**Classifier Evaluation**

- How to estimate classifier performance.
- Learning curves
- Feature curves
- Rejects and ROC curves

**Learning Curve**

- True classification error $\varepsilon$
- Sub-optimal classifier
- Bayes error $\varepsilon^*$
- Bayes consistent classifier

**The Apparent Classification Error**

The apparent (or resubstitution error) of the training set is positively biased (optimistic).

An independent test set is needed!

**Error Estimation by Test Set**

- Design Set
- Other training set $\rightarrow$ other classifier
- Other test set $\rightarrow$ other error estimate

**Training Set Size $\leftrightarrow$ Test Set Size**

- Training set should be large for good classifiers.
- Test set should be large for a reliable, unbiased error estimate.
- In practice often just a single design set is given

**Cross-validation**

- Size test set $1/n$ of design set.
- Size training set is $(n-1)/n$ of design set.
- Train and test $n$ times. Average errors. (Good choice: $n = 10$)
- All objects are tested ones $\rightarrow$ most reliable test result that is possible.
- Final classifier: trained by all objects $\rightarrow$ best possible classifier.
- Error estimate is slightly pessimistically biased.
Leave-one-out Procedure

Cross-validation in which n is total number of objects.
One object tested at a time. n classifiers to be computed.
In general unfeasible for large n.
Doable for k-NN classifier (needs no training).

Expected Learning Curves by Estimated Errors

Expected Learning Curves by Estimated Errors

Averaged Learning Curve

For obtaining ‘theoretically expected’ curves many repetitions are needed.

Repeated Learning Curves

Small sample sizes have a very large variability.

Learning Curves for Different Classifier Complexity

More complex classifiers are better in case of large training sets and worse in case of small training sets.

Peaking Phenomenon, Overtraining Curse of Dimensionality, Rao’s Paradox

Classification error

Training set size

Feature set size (dimensionality)

Classifier complexity
Example Overtraining, Polynomial Classifier

Example Overtraining (2)

Example Overtraining (3)

Example Overtraining (4)

Overtraining ↔ Increasing Bias

Example Curse of Dimensionality

Overtraining
Increasing Bias

Classification error
True error
Apparent error

feature set size (dimensionality)
classifier complexity

Fisher classifier for various feature rankings
Confusion Matrix (1)

real labels
A: \[ \begin{bmatrix} a_1 & a_2 & a_3 \\ \end{bmatrix} \]

obtained labels
L: \[ \begin{bmatrix} l_1 & l_2 & l_3 \\ \end{bmatrix} \]

Confusion matrix:
\[ C = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \]

Confusion Matrix (2)

\( N_a = 10, N_b = 30, N_c = 20 \)

Test set classified to
class A: 6 23 1
class B: 4 1 15
class C: A 1

Error:
\[ E = \frac{c_{21} + c_{22} + c_{23} + c_{31} + c_{32} + c_{33}}{N_a + N_b + N_c} \]

\[ E = \frac{14}{60} = 0.233 \]

Confusion Matrix (3)

objects from
A: \[ \begin{bmatrix} a_1 & a_2 & a_3 \\ \end{bmatrix} \]

classified to
A: 8
B: 2
C: 0

-0.20 error in class A
-0.23 error in class B
-0.25 error in class C
0.228 averaged error

Classification details are only observable in the confusion matrix!!

Conclusions on Error Estimations

- Larger training sets yield better classifiers.
- Independent test sets are needed for obtaining unbiased error estimates.
- Larger test sets yield more accurate error estimates.
- Leave-one-out cross-validation seems to be an optimal compromise, but might be computationally infeasible.
- 10-fold cross-validation is a good practice.
- More complex classifiers need larger training sets to avoid overtraining.
- This holds in particular for larger feature sizes, due to the curse of dimensionality.
- For too small training sets, more simple classifiers or smaller feature sets are needed.
- Confusion matrices allow a detailed look at the per class classification.

Reject and ROC Analysis

- Reject Types
- Reject Curves
- Performance Measures
- Varying Costs and Priors
- ROC Analysis

Outlier Reject

Reject objects for which probability density is low:
\[ P(x) = P(x|A)P(A) + P(x|B)P(B) \]
\[ P(x) \approx P(x|A)P(A) \quad \text{for} \quad \delta(x) \gg 0 \]

Note: in these area the posterior probabilities might be high!
**Ambiguity Reject**

Reject objects for which classification is unsure: about equal posterior probabilities:

\[ P(A|x) = P(B|x) \]

\[ P(d|A)P(A) = P(d|B)P(B) \]

\[ P(z|A)P(A) - P(z|B)P(B) = 0 \]

\[ S(x) = 0 \]

**Reject Curve**

The classification error \( \varepsilon \), can be reduced to \( \varepsilon_r \) by rejecting a fraction \( r \) of the objects.

