Chapter 5: Evaluating Hypotheses

CS 536: Machine Learning

Littman (Wu, TA)

Evaluating Hypotheses

[Read Ch. 5]

[Recommended exercises: 5.2, 5.3, 5.4]

- Sample error, true error
- Confidence intervals for observed hypothesis error
- Estimators
- Binomial distribution, Normal distribution, Central Limit Theorem
- Paired *t* tests
- Comparing learning methods

Two Definitions of Error

The **true error** 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*:

$$error_D(h) \equiv \Pr_{x \sim D}[f(x) \neq h(x)]$$
$$= E_{x \sim D}[\delta(f(x) \neq h(x)],$$

where $\delta(\phi)$ is 1 if ϕ is true, 0 otherwise.

The **sample error** of *h* with respect to target function *f* and data sample *S* is the proportion of examples *h* misclassifies:

$$error_S(h) \equiv 1/n\ \Sigma_{x \text{ in } S}\ \delta(f(x) \neq h(x))$$
$$= E_{x \sim S}[\delta(f(x) \neq h(x)].$$

Estimation Problem

We have $error_S(h)$.

We want to know $error_D(h)$.

How well does $error_S(h)$ estimate $error_D(h)$?

Problems Estimating Error

1. *Bias*: If S is training set, $error_S(h)$ is optimistically biased

$$bias \equiv E[error_S(h)] - error_D(h)$$

To ensure an unbiased ($bias = 0$) estimate, h and S must be chosen independently.

2. *Variance*: Even with unbiased S, $error_S(h)$ may still vary from $error_D(h)$.

To put this another way,

$$E[error_S(h)] - error_S(h) \neq 0.$$

Example

Hypothesis h misclassifies 12 of the 40 examples in S

$$error_S(h) = 12/40 = 0.3.$$

What is $error_D(h)$?

How sure are you?

Estimators

Experiment:

1. choose sample S of size n according to distribution D
2. measure $error_S(h)$

$error_S(h)$ is a random variable (that is, the result of an experiment)

$error_S(h)$ is an unbiased estimator for $error_D(h)$

Given observed $error_S(h)$, what can we conclude about $error_D(h)$?

Confidence Intervals

If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$

Then,

- With approximately 95% probability, $error_D(h)$ lies in interval

$error_S(h) \pm 1.96 \sqrt{(error_S(h)\,(1 - error_S(h))/n)}$

Confidence Intervals

General form: If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$

Then

- With approximately N% probability, $error_D(h)$ lies in interval

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

where

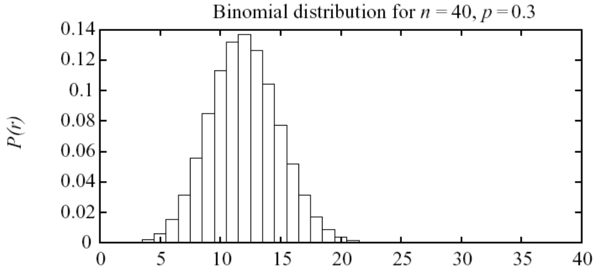
N%	50%	68%	80%	90%	**95%**	98%	99%
z_N	0.67	1.00	1.28	1.64	**1.96**	2.33	2.58

Sample Error is a Random Var.

Rerun the experiment with different randomly drawn S (of size n)

Probability of observing r misclassified examples:

$$P(r) = n! \,/\, (r!\,(n-r)!)\; error_D(h)^r \,(1 - error_D(h))^{n-r}$$



Binomial Probability Dist.

$$P(r) = n! \,/\, (r!\,(n-r)!)\; error_D(h)^r \,(1 - error_D(h))^{n-r}$$

Probability $P(r)$ of r heads in n coin flips, if $p = \Pr(heads)$

- Expected, or mean value of X, $E[X]$, is

$$E[X] \equiv \Sigma_{i=0}^{n} P(i) = np.$$

- Variance of X, σ_X^2 or $Var(X)$, is

$$Var(X) \equiv E[(X - E[X])]^2 = np(1-p).$$

- Standard deviation of X, σ_X, is

$$\sigma_X \equiv \sqrt{E[(X - E[X])]^2} = \sqrt{(np(1-p))}.$$

Normal Approximates Binomial

$error_S(h)$ follows a *Binomial* distribution, with

- mean $\mu_{errorS(h)} = error_D(h)$
- standard deviation $\sigma_{errorS(h)}$

$$\sigma_{errorS(h)} = \sqrt{(error_D(h)(1 - error_D(h))/n)}$$

Approximate this by a *Normal* distribution with

- mean $\mu_{errorS(h)} = error_D(h)$
- standard deviation $\sigma_{errorS(h)}$

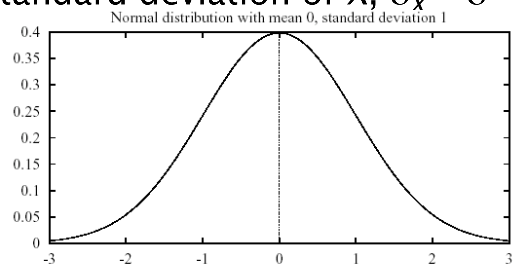
$$\sigma_{errorS(h)} \approx \sqrt{(error_S(h)(1 - error_S(h))/n)}$$

Normal Probability Dist.

$$p(x) = 1/\sqrt{(\pi\sigma^2)}\ \exp(-1/2\ ((x-\mu)/\sigma)^2)$$

The probability that X will fall into the interval (a, b) is given by $\int_a^b p(x)\, dx$.

