classification is to derive the maximum posteriori, i.e., the maximal \( P(C_i | X) \).
- This can be derived from Bayes’ theorem

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

- Since \( P(X) \) is constant for all classes, only

\[
P(C_i | X) = P(X | C_i) P(C_i)
\]

needs to be maximized.
A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

\[ P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i) \]

This greatly reduces the computation cost: Only counts the class distribution
Naïve Bayes Classifier

- A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

  \[ P(X | C_i) = \prod_{k=1}^{n} P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \ldots \times P(x_n | C_i) \]

- This greatly reduces the computation cost: Only counts the class distribution

- If \( A_k \) is categorical, \( P(x_k | C_i) \) is the number of tuples in \( C_i \) having value \( x_k \) for \( A_k \) divided by \( |C_i, D| \) (number of tuples of \( C_i \) in \( D \))
Naïve Bayes Classifier

- A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

  \[ P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i) \]

- If \( A_k \) is continuous-valued, \( P(x_k|C_i) \) is usually computed based on Gaussian distribution with a mean \( \mu \) and standard deviation \( \sigma \)

  \[ g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

  and \( P(x_k|C_i) \) is

  \[ P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \]
Naïve Bayes Classifier: Training Dataset

Class:
C1:buys_computer = ‘yes’
C2:buys_computer = ‘no’

Data to be classified:
X = (age <=30, Income = medium,
Student = yes, Credit_rating = Fair)

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31…40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31…40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31…40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31…40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>
Naïve Bayes Classifier: An Example

- $P(C_i): P(\text{buys\_computer} = \text{“yes”}) = \frac{9}{14} = 0.643$
- $P(\text{buys\_computer} = \text{“no”}) = \frac{5}{14} = 0.357$
Naïve Bayes Classifier: An Example

- \( P(C_i): \) \( P(\text{buys_computer} = \text{“yes”}) = 9/14 = 0.643 \)
  \( P(\text{buys_computer} = \text{“no”}) = 5/14 = 0.357 \)

- Compute \( P(X|C_i) \) for each class, where,
  \( X = (\text{age} \leq 30, \text{Income} = \text{medium}, \text{Student} = \text{yes}, \text{Credit_rating} = \text{Fair}) \)
  \[ P(\text{age} = \text{“\leq 30”} | \text{buys_computer} = \text{“yes”}) = 2/9 = 0.222 \]
Naïve Bayes Classifier: An Example

- $P(C_i)$: $P(\text{buys_computer} = \text{“yes”}) = 9/14 = 0.643$
  $P(\text{buys_computer} = \text{“no”}) = 5/14 = 0.357$

- Compute $P(X|C_i)$ for each class, where,
  $X = (\text{age} \leq 30, \text{Income} = \text{medium}, \text{Student} = \text{yes}, \text{Credit_rating} = \text{Fair})$

  $P(\text{age} = \text{“} \leq 30\text{”} | \text{buys_computer} = \text{“yes”}) = 2/9 = 0.222$
  $P(\text{age} = \text{“} \leq 30\text{”} | \text{buys_computer} = \text{“no”}) = 3/5 = 0.6$
  $P(\text{income} = \text{“medium”} | \text{buys_computer} = \text{“yes”}) = 4/9 = 0.444$
  $P(\text{income} = \text{“medium”} | \text{buys_computer} = \text{“no”}) = 2/5 = 0.4$
  $P(\text{student} = \text{“yes”} | \text{buys_computer} = \text{“yes”}) = 6/9 = 0.667$
  $P(\text{student} = \text{“yes”} | \text{buys_computer} = \text{“no”}) = 1/5 = 0.2$
  $P(\text{credit_rating} = \text{“fair”} | \text{buys_computer} = \text{“yes”}) = 6/9 = 0.667$
  $P(\text{credit_rating} = \text{“fair”} | \text{buys_computer} = \text{“no”}) = 2/5 = 0.4$
Naïve Bayes Classifier: An Example

- \( P(C_i): \) \( P(\text{buys\_computer} = \text{"yes"}) = 9/14 = 0.643 \)
- \( P(\text{buys\_computer} = \text{"no"}) = 5/14 = 0.357 \)

- Compute \( P(X_i|C_i) \) for each class
  - \( P(\text{age} = \text{"<=30"} | \text{buys\_computer} = \text{"yes"}) = 2/9 = 0.222 \)
  - \( P(\text{age} = \text{"<=30"} | \text{buys\_computer} = \text{"no"}) = 3/5 = 0.6 \)
  - \( P(\text{income} = \text{"medium"} | \text{buys\_computer} = \text{"yes"}) = 4/9 = 0.444 \)
  - \( P(\text{income} = \text{"medium"} | \text{buys\_computer} = \text{"no"}) = 2/5 = 0.4 \)
  - \( P(\text{student} = \text{"yes"} | \text{buys\_computer} = \text{"yes"}) = 6/9 = 0.667 \)
  - \( P(\text{student} = \text{"yes"} | \text{buys\_computer} = \text{"no"}) = 1/5 = 0.2 \)
  - \( P(\text{credit\_rating} = \text{"fair"} | \text{buys\_computer} = \text{"yes"}) = 6/9 = 0.667 \)
  - \( P(\text{credit\_rating} = \text{"fair"} | \text{buys\_computer} = \text{"no"}) = 2/5 = 0.4 \)