**Rejecting Classifier**

Original classifier

\[ S(x) > 0 \quad \Rightarrow \quad x \rightarrow A \]

\[ S(x) \leq 0 \quad \Rightarrow \quad x \rightarrow B \]

Rejecting classifier

\[ S(x) > d \quad \Rightarrow \quad x \rightarrow A \]

\[ S(x) \leq d \quad \Rightarrow \quad x \rightarrow B \]

\[ d \geq 0 \]

Rejected fraction \( r \)

\[ r = \frac{P(A) \int_{-d}^{0} P(z|A)dz + P(B) \int_{d}^{\infty} P(z|B)dz}{\int_{-d}^{\infty} P(z|A)dz + \int_{d}^{\infty} P(z|B)dz} \]

**Reject Fraction**

For reasonably well trained classifiers \( r > 2 \) (\( \varepsilon(r=0) - \varepsilon(r) \)).
To decrease the error by 1%, more than 2% has to be rejected.

**How much to reject?**

Compare the cost of a rejected object, \( c_r \), with the cost of a classification error, \( c_\varepsilon \):

\[ c = c_rP(\text{reject}) + c_\varepsilon P(\text{error}) \]

For given total cost \( c \) this is a linear function in the \((r, \varepsilon)\) space.
Shift it until a possible operating point is reached.

**Error / Performance Measures**

Objects of class A, with prior \( p(A) \), is classified by \( S(x) > 0 \) in a part \( \eta_A \) assigned to A, and a part \( c_\varepsilon \) assigned to B.

\[ p(A) = \eta_A + c_\varepsilon \]
**Error / Performance Measures**

Objects of class B, with prior \( p(B) \), is classified by \( S(x) = 0 \) in a part \( \eta_B \), assigned to B, and a part \( \epsilon_B \), assigned to A.

\[
S(x) > 0 \rightarrow \text{Class A} \\
S(x) = 0 \rightarrow \text{Class B}
\]

\( p(B) = \eta_B + \epsilon_B \)

\( \eta_B = \eta_B^1 + \eta_B^2 \)

\( \epsilon_B = \epsilon_B^1 + \epsilon_B^2 \)

\( \eta_A = \eta_A^1 + \eta_A^2 \)

\( \epsilon_A = \epsilon_A^1 + \epsilon_A^2 \)

\( p(A) = \eta_A + \epsilon_A \)

\( p(B) = \eta_B + \epsilon_B \)

\( \eta = \eta_A + \eta_B \)

\( \epsilon = \epsilon_A + \epsilon_B \)

\( p(A) + p(B) = 1 \)

\( p(A) = \frac{\eta_A}{\eta_A + \epsilon_A} \)

\( p(B) = \frac{\eta_B}{\eta_B + \epsilon_B} \)

**Sensitivity:**

\[
\frac{TP}{TP + FN} = \frac{\eta_A}{\eta_A + \epsilon_A}
\]

**Specificity:**

\[
\frac{TN}{FP + TN} = \frac{\eta_B}{\eta_B + \epsilon_B}
\]

**False Discovery Rate (FDR):**

\[
\frac{FP}{FP + TN} = \frac{\epsilon_A}{\eta_A + \epsilon_A}
\]

**ROC Curve**

For every classifier \( S(x) - d = 0 \) two errors, \( \epsilon_A \) and \( \epsilon_B \) are made. The ROC curve shows them as function of \( d \).

**Error / Performance Measures**

- **Error:** probability of erroneous classifications
- **Performance:** \( 1 - \text{error} \)
- **Sensitivity** of a target class (e.g. diseased patients): performance for objects from that target class.
- **Specificity** : performance for all objects outside the target class.
- **Precision** of a target class: fraction of correct objects among all objects assigned to that class.
- **Recall:** fraction of correctly classified objects. This is identical to the performance. It is also identical to the sensitivity when related to a particular class.

ROC is independent of class prior probabilities

Change of prior is analogous to the shift of the decision boundary.
**ROC Analysis**

ROC: Receiver-Operator Characteristic (from communication theory)

- Performance of target class
- Error in non-targets
- Error class A
- Error class B

**When are ROC Curves Useful? (1)**

- What is the performance for given cost?
- What are the cost of a preset performance?

**When are ROC Curves Useful? (2)**

- Study of the effect of changing priors

**When are ROC Curves Useful? (3)**

- Compare behavior of different classifiers

**Combine Classifiers**

Any 'virtual' classifier between $S_1(x)$ and $S_2(x)$ in the $(\epsilon_A, \epsilon_B)$ space can be realized by using at random $\alpha$ times $S_1(x)$ and $(1-\alpha)$ times $S_2(x)$.

- $\epsilon_A = \alpha \epsilon_1 + (1-\alpha) \epsilon_2$
- $\epsilon_B = \alpha \epsilon_3 + (1-\alpha) \epsilon_4$

**Summary Reject and ROC**

- Reject for solving ambiguity: reject objects close to the decision boundary → lower costs.
- Reject option for protection against outliers.
- ROC analysis for performance - cost trade-off.
- ROC analysis in case of unknown or varying priors.
- ROC analysis for comparing / combining classifiers.