Data Mining I

Outlook: Evaluation
### Accuracy and Error Rate

Most widely-used metrics:

- **Accuracy**
  
  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Error Rate**
  
  $$\text{Error Rate} = 1 - \text{Accuracy}$$
Limitation of Accuracy: Unbalanced Data

- Sometimes, classes have very unequal frequency
  - Fraud detection: 98% transactions OK, 2% fraud
  - eCommerce: 99% don’t buy, 1% buy
  - Intruder detection: 99.99% of the users are no intruders
  - Security: >99.99% of Americans are not terrorists

- The class of interest is commonly called the **positive class**, and the rest **negative classes**.

- Consider a 2-class problem
  - Number of Class 0 examples = 9990, Number of Class 1 examples = 10
  - If 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
Precision and Recall

Alternative: Use measures from information retrieval which are biased towards the positive class.

<table>
<thead>
<tr>
<th></th>
<th>Classified Positive</th>
<th>Classified Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Actual Negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

\[ p = \frac{TP}{TP + FP} \quad \text{Precision} \]

\[ r = \frac{TP}{TP + FN} \quad \text{Recall} \]

Precision \( p \) is the number of \textbf{correctly classified positive examples} divided by the total number of examples that are classified as positive.

Recall \( r \) is the number of \textbf{correctly classified positive examples} divided by the total number of actual positive examples in the test set.
Performance

- Standard Measures
  - Accuracy
  - Precision
  - Recall

- Task Specific
  - Misclassification Costs