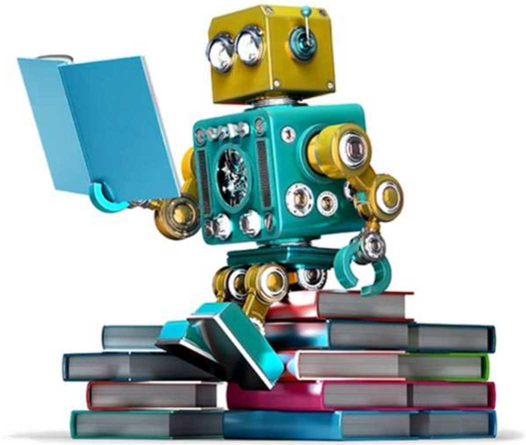


MACHINE LEARNING

MODULE-V CHAPTER 5
EVALUATION HYPOTHESIS,

BY
HARIVINOD N
VIVEKANANDA COLLEGE OF ENGINEERING TECHNOLOGY, PUTTUR

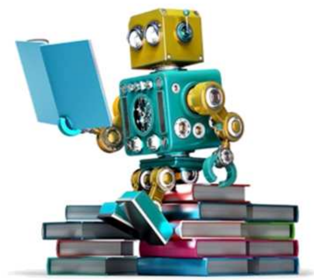
Overview

- This chapter presents an introduction to statistical methods for estimating hypothesis accuracy
- Focuses on three questions.
  1. Given the observed accuracy of a hypothesis over a limited sample of data, how well does this estimate its accuracy over additional examples?
  2. Given that one hypothesis outperforms another over some sample of data, how probable is it that this hypothesis is more accurate in general?
  3. When data is limited what is the best way to use this data to both learn a hypothesis and estimate its accuracy?

Module 5 - Outline

Chapter 5: Evaluating Hypothesis

1. **Motivation**
2. Estimating hypothesis accuracy
3. Basics of sampling theorem
4. General approach for deriving confidence intervals
5. Difference in error of two hypothesis
6. Comparing learning algorithms
7. Summary



Motivation..(1)

Importance of evaluate the performance,

1. To understand whether to use the hypothesis.
    - For instance, when learning from a limited-size database indicating the effectiveness of different medical treatments, it is important to understand as precisely as possible the accuracy of the learned hypotheses.

2. Evaluating hypotheses is an integral component of many learning methods.
    - For example, in post-pruning decision trees to avoid overfitting, we must evaluate resultant trees

Motivation..(2)

- Data is plentiful – Accuracy is straightforward.
- Difficulties arise given limited set of data. They are

1. Bias in the estimate.

- the observed accuracy of the learned hypothesis over the training examples is often a poor estimator of its accuracy over future examples.
- i.e it is a biased estimate of hypothesis accuracy over future examples.
- To obtain an unbiased estimate of future accuracy, we typically test the hypothesis on some set of test examples chosen independently of the training examples and the hypothesis.

Motivation…(3)

2. Variance in the estimate.

- Even if the hypothesis accuracy is measured over an unbiased set of test examples independent of the training examples,

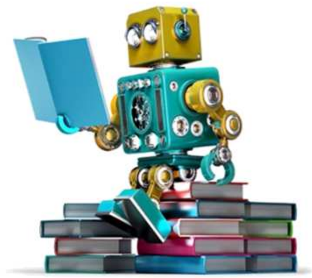
  the measured accuracy can still vary from the true accuracy, depending on the makeup of the particular set of test examples.

- The smaller the set of test examples, the greater the expected variance.

Module 5 - Outline

Chapter 5: Evaluating Hypothesis

1. Motivation
2. **Estimating hypothesis accuracy**
3. Basics of sampling theorem
4. General approach for deriving confidence intervals
5. Difference in error of two hypothesis
6. Comparing learning algorithms
7. Summary



Estimating Hypothesis Accuracy

- Set of possible instances – X
  - Ex: Set of people
- Various target functions may be defined over X
  - Ex: People who plan to buy cell phone is this year
- The target function f : X → {0,1} classifies each person according to whether or not they plan to purchase cell phone this year.
- The learning task is to learn the target concept or target function f by considering a space H of possible hypotheses.
- Assume there is some unknown probability distribution ***D*** that defines the probability of encountering each instance in X
  - *D* might assign a higher probability to encountering 19-year-old people than 91-year-old people.

