Classification
Part 4

Dr. Sanjay Ranka
Professor
Computer and Information Science and Engineering
University of Florida, Gainesville
Model Evaluation

• Metrics for Performance Evaluation
  – How to evaluate the performance of a model

• Methods for Performance Evaluation
  – How to obtain reliable estimates

• Methods for Model Comparison
  – How to compare the relative performance among competing models
Metrics for Performance Evaluation

• Focus on the predictive capability of a model
  – Rather than how fast it takes to classify or build models, scalability, etc.

• Confusion Matrix:

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class=Yes</td>
<td>Class=Yes</td>
<td>Class=No</td>
</tr>
<tr>
<td>Class=Yes</td>
<td>a</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>Class=No</td>
<td>c</td>
<td>d</td>
<td></td>
</tr>
</tbody>
</table>

a: TP (true positive)  
b: FN (false negative)  
c: FP (false positive)  
d: TN (true negative)
Metrics for Performance Evaluation

<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=Yes</td>
<td>B (FN)</td>
</tr>
<tr>
<td>Class=No</td>
<td>C (FP)</td>
</tr>
<tr>
<td>Class=No</td>
<td>d (TN)</td>
</tr>
</tbody>
</table>

- Most widely-used metric:

\[
\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}
\]
Cost Matrix

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
<th>Class=Yes</th>
<th>Class=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class=Yes</td>
<td>C(Yes</td>
<td>Yes)</td>
<td>C(No</td>
</tr>
<tr>
<td>Class=No</td>
<td>C(Yes</td>
<td>No)</td>
<td>C(No</td>
</tr>
</tbody>
</table>

C(i | j): Cost of misclassifying class j example as class i

- Accuracy is a useful measure if
  - \(C(Yes | No) = C(No | Yes)\) and \(C(Yes | Yes) = C(No | No)\)
  - \(P(Yes) = P(No)\) (class distribution are equal)
## Cost vs. Accuracy

### Cost Matrix

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C(i</td>
<td>j)</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>-1</td>
</tr>
<tr>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

---

### Model M₁

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C(i</td>
<td>j)</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>150</td>
</tr>
<tr>
<td>-</td>
<td>60</td>
</tr>
</tbody>
</table>

**Accuracy = 80%**  
**Cost = 3910**

### Model M₂

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C(i</td>
<td>j)</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>250</td>
</tr>
<tr>
<td>-</td>
<td>5</td>
</tr>
</tbody>
</table>

**Accuracy = 90%**  
**Cost = 4255**

---

Data Mining, Sanjay Ranka  Spring 2011
Cost-Sensitive Measures

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

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

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

- Precision is biased towards C(Yes \mid Yes) & C(Yes \mid No)
- Recall is biased towards C(Yes \mid Yes) & C(No \mid Yes)
- F\text{-}measure is biased towards all except C(No \mid No)

Weighted Accuracy = \frac{w_1a + w_4d}{w_1a + w_2b + w_3c + w_4d}
Methods for Performance Evaluation

• How to obtain a reliable estimate of performance

• Performance of a model may depend on other factors besides the learning algorithm:
  – Class distribution
  – Cost of misclassification
  – Size of training and test sets
Learning Curve

- Learning curve shows how accuracy changes with varying sample size
- Requires a sampling schedule for creating a learning curve
  - Arithmetic sampling
  - Geometric sampling
- Effect of small sample size
  - Bias in the estimate
  - Variance of the estimate
Methods for Estimation

• Holdout
  – Reserve 2/3 for training and 1/3 for testing

• Random sub-sampling
  – Repeated holdout

• Cross validation
  – Partition data into k disjoint subsets
  – k-fold: train on k-1 partitions, test on the remaining one
  – Leave-one-out: k=n

• Stratified sampling
  – Over-sampling vs. Under-sampling

• Bootstrap
  – Sampling with replacement
Receiver Operating Characteristic (ROC)

• Developed in 1950s for signal detection theory to analyze noisy signals
  – Characterize the trade-off between positive hits and false alarms

• ROC curve plots TP (on the y-axis) against FP (on the x-axis)

• Performance of each classifier represented as a point on the ROC curve
  – changing the threshold of algorithm, sample distribution or cost matrix changes the location of the point
ROC Curve

- 1-dimensional data set containing 2 classes (positive and negative)
- Any point located at \( x > t \) is classified as positive

At threshold \( t \):
\[
TP=0.5, \ FN=0.5, \ FP=0.12, \ FN=0.88
\]
ROC Curve

(TP,FP):
• (0,0): declare everything to be negative class
• (1,1): declare everything to be positive class
• (1,0): ideal

• Diagonal line:
  – Random guessing
  – Below diagonal line:
    • prediction is opposite of the true class
Using ROC for Model Comparison

- No model consistently outperforms the other
  - M1 is better for small FPR
  - M2 is better for large FPR
- Area under the ROC curve
  - Ideal, area = 1
  - Random guess, area = 0.5