- Expected, or mean value of X, $E[X] = \mu$.
- Variance of X is $Var(X) = \sigma^2$
- Standard deviation of X, $\sigma_X = \sigma$

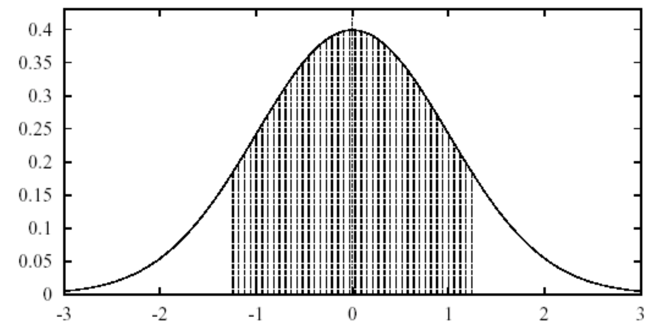


Normal Probability Dist.

80% of area (probability) lies in $\mu \pm 1.28\sigma$.

N% of area (probability) lies in $\mu \pm z_N \sigma$.

N%:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58



Confidence, More Correctly

If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$

Then,

- With approximately 95% probability, $error_S(h)$ lies in interval

$$error_D(h) \pm 1.96 \sqrt{(error_D(h)\ (1 - error_D(h))/n)}$$

equivalently, $error_D(h)$ lies in interval

$$error_S(h) \pm 1.96 \sqrt{(error_D(h)\ (1 - error_D(h))/n)},$$

which is approximately

$$error_S(h) \pm 1.96 \sqrt{(error_S(h)\ (1 - error_S(h))/n)}$$

Central Limit Theorem

Consider a set of independent, identically distributed random variables $Y_1, \ldots, Y_n$, all governed by an arbitrary probability distribution with mean μ and finite variance σ^2. Define the sample mean,

$$\bar{Y} \equiv 1/n\ \Sigma_{i=1}^{n} Y_i\,.$$

Central Limit Theorem. As $n \to \infty$, the distribution governing $\bar{Y}$ approaches a Normal distribution, with mean μ and variance σ^2/n.

Calculating Conf. Intervals

1. Pick parameter p to estimate
- $error_D(h)$.

2. Choose an estimator
- $error_S(h)$.

3. Determine probability distribution that governs estimator
- $error_S(h)$ governed by Binomial distribution, approximated by Normal when $n \geq 30$.

4. Find interval (L, U) such that N% of probability mass falls in the interval
- Use table of z_N values

Difference Between Hypotheses

1. Pick parameter to estimate

$$d \equiv error_D(h_1) - error_D(h_2)$$

2. Choose an estimator

$$\hat{d} \equiv error_{S1}(h_1) - error_{S2}(h_2)$$

3. Determine probability distribution that governs estimator

$$\sigma_{\hat{d}} \approx \sqrt{[(error_{S1}(h_1)\,(1 - error_{S1}(h_1))/n_1) + (error_{S2}(h_2)\,(1 - error_{S2}(h_2))/n_2)]}$$

4. Find interval (L, U) such that N% of probability mass falls in the interval

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

Paired t Test

Can be used to compare h_A, h_B as follows.

1. Partition data into k disjoint test sets T_1, T_2, … , T_k of equal size, where this size is at least 30.
2. For i from 1 to k, do

   $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$
3. Return the value $\bar{\delta}$, where

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

Confidence

N% confidence interval estimate for d:

$$\bar{\delta} \pm t_{N,k-1}\, s_{\bar{\delta}}$$

$$s_{\bar{\delta}} \equiv \sqrt{[\ 1/(k(k-1))\ \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2]}$$

Note δ_i approximately Normally distributed.

Use Student's t distribution.

Comparing Learning Algorithms

Want to compare learning algorithms L_A and L_B

What we'd like to estimate:

$$E_{S \subset D}[error_D(L_A(S)) - error_D(L_B(S))]$$

where $L(S)$ is the hypothesis output by learner L using training set S.

That is, the expected difference in true error between hypotheses output by learners L_A and L_B, when trained using randomly selected training sets S drawn according to distribution D.

An Estimator

But, given limited data D_0, what is a good estimator?

- could partition D_0 into training set S_0 and testing set T_0, and measure

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$$

- even better, repeat this many times and average the results (next slide)

Using Fixed Data to Compare

1. Partition data D_0 into k (10?) disjoint test sets $T_1, T_2, \ldots, T_k$ 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

   - $S_i \leftarrow \{D_0 - T_i\}$
   - $\delta_i \leftarrow error_{T_i}(L_A(S_i)) - error_{T_i}(L_B(S_i))$
3. Return the value $\bar{\delta}$, where

$$\bar{\delta} \equiv 1/k \ \sum_{i=1}^{k} \delta_i .$$

Statistical Correctness

Notice we'd like to use the paired t test on $\bar{\delta}$ to obtain a confidence interval

But it's not really correct, because the training sets in this algorithm are not independent (they overlap!)

More correct to view algorithm as producing an estimate of

$$E_{S \subset D_0}[error_D(L_A(S)) - error_D(L_B(S))]$$

instead of

$$E_{S \subset D}[error_D(L_A(S)) - error_D(L_B(S))],$$

but even this approximation is better than no comparison!