Estimating Hypothesis Accuracy

Within this general setting we are interested in the following two questions:

1. Given a hypothesis h and a data sample containing n examples drawn at random according to the distribution D, what is the best estimate of the accuracy of h over future instances drawn from the same distribution?

2. What is the probable error in this accuracy estimate?

Sample error and True error

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

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

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: The **true error** (denoted $error_{\mathcal{D}}(h)$) of hypothesis h with respect to target function f and distribution $\mathcal{D}$, is the probability that h will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}} [f(x) \neq h(x)]$$

How to compute **$error_D(h)$?**

Confidence Intervals for Discrete-Valued Hypotheses

- Here we give an answer to the question
- *"How good an estimate of **$error_D(h)$** is provided by **$error_S(h)$**? "* for the case in which ***h*** is a discrete-valued hypothesis.

Confidence Intervals for Discrete-Valued Hypotheses

More specifically, suppose we wish to estimate the true error for some discrete-valued hypothesis h, based on its observed sample error over a sample S, where

- the sample S contains n examples drawn independent of one another, and independent of h, according to the probability distribution $\mathcal{D}$
- $n \geq 30$
- hypothesis h commits r errors over these n examples (i.e., $error_S(h) = r/n$).

Under these conditions, statistical theory allows us to make the following assertions:

1. Given no other information, the most probable value of $error_{\mathcal{D}}(h)$ is $error_S(h)$
2. With approximately 95% probability, the true error $error_{\mathcal{D}}(h)$ lies in the interval

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

Example

- To illustrate, suppose the data sample ***S*** contains ***n*** = ***40*** examples and that hypothesis ***h*** commits ***r*** = ***12*** errors over this data. In this case, the sample error ***errors(h)*** = ***12/40*** = ***.30.***
- Given no other information, the best estimate of the true error ***errorD(h)*** is the observed sample error ***.30.***
- Suppose, the ***95%*** is the confidence interval,
- according to the above expression,

$$0.30 \pm (1.96 \cdot .07) = 0.30 \pm .14.$$

General expression

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

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

Confidence level $N\%$:	50%	68%	80%	90%	95%	98%	99%
Constant z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

TABLE 5.1
Values of z_N for two-sided $N\%$ confidence intervals.

Note on Error computation

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

- The expression provides only an approximate confidence interval,
  - the approximation is quite good when the sample contains at least ***30*** examples, and
  - ***errors(h)*** is not too close to ***0*** or ***1.***
- ***A*** more accurate rule of thumb is that the above approximation works well when

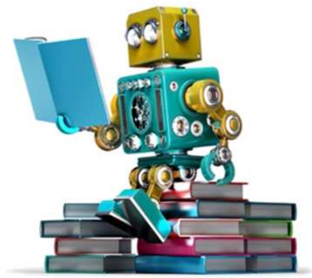
$$n\ error_S(h)(1 - error_S(h)) \geq 5$$

Module 5 - Outline

MACHINE LEARNING

Chapter 5: Evaluating Hypothesis

1. Motivation
2. Estimating hypothesis accuracy
3. **Basics of sampling theorem**
4. General approach for deriving confidence intervals
5. Difference in error of two hypothesis
6. Comparing learning algorithms
7. Summary



Basic definitions and facts from statistics

- A *random variable* can be viewed as the name of an experiment with a probabilistic outcome. Its value is the outcome of the experiment.
- A *probability distribution* for a random variable Y specifies the probability $\Pr(Y = y_i)$ that Y will take on the value y_i, for each possible value y_i.
- The *expected value*, or *mean*, of a random variable Y is $E[Y] = \sum_i y_i \Pr(Y = y_i)$. The symbol μ_Y is commonly used to represent E[Y].
- The *variance* of a random variable is $Var(Y) = E[(Y - \mu_Y)^2]$. The variance characterizes the width or dispersion of the distribution about its mean.
- The *standard deviation* of Y is $\sqrt{Var(Y)}$. The symbol σ_Y is often used used to represent the standard deviation of Y.
- The *Binomial distribution* gives the probability of observing r heads in a series of n independent coin tosses, if the probability of heads in a single toss is p.