- \( X = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair}) \)

\[
P(X | C_i) = P(X | \text{buys\_computer} = \text{"yes"}) = P(\text{age} = \text{"<=30"} | \text{buys\_computer} = \text{"yes"}) \times P(\text{income} = \text{"medium"} | \text{buys\_computer} = \text{"yes"}) \times P(\text{student} = \text{"yes"} | \text{buys\_computer} = \text{"yes"}) \times P(\text{credit\_rating} = \text{"fair"} | \text{buys\_computer} = \text{"yes"})
\]
Naïve Bayes Classifier: An Example

- P(Ci): P(buys_computer = “yes”) = 9/14 = 0.643
  P(buys_computer = “no”) = 5/14 = 0.357

- Compute P(Xi | Ci) for each class
  
  \[
P(\text{age} = \text{"<=30"} | \text{buys_computer} = \text{"yes"}) = 2/9 = 0.222
  
P(\text{age} = \text{"<=30"} | \text{buys_computer} = \text{"no"}) = 3/5 = 0.6
  
P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"yes"}) = 4/9 = 0.444
  
P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.4
  
P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667
  
P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"no"}) = 1/5 = 0.2
  
P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667
  
P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.4
  
\]

- X = (age <= 30, income = medium, student = yes, credit_rating = fair)
  
P(X | Ci) : P(X | buys_computer = “yes”) = P(\text{age} = \text{"<=30"} | \text{buys_computer} = \text{"yes"}) * P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"yes"}) * P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"yes"}) * P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"yes"}) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
Naïve Bayes Classifier: An Example

- $P(C_i): P(\text{buys\_computer} = \text{“yes”}) = 9/14 = 0.643$
  
  $P(\text{buys\_computer} = \text{“no”}) = 5/14 = 0.357$

- Compute $P(X_i|C_i)$ for each class
  
  $P(\text{age} = \text{“<=30”} | \text{buys\_computer} = \text{“yes”}) = 2/9 = 0.222$
  
  $P(\text{age} = \text{“<=30”} | \text{buys\_computer} = \text{“no”}) = 3/5 = 0.6$
  
  $P(\text{income} = \text{“medium”} | \text{buys\_computer} = \text{“yes”}) = 4/9 = 0.444$
  
  $P(\text{income} = \text{“medium”} | \text{buys\_computer} = \text{“no”}) = 2/5 = 0.4$
  
  $P(\text{student} = \text{“yes”} | \text{buys\_computer} = \text{“yes”}) = 6/9 = 0.667$
  
  $P(\text{student} = \text{“yes”} | \text{buys\_computer} = \text{“no”}) = 1/5 = 0.2$
  
  $P(\text{credit\_rating} = \text{“fair”} | \text{buys\_computer} = \text{“yes”}) = 6/9 = 0.667$
  
  $P(\text{credit\_rating} = \text{“fair”} | \text{buys\_computer} = \text{“no”}) = 2/5 = 0.4$

- $X = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair})$

  $P(X | \text{buys\_computer} = \text{“yes”}) = P(\text{age} = \text{“<=30”} | \text{buys\_computer} = \text{“yes”}) \times P(\text{income} = \text{“medium”} | \text{buys\_computer} = \text{“yes”}) \times P(\text{student} = \text{“yes”} | \text{buys\_computer} = \text{“yes”}) \times P(\text{credit\_rating} = \text{“fair”} | \text{buys\_computer} = \text{“yes”}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$

  $P(X | \text{buys\_computer} = \text{“no”}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$
Naïve Bayes Classifier: An Example

- P(C_i): P(buys_computer = “yes”) = 9/14 = 0.643
  P(buys_computer = “no”) = 5/14 = 0.357

- Compute P(X_i | C_i) for each class
  P(age = “<=30” | buys_computer = “yes”) = 2/9 = 0.222
  P(age = “<= 30” | buys_computer = “no”) = 3/5 = 0.6
  P(income = “medium” | buys_computer = “yes”) = 4/9 = 0.444
  P(income = “medium” | buys_computer = “no”) = 2/5 = 0.4
  P(student = “yes” | buys_computer = “yes”) = 6/9 = 0.667
  P(student = “yes” | buys_computer = “no”) = 1/5 = 0.2
  P(credit_rating = “fair” | buys_computer = “yes”) = 6/9 = 0.667
  P(credit_rating = “fair” | buys_computer = “no”) = 2/5 = 0.4

