CHAPTER 19:
Design and Analysis of Machine Learning Experiments
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
Factors and Response

![Diagram showing input, controllable factors, uncontrollable factors, and output]
Strategies of Experimentation

Response surface design for approximating and maximizing the response function in terms of the controllable factors
Guidelines for ML experiments

A. Aim of the study
B. Selection of the response variable
C. Choice of factors and levels
D. Choice of experimental design
E. Performing the experiment
F. Statistical Analysis of the Data
G. Conclusions and Recommendations
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 \(k\), \(X_i, i=1,\ldots,K\)

\[
\begin{align*}
\mathcal{V}_1 &= X_1 & \mathcal{T}_1 &= X_2 \cup X_3 \cup \cdots \cup X_K \\
\mathcal{V}_2 &= X_2 & \mathcal{T}_2 &= X_1 \cup X_3 \cup \cdots \cup X_K \\
\vdots \\
\mathcal{V}_K &= X_K & \mathcal{T}_K &= X_1 \cup X_2 \cup \cdots \cup X_{K-1}
\end{align*}
\]

- \(T_i\) share \(K-2\) parts
5×2 Cross-Validation

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

\[
\begin{align*}
\mathcal{T}_1 &= \mathcal{X}_1^{(1)} & \mathcal{V}_1 &= \mathcal{X}_1^{(2)} \\
\mathcal{T}_2 &= \mathcal{X}_1^{(2)} & \mathcal{V}_2 &= \mathcal{X}_1^{(1)} \\
\mathcal{T}_3 &= \mathcal{X}_2^{(1)} & \mathcal{V}_3 &= \mathcal{X}_2^{(2)} \\
\mathcal{T}_4 &= \mathcal{X}_2^{(2)} & \mathcal{V}_4 &= \mathcal{X}_2^{(1)} \\
\vdots & & \\
\mathcal{T}_9 &= \mathcal{X}_5^{(1)} & \mathcal{V}_9 &= \mathcal{X}_5^{(2)} \\
\mathcal{T}_{10} &= \mathcal{X}_5^{(2)} & \mathcal{V}_{10} &= \mathcal{X}_5^{(1)}
\end{align*}
\]
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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>TP: True Positive</td>
<td>FN: False Negative</td>
</tr>
<tr>
<td>No</td>
<td>FP: False Positive</td>
<td>TN: True Negative</td>
</tr>
</tbody>
</table>

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

Hit rate: $\frac{TP}{(TP + FN)}$

False alarm rate: $\frac{FP}{(FP + TN)}$

Sensitivity (Hit rate)

Specificity = 1 - False alarm rate
(a) Example ROC curve

(b) Different ROC curves for different classifiers
Precision and Recall

\[
\text{Precision: } \frac{a}{a + b}
\]

\[
\text{Recall: } \frac{a}{a + c}
\]

(a) Precision and recall

(b) Precision = 1

(c) Recall = 1
Interval Estimation

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

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

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

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

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

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

\[
P\left\{ \sqrt{N} \frac{(m-\mu)}{\sigma} < 1.64 \right\} = 0.95
\]
\[
P\left\{ m - 1.64 \frac{\sigma}{\sqrt{N}} < \mu \right\} = 0.95
\]
\[
P\left\{ m - z_\alpha \frac{\sigma}{\sqrt{N}} < \mu \right\} = 1 - \alpha
\]

\[
S^2 = \sum_t (x^t - m)^2 / (N - 1)
\]
\[
\frac{\sqrt{N}(m-\mu)}{S} \sim t_{N-1}
\]
\[
P\left\{ m - t_{\alpha/2, N-1} \frac{S}{\sqrt{N}} < \mu < m + t_{\alpha/2, N-1} \frac{S}{\sqrt{N}} \right\} = 1 - \alpha
\]
Hypothesis Testing

- Reject a null hypothesis if not supported by the sample with enough confidence
- \( X = \{ x^t \}_t \) where \( x^t \sim N(\mu, \sigma^2) \)
  
  \[ H_0: \mu = \mu_0 \text{ vs. } H_1: \mu \neq \mu_0 \]

