Classifier Evaluation
Today

• Evaluation Measures
  – Accuracy
  – Significance Testing
  – Cross-validation
  – F-Measure
  – Error Types
    • ROC Curves
    • Equal Error Rate
  – AIC/BIC
How do you know that you have a good classifier?

• Is a feature contributing to overall performance?
• Is classifier A better than classifier B?
• Internal Evaluation:
  – Measure the performance of the classifier.
• External Evaluation:
  – Measure the performance on a downstream task
Accuracy

- Easily the most common and intuitive measure of classification performance.

\[ \text{Accuracy} = \frac{\# \text{correct}}{N} \]
Significance testing

• Say I have two classifiers.

• A = 50% accuracy
• B = 75% accuracy

• B is better, right?
Significance Testing

• Say I have another two classifiers

• A = 50% accuracy
• B = 50.5% accuracy

• Is B better?
Basic Evaluation

- Training data – used to identify model parameters
- Testing data – used for evaluation
- Optionally: Development / tuning data – used to identify model hyperparameters.

- Difficult to get significance or confidence values
Cross validation

• Identify \( n \) “folds” of the available data.
• Train on \( n-1 \) folds
• Test on the remaining fold.

• In the extreme \((n=N)\) this is known as “leave-one-out” cross validation

• \( n \)-fold cross validation (xval) gives \( n \) samples of the performance of the classifier.
Cross-validation visualized

Available Labeled Data

Identify n partitions

Fold 1
Train  Train  Train  Train  Dev  Test
Cross-validation visualized

Available Labeled Data

Identify n partitions

Test Train Train Train Train Dev

Fold 2
Cross-validation visualized

Available Labeled Data

Identify n partitions

Fold 3
Dev  Test  Train  Train  Train  Train
Cross-validation visualized

Available Labeled Data

Identify n partitions

Fold 4

Train  Dev  Test  Train  Train  Train
Cross-validation visualized

Available Labeled Data

Identify n partitions

Fold 5

Train  Train  Dev  Test  Train  Train
Cross-validation visualized

Available Labeled Data

Identify n partitions

Fold 6
Train, Train, Train, Dev, Test, Train

Calculate Average Performance
Some criticisms of cross-validation

• While the test data is independently sampled, there is a lot of overlap in training data.
  – The model performance may be correlated.
  – Underestimation of variance.
  – Overestimation of significant differences.

• One proposed solution is to repeat 2-fold cross-validation 5 times rather than 10-fold cross validation
Significance Testing

• Is the performance of two classifiers different with statistical significance?
• Means testing
  – If we have two samples of classifier performance (accuracy), we want to determine if they are drawn from the same distribution (no difference) or two different distributions.
T-test

- One Sample t-test

\[ t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \]

- Independent t-test
  - Unequal variances and sample sizes

\[ t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1-\bar{X}_2}} \]

\[ s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \]

Once you have a t-value, look up the significance level on a table, keyed on the t-value and degrees of freedom.
Significance Testing

- Run Cross-validation to get n-samples of the classifier mean.
- Use this distribution to compare against either:
  - A known (published) level of performance
    - one sample t-test
  - Another distribution of performance
    - two sample t-test
- If at all possible, results should include information about the variance of classifier performance.
Significance Testing

• Caveat – including more samples of the classifier performance can artificially inflate the significance measure.
  – If $x$ and $s$ are constant (the sample represents the population mean and variance) then raising $n$ will increase $t$.
  – If these samples are real, then this is fine. Often cross-validation fold assignment is not truly random. Thus subsequent xval runs only resample the same information.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
Confidence Bars

• Variance information can be included in plots of classifier performance to ease visualization.

\[ \mu = 10 \quad \sigma = 1 \quad n = 10 \]

• Plot standard deviation, standard error or confidence interval?

\[ SD = \sigma \quad SE = \frac{\sigma}{\sqrt{n}} \]

\[ CI_{95\%} = \mu \pm 1.96 \times \frac{\sigma}{\sqrt{n}} \]
Confidence Bars

- Most important to be clear about what is plotted.
- 95% confidence interval has the clearest interpretation.
Baseline Classifiers

• Majority Class baseline
  – Every data point is classified as the class that is most frequently represented in the training data

• Random baseline
  – Randomly assign one of the classes to each data point.
    • with an even distribution
    • with the training class distribution
Problems with accuracy

- Contingency Table

<table>
<thead>
<tr>
<th>Hyp Values</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]
Problems with accuracy

• Information Retrieval Example
  – Find the 10 documents related to a query in a set of 100 documents

<table>
<thead>
<tr>
<th>Hyp Values</th>
<th>True Values</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

Accuracy = 90%
Problems with accuracy

- **Precision**: how many hypothesized events were true events

- **Recall**: how many of the true events were identified

- **F-Measure**: Harmonic mean of precision and recall

\[
P = \frac{TP}{TP + FP} \quad \quad R = \frac{TP}{TP + FN} \quad \quad F = \frac{2PR}{P + R}
\]

<table>
<thead>
<tr>
<th>True Values</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyp Values</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>
F-Measure

• F-measure can be weighted to favor Precision or Recall
  – beta > 1 favors recall
  – beta < 1 favors precision

\[ F_\beta = \frac{(1 + \beta^2)PR}{(\beta^2 P) + R} \]

<table>
<thead>
<tr>
<th>Hyp Values</th>
<th>True Values</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

\[ P = 0 \]
\[ R = 0 \]
\[ F_1 = 0 \]
# F-Measure

\[ F_\beta = \frac{(1 + \beta^2)PR}{(\beta^2 P) + R} \]

<table>
<thead>
<tr>
<th>Hyp Values</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>9</td>
<td>90</td>
</tr>
</tbody>
</table>

\[ P = 1 \]
\[ R = \frac{1}{10} \]
\[ F_1 = .18 \]
\[ F_\beta = \frac{(1 + \beta^2)PR}{(\beta^2 P) + R} \]

<table>
<thead>
<tr>
<th>Hyp Values</th>
<th>True Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>10</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
</tr>
<tr>
<td>Positive</td>
<td>50</td>
</tr>
<tr>
<td>Negative</td>
<td>40</td>
</tr>
</tbody>
</table>

\[ P = \frac{10}{60} \]
\[ R = 1 \]
\[ F_1 = .29 \]
### F-Measure

\[
F_\beta = \frac{(1 + \beta^2)PR}{(\beta^2 P) + R}
\]

<table>
<thead>
<tr>
<th>Hyp Values</th>
<th>True Values</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>1</td>
<td>89</td>
</tr>
</tbody>
</table>

\[
P = .9 \\
R = .9 \\
F_1 = .9
\]
F-Measure

- Accuracy is weighted towards majority class performance.
- F-measure is useful for measuring the performance on minority classes.
Types of Errors

- **False Positives**
  - The system predicted **TRUE** but the value was **FALSE**
  - aka “False Alarms” or Type I error

- **False Negatives**
  - The system predicted **FALSE** but the value was **TRUE**
  - aka “Misses” or Type II error
ROC curves

• It is common to plot classifier performance at a variety of settings or thresholds.
• Receiver Operating Characteristic (ROC) curves plot true positives against false positives.
• The overall performance is calculated by the Area Under the Curve (AUC).
Equal Error Rate (EER) is commonly reported.

EER represents the highest accuracy of the classifier.

Curves provide more detail about performance.

Gauvain et al. 1995