Basic definitions and facts from statistics

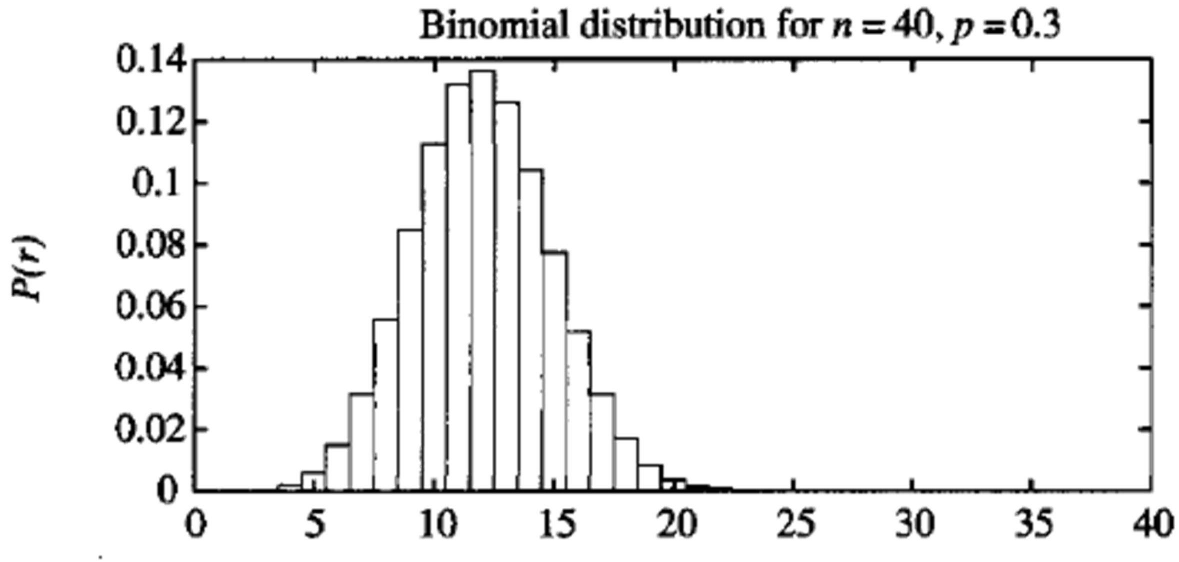
- The *Normal distribution* is a bell-shaped probability distribution that covers many natural phenomena.
- The *Central Limit Theorem* is a theorem stating that the sum of a large number of independent, identically distributed random variables approximately follows a Normal distribution.
- An *estimator* is a random variable Y used to estimate some parameter p of an underlying population.
- The *estimation bias* of Y as an estimator for p is the quantity $(E[Y] - p)$. An unbiased estimator is one for which the bias is zero.
- A *N% confidence interval* estimate for parameter p is an interval that includes p with probability $N\%$.

Error Estimation and Estimating Binomial Proportions

- Precisely how does the deviation between sample error and true error depend on the size of the data sample?
- The key to answering this question is to note that when we measure the sample error we are performing an experiment with a random outcome.

- Imagine k random experiments, with $errors_{s1}(h)$, $errors_{s2}(h)$ . . . $errors_{sk}(h)$.
- Plot a histogram displaying the frequency with which we observed each possible error value.
- As we allowed k to grow, the histogram would approach the form of the distribution called the Binomial distribution.

