Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN
© The MIT Press, 2004

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml
CHAPTER 14:
Assessing and Comparing Classification Algorithms
Introduction

- Questions:
  - Assessment of the expected error of a learning algorithm: Is the error rate of 1-NN less than 2%?
  - Comparing the expected errors of two algorithms: Is $k$-NN more accurate than MLP?

- Training/validation/test sets
- Resampling methods: $K$-fold cross-validation
Algorithm Preference

- Criteria (Application-dependent):
  - Misclassification error, or risk (loss functions)
  - Training time/space complexity
  - Testing time/space complexity
  - Interpretability
  - Easy programmability
- Cost-sensitive learning
Resampling and K-Fold Cross-Validation

- The need for multiple training/validation sets \( \{X_i, V_i\}_i \): Training/validation sets of fold \( i \)
- \( K \)-fold cross-validation: Divide \( X \) into \( k \), \( X_i, i=1,\ldots,K \)
  \[
  \begin{align*}
  V_1 &= X_1 & T_1 &= X_2 \cup X_3 \cup \cdots \cup X_K \\
  V_2 &= X_2 & T_2 &= X_1 \cup X_3 \cup \cdots \cup X_K \\
  &\vdots & &\vdots \\
  V_K &= X_K & T_K &= X_1 \cup X_2 \cup \cdots \cup X_{K-1}
  \end{align*}
  \]
- \( T_i \) share \( K-2 \) parts
### 5x2 Cross-Validation

- 5 times 2 fold cross-validation (Dietterich, 1998)

<table>
<thead>
<tr>
<th>Fold</th>
<th>Training Set $T_k$</th>
<th>Validation Set $V_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$X_1^{(1)}$</td>
<td>$X_1^{(2)}$</td>
</tr>
<tr>
<td>2</td>
<td>$X_1^{(2)}$</td>
<td>$X_1^{(1)}$</td>
</tr>
<tr>
<td>3</td>
<td>$X_2^{(1)}$</td>
<td>$X_2^{(2)}$</td>
</tr>
<tr>
<td>4</td>
<td>$X_2^{(2)}$</td>
<td>$X_2^{(1)}$</td>
</tr>
<tr>
<td>5</td>
<td>$X_3^{(1)}$</td>
<td>$X_3^{(2)}$</td>
</tr>
<tr>
<td>6</td>
<td>$X_3^{(2)}$</td>
<td>$X_3^{(1)}$</td>
</tr>
<tr>
<td>7</td>
<td>$X_4^{(1)}$</td>
<td>$X_4^{(2)}$</td>
</tr>
<tr>
<td>8</td>
<td>$X_4^{(2)}$</td>
<td>$X_4^{(1)}$</td>
</tr>
<tr>
<td>9</td>
<td>$X_5^{(1)}$</td>
<td>$X_5^{(2)}$</td>
</tr>
<tr>
<td>10</td>
<td>$X_5^{(2)}$</td>
<td>$X_5^{(1)}$</td>
</tr>
</tbody>
</table>
Bootstrapping

- Draw instances from a dataset with replacement
- Prob that we do not pick an instance after N draws

\[
\left(1 - \frac{1}{N}\right)^N \approx e^{-1} = 0.368
\]

that is, only 36.8% is new!
# Measuring Error

<table>
<thead>
<tr>
<th>True Class</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>TP: True Positive</td>
</tr>
<tr>
<td></td>
<td>FP: False Positive</td>
</tr>
</tbody>
</table>

- **Error rate** = \# of errors / \# of instances = \( \frac{(FN+FP)}{N} \)
- **Recall** = \# of found positives / \# of positives = \( \frac{TP}{TP+FN} \) = **sensitivity** = hit rate
- **Precision** = \# of found positives / \# of found = \( \frac{TP}{TP+FP} \)
- **Specificity** = \( \frac{TN}{TN+FP} \)
- **False alarm rate** = \( \frac{FP}{FP+TN} \) = 1 - Specificity
**ROC Curve**

- Hit rate: \( \frac{|TP|}{|TP|+|FN|} \)
- False alarm rate: \( \frac{|FP|}{|FP|+|TN|} \)
- Sensitivity (Hit rate)
- Specificity = 1 - False alarm rate
Interval Estimation

- $X = \{ x^t \}_t$ where $x^t \sim \mathcal{N}(\mu, \sigma^2)$
- $m \sim \mathcal{N}(\mu, \sigma^2/N)$

\[
\sqrt{N} \frac{(m-\mu)}{\sigma} \sim Z
\]

\[
P\left\{-1.96 < \sqrt{N} \frac{(m-\mu)}{\sigma} < 1.96\right\} = 0.95
\]

\[
P\left\{m - 1.96 \frac{\sigma}{\sqrt{N}} < \mu < m + 1.96 \frac{\sigma}{\sqrt{N}}\right\} = 0.95
\]

\[
P\left\{m - z_{\alpha/2} \frac{\sigma}{\sqrt{N}} < \mu < m + z_{\alpha/2} \frac{\sigma}{\sqrt{N}}\right\} = 1 - \alpha
\]