- X = (age <= 30, income = medium, student = yes, credit_rating = fair)
  P(X | C_i): P(X | buys_computer = “yes”) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
  P(X | buys_computer = “no”) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

P(X | C_i) * P(C_i): P(X | buys_computer = “yes”) * P(buys_computer = “yes”) = 0.028
P(X | buys_computer = “no”) * P(buys_computer = “no”) = 0.007
Naïve Bayes Classifier: An Example

- **P(C_i):**
  - $P(buys\_computer = \text{“yes”}) = 9/14 = 0.643$
  - $P(buys\_computer = \text{“no”}) = 5/14 = 0.357$

- Compute $P(X_i | C_i)$ for each class
  - $P(\text{age} = \text{“<=30”} | buys\_computer = \text{“yes”}) = 2/9 = 0.222$
  - $P(\text{age} = \text{“<=30”} | buys\_computer = \text{“no”}) = 3/5 = 0.6$
  - $P(\text{income} = \text{“medium”} | buys\_computer = \text{“yes”}) = 4/9 = 0.444$
  - $P(\text{income} = \text{“medium”} | buys\_computer = \text{“no”}) = 2/5 = 0.4$
  - $P(\text{student} = \text{“yes”} | buys\_computer = \text{“yes”}) = 6/9 = 0.667$
  - $P(\text{student} = \text{“yes”} | buys\_computer = \text{“no”}) = 1/5 = 0.2$
  - $P(\text{credit\_rating} = \text{“fair”} | buys\_computer = \text{“yes”}) = 6/9 = 0.667$
  - $P(\text{credit\_rating} = \text{“fair”} | buys\_computer = \text{“no”}) = 2/5 = 0.4$

- $X = (\text{age <= 30, income = medium, student = yes, credit\_rating = fair})$
  - $P(X | C_i) : P(X | buys\_computer = \text{“yes”}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$
  - $P(X | buys\_computer = \text{“no”}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$
  - $P(X | C_i) \times P(C_i) : P(X | buys\_computer = \text{“yes”}) \times P(buys\_computer = \text{“yes”}) = 0.028$
  - $P(X | buys\_computer = \text{“no”}) \times P(buys\_computer = \text{“no”}) = 0.007$

Since Red > Blue here, $X$ belongs to class (“buys\_computer = yes”)

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31...40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31...40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>31...40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31...40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>
Avoiding the Zero-Probability Problem

- Naïve Bayesian prediction requires each conditional prob. be non-zero. Otherwise, the predicted prob. will be zero.

\[ P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \]

- Ex. Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)

- Use **Laplacian correction** (or Laplacian estimator)
  - Adding 1 to each case
    - Prob(income = low) = 1/1003
    - Prob(income = medium) = 991/1003
    - Prob(income = high) = 11/1003
  - The “corrected” prob. estimates are close to their “uncorrected” counterparts
Naïve Bayes Classifier: Comments

- **Advantages**
  - Easy to implement
  - Good results obtained in most of the cases

- **Disadvantages**
  - Assumption: class conditional independence, therefore loss of accuracy
  - Practically, dependencies exist among variables
    - E.g., hospitals: patients: Profile: age, family history, etc.
      - Symptoms: fever, cough etc., Disease: lung cancer, diabetes, etc.
    - Dependencies among these cannot be modeled by Naïve Bayes Classifier

- How to deal with these dependencies? Bayesian Belief Networks (Chapter 9)
Classification: Basic Concepts

- Classification: Basic Concepts
- Decision Tree Induction
- Model Evaluation and Selection
- Practical Issues of Classification
- Bayes Classification Methods
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Summary
Ensemble Methods: Increasing the Accuracy

- **Ensemble methods**
  - Use a combination of models to increase accuracy
  - Combine a series of $k$ learned models, $M_1, M_2, \ldots, M_k$, with the aim of creating an improved model $M^*$
Ensemble Methods: Increasing the Accuracy

- **Ensemble methods**
  - Use a combination of models to increase accuracy
  - Combine a series of $k$ learned models, $M_1, M_2, \ldots, M_k$, with the aim of creating an improved model $M^*$

- **Popular ensemble methods**
  - **Bagging**: averaging the prediction over a collection of classifiers
  - **Boosting**: weighted vote with a collection of classifiers
  - **Random forests**: Imagine that each of the classifiers in the ensemble is a decision tree classifier so that the collection of classifiers is a “forest”
Classification of Class-Imbalanced Data Sets

