Evaluating Classifiers

Lecture 2
Instructor: Max Welling
Evaluation

Given:
• Hypothesis h(x): X\(\rightarrow\)C, in hypothesis space H, mapping attributes x to classes c=[1,2,3,...C]
• A data-sample S(n) of size n.

Questions:
• What is the error of “h” on unseen data?
• If we have two competing hypotheses, which one is better on unseen data?
• How do we compare two learning algorithms in the face of limited data?
• How certain are we about our answers?
Sample and True Error

We can define two errors:

1) Error(h|S) is the error on the sample S:

\[
error(h \mid S) = \frac{1}{n} \sum_{i=1}^{n} [h(x_i) \neq y_i]
\]

2) Error(h|P) is the true error on the unseen data sampled from the distribution P(x):

\[
error(h \mid P) = \int dx \ P(x) \ [h(x) \neq f(x)]
\]

where f(x) is the true hypothesis.
Flow of Thought

• Determine the property you want to know about the future data (e.g. error(h|P))

• Find an unbiased estimator E for this quantity based on observing data X (e.g. error(h|X))

• Determine the distribution P(error) under the assumption you have infinitely many sample sets X1,X2,...of some size n.

• Estimate the parameters of P(error) from an actual data sample S.

• Compute mean and variance of P(error) and pray P(error) it is close to a Normal distribution. (sums of random variables converge to normal distributions – central limit theorem)

• State your confidence interval as: with confidence N%, error(h|P) is contained in the interval

\[ Y = \text{mean} \pm z_N \sqrt{\text{var}} \]
Assumptions

• We only consider discrete valued hypotheses (i.e. classes)

• Training data and test data are drawn IID from the same distribution $P(x)$. (IID: independently & identically distributed)

• The hypothesis must be chosen independently from the data sample $S$!

• When you obtain a hypothesis from a learning algorithm, split the data into a training set and a testing set. Find the hypothesis using the training set and estimate error on the testing set.
Paired Tests

• Consider the following data:

\[
\begin{align*}
\text{error}(h_1|s_1) &= 0.1 & \text{error}(h_2|s_1) &= 0.11 \\
\text{error}(h_1|s_2) &= 0.2 & \text{error}(h_2|s_2) &= 0.21 \\
\text{error}(h_1|s_3) &= 0.66 & \text{error}(h_2|s_3) &= 0.67 \\
\text{error}(h_1|s_4) &= 0.45 & \text{error}(h_2|s_4) &= 0.46
\end{align*}
\]
and so on.

• We have \( \text{var}(\text{error}(h_1)) = \text{large} \), \( \text{var}(\text{error}(h_2)) = \text{large} \).
  The total variance of \( \text{error}(h_1) - \text{error}(h_2) \) is their sum. 
  \textit{However, \( h_1 \) is consistently better than \( h_2 \).}

• We ignored the fact that we compare on the same data.
  We want a different estimator that compares data one by one.

• You can use a \textit{"paired t-test"} (e.g. in matlab) to see if the two errors are significantly different, or if one error is significantly larger than the other.
Paired t-test

• On each subset $T_1, \ldots, T_k$ with $|T_i| > 30$ compute: $\delta_i = \text{error}(h1 \mid T_i) - \text{error}(h2 \mid T_i)$

• Now compute:

$$\bar{\delta} = \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

$$s(\bar{\delta}) = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

• State: With $N\%$ confidence the difference in error between $h1$ and $h2$ is:

$$\bar{\delta} \pm t_{N,k-1} s(\bar{\delta})$$

• “$t$” is the t-statistic which is related to the student-t distribution.
Comparing Learning Algorithms

• In general it is a really bad idea to estimate error rates on the same data on which a learning algorithm is trained.

• So just as in x-validation, we split the data into k subsets: \( S \rightarrow \{T_1, T_2, \ldots, T_k\} \).

• Train both learning algorithm 1 (L1) and learning algorithm 2 (L2) on the complement of each subset: \( \{S-T_1, S-T_2, \ldots\} \) to produce hypotheses \( \{L_1(S-T_i), L_2(S-T_i)\} \) for all \( i \).

• Compute for all \( i \) : 
  \[ \delta_i = \text{error}(L_1(S-T_i) \mid T_i) - \text{error}(L_2(S-T_i) \mid T_i) \]

• Note: we train on \( S-T_i \), but test on \( T_i \).

• As in the last slide perform a paired t-test on these differences to compute an estimate and a confidence interval for the relative error of the hypothesis produced by L1 and L2.
Conclusion

Never (ever) draw error-curves without confidence intervals

(The second most important sentence of this course)