Machine Learning: Lecture 5

Experimental Evaluation of
Learning Algorithms

(Based on Chapter 5 of Mitchell T..,
Machine Learning, 1997)

Motivation

- Evaluating the performance of learning systems is important because:
  - Learning systems are usually designed to predict the class of "future" unlabeled data points.
  - In some cases, evaluating hypotheses is an integral part of the learning process (example, when pruning a decision tree)

Difficulties in Evaluating Hypotheses when only limited data are available

- ***Bias in the estimate:*** The observed accuracy of the learned hypothesis over the training examples is a poor estimator of its accuracy over future examples ==> we test the hypothesis on a test set chosen independently of the training set and the hypothesis.
- ***Variance in the estimate:*** Even with a separate test set, the measured accuracy can vary from the true accuracy, depending on the makeup of the particular set of test examples. The smaller the test set, the greater the expected variance.

Questions Considered

- Given the observed accuracy of a hypothesis over a limited sample of data, how well does this estimate its accuracy over additional examples?
- Given that one hypothesis outperforms another over some sample data, how probable is it that this hypothesis is more accurate, in general?
- When data is limited what is the best way to use this data to both learn a hypothesis and estimate its accuracy?

Estimating Hypothesis Accuracy

Two Questions of Interest:

- Given a hypothesis *h* and a data sample containing *n* examples drawn at random according to distribution *D*, what is the best estimate of the accuracy of *h* over future instances drawn from the same distribution? ==> ***sample*** vs. ***true error***
- What is the probable error in this accuracy estimate? ==> ***confidence intervals***

Sample Error and True Error

- **Definition 1:** The ***sample error*** (denoted $error_S(h)$) of hypothesis h with respect to target function f and data sample S is:

$$error_S(h) = 1/n \sum_{x \in S} \delta(f(x), h(x))$$

where n is the number of examples in S, and the quantity $\delta(f(x), h(x))$ is 1 if $f(x) \neq h(x)$, and 0, otherwise.

- **Definition 2:** The ***true error*** (denoted $error_D(h)$) of hypothesis h with respect to target function f and distribution D, is the probability that h will misclassify an instance drawn at random according to D.

$$error_D(h) = Pr_{x \in D}[f(x) \neq h(x)]$$

Confidence Intervals for Discrete-Valued Hypotheses

- The general expression for approximate $N\%$ ***confidence intervals*** for $error_D(h)$ is:

$$\boldsymbol{error_S(h) \pm z_N \sqrt{error_S(h)(1-error_S(h))/n}}$$

where Z_N is given in [Mitchell, table 5.1]

- This approximation is quite good when

$$n\, error_S(h)(1 - error_S(h)) \geq 5$$

Mean and Variance

- **Definition 1:** Consider a random variable Y that takes on possible values $y_1, \ldots, y_n$. The ***expected value*** (or ***mean value***) of Y, ***E[Y],*** is:
  $$E[Y] = \Sigma_{i=1}^{n} y_i \, Pr(Y=y_i)$$
- **Definition 2:** The ***variance*** of a random variable Y, ***Var[Y],*** is:
  $$Var[Y] = E[(Y-E[Y])^2]$$
- **Definition 3:** The ***standard deviation*** of a random variable Y is the square root of the variance.

Estimators, Bias and Variance

- Since $error_S(h)$ (an ***estimator*** for the true error) obeys a Binomial distribution (See, [Mitchell, Section 5.3]), we have: $\boldsymbol{error_S(h) = r/n}$ and $\boldsymbol{error_D(h) = p}$

  where n is the number of instances in the sample S, r is the number of instances from S misclassified by h, and p is the probability of misclassifying a single instance drawn from D.

- Definition: The ***estimation bias*** (≠ from the inductive bias) of an estimator Y for an arbitrary parameter p is

$$\boldsymbol{E[Y] - p}$$

- The ***standard deviation*** for $error_S(h)$ is given by

$$\boldsymbol{\sqrt{p(1-p)/n} \approx \sqrt{error_S(h)(1-error_S(h))/n}}$$

Difference in Error of two Hypotheses

- Let h_1 and h_2 be two hypotheses for some discrete-valued target function. H_1 has been tested on a sample S_1 containing n_1 randomly drawn examples, and h_2 has been tested on an independent sample S_2 containing n_2 examples drawn from the same distribution.
- Let's estimate the difference between the true errors of these two hypotheses, ***d***, by computing the difference between the sample errors: $\hat{d} = error_{S1}(h1) - error_{S2}(h2)$
- The approximate ***N%*** confidence interval for ***d*** is:

$$\hat{d} \pm Z_N \sqrt{error_{S1}(h_1)(1-error_{S1}(h_1))/n_1 + error_{S2}(h_2)(1-error_{S2}(h_2))/n_2}$$

Comparing Learning Algorithms

- Which of L_A and L_B is the better learning method on average for learning some particular target function f?
- To answer this question, we wish to estimate the expected value of the difference in their errors:
  $E_{S \subset D}[error_D(L_A(S)) - error_D(L_B(S))]$
- Of course, since we have only a limited sample D_0 we estimate this quantity by dividing D_0 into a ***training set*** S_0 and a ***testing set*** T_0 and measure:
  $error_{T0}(L_A(S_0)) - error_{T0}(L_B(S_0))$
- **Problem:** We are only measuring the difference in errors for one training set S_0 rather than the expected value of this difference over all samples S drawn from D

Solution: ***k-fold Cross-Validation***

k-Fold Cross-Validation

1. Partition the available data D_0 into k disjoint subsets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30.
2. For i from **1** to k, do

   use T_i for the test set, and the remaining data for training set S_i

   - $S_i \leftarrow \{D_0 - T_i\}$
   - $h_A \leftarrow L_A(S_i)$
   - $h_B \leftarrow L_B(S_i)$
   - $\delta i \leftarrow error_{Ti}(h_A) - error_{Ti}(h_B)$

3. Return the value $avg(\delta)$,

   where .

Confidence of the k-fold Estimate

- The approximate N% confidence interval for estimating $E_{S \subset D0}[error_D(L_A(S)) - error_D(L_B(S))]$ using $avg(\delta)$, is given by:

$$avg(\delta) \pm t_{N,k-1} s_{avg(\delta)}$$

where $t_{N,k-1}$ is a constant similar to Z_N (See [Mitchell, Table 5.6]) and $s_{avg(\delta)}$ is an estimate of the standard deviation of the distribution governing $avg(\delta)$

$$s_{avg(\delta))} = \sqrt{1/k(k-1) \sum_{i=1}^{k} (\delta_i - avg(\delta))^2}$$