$100(1- \alpha)$ percent confidence interval
When $\sigma^2$ is not known:

$$S^2 = \sum_{t} (x^t - m)^2 / (N - 1) \quad \frac{\sqrt{N}(m - \mu)}{S} \sim t_{N-1}$$

$$P\left\{ m - t_{\alpha/2,N-1} \frac{S}{\sqrt{N}} < \mu < m + t_{\alpha/2,N-1} \frac{S}{\sqrt{N}} \right\} = 1 - \alpha$$
Hypothesis Testing

- Reject a null hypothesis if not supported by the sample with enough confidence
- $\mathbf{X} = \{ x^t \}_t$ where $x^t \sim \mathcal{N}(\mu, \sigma^2)$

$$H_0: \mu = \mu_0 \text{ vs. } H_1: \mu \neq \mu_0$$

Accept $H_0$ with level of significance $\alpha$ if $\mu_0$ is in the 100(1 - $\alpha$) confidence interval

$$\frac{\sqrt{N}(m - \mu_0)}{\sigma} \in (-z_{\alpha/2}, z_{\alpha/2})$$

Two-sided test
<table>
<thead>
<tr>
<th>Truth</th>
<th>Accept</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct</td>
<td>Type I error</td>
</tr>
<tr>
<td>False</td>
<td>Type II error</td>
<td>Correct (Power)</td>
</tr>
</tbody>
</table>

- **One-sided test:** \( H_0: \mu \leq \mu_0 \) vs. \( H_1: \mu > \mu_0 \)
  Accept if
  \[
  \frac{\sqrt{N}(m - \mu_0)}{\sigma} \in (-\infty, Z_\alpha)
  \]

- **Variance unknown:** Use \( t \), instead of \( z \)
  Accept \( H_0: \mu = \mu_0 \) if
  \[
  \frac{\sqrt{N}(m - \mu_0)}{S} \in \left(-t_{\alpha/2,N-1}, t_{\alpha/2,N-1}\right)
  \]
Assessing Error:

\( H_0: p \leq p_0 \) vs. \( H_1: p > p_0 \)

- Single training/validation set: Binomial Test
  
  If error prob is \( p_0 \), prob that there are \( e \) errors or less in \( N \) validation trials is
  
  \[
  P\{X \leq e\} = \sum_{j=1}^{e} \binom{N}{j} p_0^j (1 - p_0)^{N-j}
  \]

  Accept if this prob is less than \( 1 - \alpha \)
Normal Approximation to the Binomial

- Number of errors $X$ is approx $\mathcal{N}$ with mean $Np_0$ and var $Np_0(1-p_0)$

$$\frac{X - Np_0}{\sqrt{Np_0(1-p_0)}} \sim Z$$

Accept if this prob for $X = e$ is less than $z_{1-\alpha}$
Paired t Test

- Multiple training/validation sets
- \( x_i^t = 1 \) if instance \( t \) misclassified on fold \( i \)
- Error rate of fold \( i \):
  \[
  p_i = \frac{1}{N} \sum_{t=1}^{N} x_i^t
  \]
- With \( m \) and \( s^2 \) average and var of \( p_i \)
  we accept \( p_0 \) or less error if
  \[
  \frac{\sqrt{K(m - p_0)}}{s} \sim t_{K-1}
  \]
  is less than \( t_{\alpha,K-1} \)
Comparing Classifiers:

\[ H_0: \mu_0 = \mu_1 \text{ vs. } H_1: \mu_0 \neq \mu_1 \]

- Single training/validation set: McNemar’s Test

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(e_{00}):</td>
<td>Number of examples misclassified by both</td>
<td>(e_{01}): Number of examples misclassified by 1 but not 2</td>
</tr>
<tr>
<td>(e_{10}):</td>
<td>Number of examples misclassified by 2 but not 1</td>
<td>(e_{11}): Number of examples correctly classified by both</td>
</tr>
</tbody>
</table>

- Under \(H_0\), we expect 

\[
\frac{\left|e_{01} - e_{10}\right| - 1}{e_{01} + e_{10}}^2 \sim \chi^2_1
\]