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
### Decision Table

<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
  \[
  \frac{\sqrt{N}(m - \mu_0)}{S} \in (-t_{\alpha/2,N-1}, t_{\alpha/2,N-1})
  \]
Assessing Error: $H_0: \rho \leq \rho_0$ vs. $H_1: \rho > \rho_0$

- Single training/validation set: Binomial Test

If error prob is $\rho_0$, prob that there are $e$ errors or less in $N$ validation trials is

$$P\{X \leq e\} = \sum_{j=1}^{e} \binom{N}{j} \rho_0^j (1 - \rho_0)^{N-j}$$

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

- Number of errors $X$ is approx $N$ 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 $z_{1-\alpha}$
t Test

- Multiple training/validation sets
- $x_{ti} = 1$ if instance $t$ misclassified on fold $i$
- Error rate of fold $i$: 
  \[ p_i = \frac{\sum_{t=1}^{N} x_{ti}}{N} \]
- 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:

$H_0: \mu_0 = \mu_1$ vs. $H_1: \mu_0 \neq \mu_1$

- Single training/validation set: McNemar’s Test

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

$$\frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \sim \chi^2_1$$

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

- Use K-fold cv to get K training/validation folds
- $p_{i1}, p_{i2}$: Errors of classifiers 1 and 2 on fold $i$
- $p_i = p_{i1} - p_{i2}$: Paired difference on fold $i$
- The null hypothesis is whether $p_i$ has mean 0

$$H_0 : \mu = 0 \ vs. \ H_0 : \mu \neq 0$$

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

$$\frac{\sqrt{K(m - 0)}}{s} = \frac{\sqrt{K \cdot m}}{s} \sim t_{K-1} \quad \text{Accept if } \text{in} \left( -t_{\alpha/2, K-1}, t_{\alpha/2, K-1} \right)$$
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 $i=1,..,5$

$$
\bar{p}_i = \left( p_i^{(1)} + p_i^{(2)} \right) / 2 \\
\sigma_i^2 = \left( p_i^{(1)} - \bar{p}_i \right)^2 + \left( p_i^{(2)} - \bar{p}_i \right)^2 \\
\frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^5 s_i^2 / 5}} \sim t_5
$$

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 $< t_{\alpha,5}$
5×2 cv Paired $F$ Test

$$\sum_{i=1}^{5} \sum_{j=1}^{2} \left( p_i^{(j)} \right)^2 \sim F_{10,5}$$

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,

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

$$\sum_j \frac{(m_j - m)^2}{\sigma^2 / K} \sim \chi^2_{L-1} \quad SSb \equiv K \sum_j (m_j - m)^2$$

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 \( S_j^2 \):

\[
S_j^2 = \frac{\sum_{i=1}^{K} (x_{ij} - m_j)^2}{K - 1}
\]

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

\[
SSw \equiv \sum_{j} \sum_{i} (x_{ij} - m_j)^2
\]

\[
(K - 1) \frac{S_j^2}{\sigma^2} \sim \chi_{K-1}^2 \quad \frac{SSw}{\sigma^2} \sim \chi_{L(K-1)}^2
\]

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

\( H_0 : \mu_1 = \mu_2 = \cdots = \mu_L \) if \( \leq F_{\alpha, L-1, L(K-1)} \)
If ANOVA rejects, we do pairwise posthoc tests

\[ H_0 : \mu_i = \mu_j \text{ vs } H_1 : \mu_i \neq \mu_j \]

\[ t = \frac{m_i - m_j}{\sqrt{\frac{2\sigma_w}{L-1}}} \sim t_{L(K-1)} \]
Comparison over Multiple Datasets

- Comparing two algorithms:
  
  **Sign test:** Count how many times \( A \) beats \( B \) over \( N \) datasets, and check if this could have been by chance if \( A \) and \( B \) did have the same error rate.

- Comparing multiple algorithms:
  
  **Kruskal-Wallis test:** Calculate the average rank of all algorithms on \( N \) datasets, and check if these could have been by chance if they all had equal error. If KW rejects, we do pairwise posthoc tests to find which ones have significant rank difference.