- **Class-imbalance problem**: Rare positive example but numerous negative ones, e.g., medical diagnosis, fraud, oil-spill, fault, etc.
- Traditional methods assume a balanced distribution of classes and equal error costs: not suitable for class-imbalanced data

- Typical methods in two-class classification:
  - **Oversampling**: re-sampling of data from positive class
  - **Under-sampling**: randomly eliminate tuples from negative class
  - **Threshold-moving**: move the decision threshold, \( t \), so that the rare class tuples are easier to classify, and hence, less chance of costly false negative errors
  - **Ensemble techniques**: Ensemble multiple classifiers

- Still difficult for class imbalance problem on multiclass tasks
Other Issues

- Data Fragmentation
- Search Strategy
- Expressiveness
Data Fragmentation

- Number of instances gets smaller as you traverse down the tree

- Number of instances at the leaf nodes could be too small to make any statistically significant decision
Finding an optimal decision tree is NP-hard

The algorithm presented so far uses a greedy, top-down, recursive partitioning strategy to induce a reasonable solution

Other strategies?

- Bottom-up
- Bi-directional
Expressiveness

- Decision tree provides expressive representation for learning discrete-valued function
  - But they do not generalize well to certain types of Boolean functions
    - Example: XOR or Parity functions (example in book)

- Not expressive enough for modeling continuous variables
  - Particularly when test condition involves only a single attribute at-a-time
Expressiveness: Oblique Decision Trees

- Test condition may involve multiple attributes
- More expressive representation
- Finding optimal test condition is computationally expensive
- Needs multi-dimensional discretization

\[ x + y < 1 \]
Bagging: Bootstrap Aggregation

- **Analogy**: Diagnosis based on multiple doctors’ majority vote
- **Training**
  - Given a set D of d tuples, at each iteration i, a training set Di of d tuples is sampled with replacement from D (i.e., bootstrap)
  - A classifier model Mi is learned for each training set Di
- **Classification**: classify an unknown sample X
  - Each classifier Mi returns its class prediction
  - The bagged classifier M* counts the votes and assigns the class with the most votes to X
- **Prediction**: can be applied to the prediction of continuous values by taking the average value of each prediction for a given test tuple
- **Accuracy**: Proved improved accuracy in prediction
  - Often significantly better than a single classifier derived from D
  - For noise data: not considerably worse, more robust
Boosting

- Analogy: Consult several doctors, based on a combination of weighted diagnoses—weight assigned based on the previous diagnosis accuracy

- How boosting works?
  - **Weights** are assigned to each training tuple
  - A series of $k$ classifiers is iteratively learned
  - After a classifier $M_i$ is learned, the weights are updated to allow the subsequent classifier, $M_{i+1}$, to **pay more attention to the training tuples that were misclassified** by $M_i$
  - The final $M^*$ **combines the votes** of each individual classifier, where the weight of each classifier's vote is a function of its accuracy

- Boosting algorithm can be extended for numeric prediction

- Comparing with bagging: Boosting tends to have greater accuracy, but it also risks overfitting the model to misclassified data
Adaboost (Freund and Schapire, 1997)

- Given a set of \( d \) class-labeled tuples, \((X_1, y_1), \ldots, (X_d, y_d)\)
- Initially, all the weights of tuples are set the same (1/d)
- Generate \( k \) classifiers in \( k \) rounds. At round \( i \),
  - Tuples from \( D \) are sampled (with replacement) to form a training set \( D_i \) of the same size
  - Each tuple’s chance of being selected is based on its weight
  - A classification model \( M_i \) is derived from \( D_i \)
  - Its error rate is calculated using \( D_i \) as a test set
  - If a tuple is misclassified, its weight is increased, o.w. it is decreased
- Error rate: \( \text{err}(X_j) \) is the misclassification error of tuple \( X_j \). Classifier \( M_i \) error rate is the sum of the weights of the misclassified tuples:
  \[
  \text{error}(M_i) = \sum_j w_j \times \text{err}(X_j)
  \]
- The weight of classifier \( M_i \)'s vote is
  \[
  \log \frac{1 - \text{error}(M_i)}{\text{error}(M_i)}
  \]
Random Forest (Breiman 2001)

- **Random Forest:**
  - Each classifier in the ensemble is a decision tree classifier and is generated using a random selection of attributes at each node to determine the split.
  - During classification, each tree votes and the most popular class is returned.

- **Two Methods to construct Random Forest:**
  - Forest-RI (*random input selection*): Randomly select, at each node, F attributes as candidates for the split at the node. The CART methodology is used to grow the trees to maximum size.
  - Forest-RC (*random linear combinations*): Creates new attributes (or features) that are a linear combination of the existing attributes (reduces the correlation between individual classifiers).

- Comparable in accuracy to Adaboost, but more robust to errors and outliers.
- Insensitive to the number of attributes selected for consideration at each split, and faster than bagging or boosting.