Accept if \(< \chi^2_{\alpha,1}\)
**K-Fold CV Paired t Test**

- Use K-fold cv to get K training/validation folds
- $p_i^1, p_i^2$: Errors of classifiers 1 and 2 on fold $i$
- $p_i = p_i^1 - p_i^2$: Paired difference on fold $i$
- The null hypothesis is whether $p_i$ has mean 0

\[
H_0 : \mu = 0 \text{ vs. } H_0 : \mu \neq 0
\]

\[
m = \frac{\sum_{i=1}^{K} p_i}{K}, \quad s^2 = \frac{\sum_{i=1}^{K} (p_i - m)^2}{K - 1}
\]

\[
\frac{\sqrt{K(m - 0)}}{s} = \frac{\sqrt{K} \cdot m}{s} \sim t_{K-1} \text{ Accept if } \in (-t_{\alpha/2,K-1}, t_{\alpha/2,K-1})
\]
5×2 cv Paired t Test

- Use 5×2 cv to get 2 folds of 5 tra/val replications (Dietterich, 1998)
- $p_i^{(j)}$: difference btw errors of 1 and 2 on fold $j=1, 2$ of replication $i=1,...,5$
  \[\bar{p}_i = \left( p_i^{(1)} + p_i^{(2)} \right) / 2 \]
  \[s_i^2 = (p_i^{(1)} - \bar{p}_i)^2 + (p_i^{(2)} - \bar{p}_i)^2 \]
  \[\frac{p_i^{(1)}}{\sqrt{\sum_{i=1}^{5} s_i^2 / 5}} \sim t_5\]

Two-sided test: Accept $H_0: \mu_0 = \mu_1$ if in $(-t_{\alpha/2,5}, t_{\alpha/2,5})$
One-sided test: Accept $H_0: \mu_0 \leq \mu_1$ if $< t_{\alpha,5}$
$5 \times 2$ cv Paired F Test

$$\frac{\sum_{i=1}^{5} \sum_{j=1}^{2} (p_{ij})^2}{2 \sum_{i=1}^{5} s_i^2} \sim F_{10,5}$$

Two-sided test: Accept $H_0$: $\mu_0 = \mu_1$ if $< F_{\alpha,10,5}$
Comparing $L > 2$ Algorithms: Analysis of Variance (Anova)

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_L$

- Errors of $L$ algorithms on $K$ folds
  
  $X_{ij} \sim \mathcal{N}(\mu_j, \sigma^2), j = 1, \ldots, L, i = 1, \ldots, K$

- We construct two estimators to $\sigma^2$.
  
  One is valid if $H_0$ is true, the other is always valid.
  
  We reject $H_0$ if the two estimators disagree.
If $H_0$ is true:

\[ m_j = \sum_{i=1}^{K} \frac{X_{ij}}{K} \sim \mathcal{N}(\mu, \sigma^2 / K) \]

\[ m = \frac{\sum_{j=1}^{L} m_j}{L} \quad S^2 = \frac{\sum_{j} (m_j - m)^2}{L - 1} \]

Thus an estimator of $\sigma^2$ is $K \cdot S^2$, namely,

\[ \hat{\sigma}^2 = K \sum_{j=1}^{L} \frac{(m_j - m)^2}{L - 1} \]

\[ \sum_{j} \frac{(m_j - m)^2}{\sigma^2 / K} \sim \chi^2_{L-1} \quad SSb = K \sum_{j} (m_j - m)^2 \]

So when $H_0$ is true, we have

\[ \frac{SSb}{\sigma^2} \sim \chi^2_{L-1} \]
Regardless of $H_0$ our second estimator to $\sigma^2$ is the average of group variances $S_j^2$:

$$S_j^2 = \frac{\sum_{i=1}^{K} (X_{ij} - m_j)^2}{K - 1} \quad \hat{\sigma}^2 = \sum_{j=1}^{L} \frac{S_j^2}{L} = \sum_{j} \sum_{i} \frac{(X_{ij} - m_j)^2}{L(K - 1)}$$

$$SSw \equiv \sum_{j} \sum_{i} (X_{ij} - m_j)^2$$

$$\left( K - 1 \right) \frac{S_j^2}{\sigma^2} \sim \chi^2_{K-1} \quad \frac{SSw}{\sigma^2} \sim \chi^2_{L(K-1)}$$

$$\left( \frac{SSb / \sigma^2}{L - 1} \right) / \left( \frac{SSw / \sigma^2}{L(K - 1)} \right) = \frac{SSb / (L - 1)}{SSw / (L(K - 1))} \sim F_{L-1,L(K-1)}$$

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_L$ if $< F_{\alpha,L-1,L(K-1)}$
Other Tests

- Range test (Newman-Keuls): 1 4 5 2 3
- Nonparametric tests (Sign test, Kruskal-Wallis)
- Contrasts: Check if 1 and 2 differ from 3, 4, and 5
- Multiple comparisons require Bonferroni correction
  If there are $m$ tests, to have an overall significance of $\alpha$, each test should have a significance of $\alpha/m$.
- Regression: CLT states that the sum of iid variables from any distribution is approximately normal and the preceding methods can be used.
- Other loss functions?