Evaluating Hypothesis - Part 2

Section 5.4 – 5.7

Shen Wan

March 26, 2008

Review of Part 1

- Why do we want to evaluate accuracy of hypotheses?
- What is the accuracy of hypotheses?
  - Bias and Variance in the estimate
- Sample Error vs True Error
- Confidence Intervals and How to Calculate
  - 2-Sided vs 1-Sided Bounds

General Approach for Estimating

1. Identify estimatee.

2. Define estimator Y.

3. Determine distribution $\mathcal{D}_Y$, its mean and variance.

4. Determine the $N\%$ confidence interval.

Central Limit Theorem

- For n *iid* (independent, identically distributed) random variables $Y_1, \ldots, Y_n$ of arbitrary distribution with mean μ and variance σ^2, $\lim_{n\to\infty} \bar{Y} = \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$.

- The theorem says: summation or average of many *iid* is (almost) Normal distribution.

- When $n \geq 30$, it is considered large enough.

- The mean and variance of $\bar{Y}$ can be used to determine the mean and variance of Y_i.

Comparing Two Hypotheses

- Use $\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$ to estimate $\hat{d} \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$.

- When $|S_1| \geq 30$ and $|S_2| \geq 30$, $\hat{d}$ is approximately Normal with mean d and variance

$$\sigma_{\hat{d}}^2 \approx \sigma_{error_{S_1}(h_1)} + \sigma_{error_{S_2}(h_2)}.$$

- The approximate $N\%$ confidence interval is $\hat{d} \pm z_N \sigma_{\hat{d}}$.

- When $S_1 = S_2$, $\sigma_{\hat{d}}$ will usually be smaller.

Comparing Learning Algorithms

- We want to estimate

$$d \equiv E_{S \subset \mathcal{D}, |S|=n} \left[error_{\mathcal{D}} \left(L_A(S) \right) - error_{\mathcal{D}} \left(L_B(S) \right) \right].$$

- But we only have a limited sample $D_0 \subset \mathcal{D}$.

- So we need to divide D_0 into a training set S_0 and a disjoint test set T_0,

- and measure $error_{T_0} \left(L_A(S_0) \right) - error_{T_0} \left(L_B(S_0) \right)$.

k-fold Method

1. Partition D_0 into k disjoint subsets $T_1, \ldots, T_k$ of equal size. (≥ 30)

2. Train and test k times, using each subset T_i to test and all remaining $S_i = D_0 - T_i$ as training set.

$$\delta_i \leftarrow error_{T_i}(L_A(S_i)) - error_{T_i}(L_B(S_i))$$

3. The mean difference in errors $\bar{\delta} \equiv \frac{1}{k}\sum_{i=1}^{k}\delta_i$ is an estimator of d.

Analysis of k-fold Method

- $\overline{\delta}$ is an estimate of $E_{S \subset D_0}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$, where S is a random sample of size $\frac{k-1}{k}|D_0|$ drawn uniformly from D_0.

- The N% confidence interval is $\overline{\delta} \pm t_{N,k-1} s_{\overline{\delta}}$, where

$$s_{\overline{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \overline{\delta})^2}.$$

-

$$\lim_{k \to \infty} t_{N,k-1} = z_N$$

Paired Test

- We can actually evaluate the error of the learning algorithms independently, and then compare their estimated error rates.

- To evaluate hypotheses over identical samples: *paired test*.

- k-fold method is a *paired test*.

- Paired tests typically produce tighter confidence intervals.

Another Method: Randomly Partition

- Randomly choose a test set with more than 30 examples and use remaining examples for training.
- Repeat the proces until you are bored :)
- Pro: Able to shrink confidence interval to desired width.
- Con: Test sets not independently drawn with respect to $\mathcal{D}$.

Paired t Tests

- To estimate the sample mean of k *iid* Normal distributions.

- $\mu = \bar{Y} \pm t_{N,k-1} S_{\bar{Y}}$

- Where $S_{\bar{Y}}$ is the estimated standard deviation of the sample mean

$$S_{\bar{Y}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (Y_i - \bar{Y})^2}.$$