Binomial Distribution



- ***A Binomial distribution*** gives the probability of observing ***r*** heads in a sample of ***n*** independent coin tosses, when the probability of heads on a single coin toss is ***p.***

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

Binomial Distribution

If the random variable X follows a Binomial distribution, then:

- The probability $\Pr(X = r)$ that X will take on the value r is given by $P(r)$
- The expected, or mean value of X, $E[X]$, is

$$E[X] = np$$

- The variance of X, $Var(X)$, is

$$Var(X) = np(1 - p)$$

- The standard deviation of X, σ_X, is

$$\sigma_X = \sqrt{np(1 - p)}$$

Binomial Distribution

Definition: Consider a random variable Y that takes on the possible values $y_1, \ldots y_n$. The **expected value** of Y, $E[Y]$, is

$$E[Y] \equiv \sum_{i=1}^{n} y_i \Pr(Y = y_i) \tag{5.3}$$

For example, if Y takes on the value 1 with probability .7 and the value 2 with probability .3, then its expected value is $(1 \cdot 0.7 + 2 \cdot 0.3 = 1.3)$. In case the random variable Y is governed by a Binomial distribution, then it can be shown that

$$E[Y] = np \tag{5.4}$$

where n and p are the parameters of the Binomial distribution defined in Equation (5.2).

Binomial Distribution

Definition: The **variance** of a random variable Y, $Var[Y]$, is

$$Var[Y] \equiv E[(Y - E[Y])^2]$$

Definition: The **standard deviation** of a random variable Y, σ_Y, is

$$\sigma_Y \equiv \sqrt{E[(Y - E[Y])^2]}$$

In case the random variable Y is governed by a Binomial distribution, then the variance and standard deviation are given by

$$Var[Y] = np(1 - p)$$
$$\sigma_Y = \sqrt{np(1 - p)} \qquad (5.7)$$

Estimator, Bias, Confidence Interval

Statisticians call $error_S(h)$ an *estimator* for the true error $error_{\mathcal{D}}(h)$.

Definition: The **estimation bias** of an estimator Y for an arbitrary parameter p is

$$E[Y] - p$$

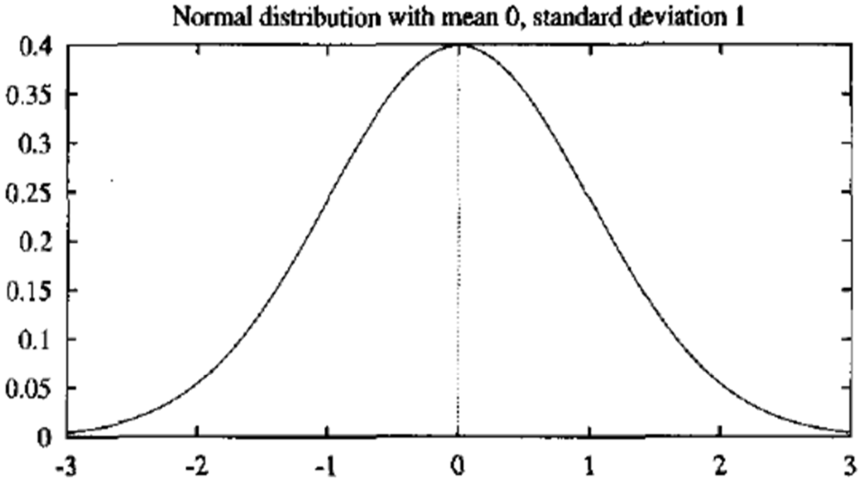
Definition: An $N\%$ **confidence interval** for some parameter p is an interval that is expected with probability $N\%$ to contain p.

For example, if we observe $r = 12$ errors in a sample of $n = 40$ independently drawn examples, we can say with approximately 95% probability that the interval 0.30 ± 0.14 contains the true error $error_{\mathcal{D}}(h)$.

How to find interval?

- For a given value of ***N*** how can we find the size of the interval that contains ***N%*** of the probability mass?
- Unfortunately, for the Binomial distribution this calculation can be quite tedious.
- Fortunately, however, an easily calculated and very good approximation can be found in most cases, based on the fact that
  - for sufficiently large sample sizes the Binomial distribution can be closely approximately the Normal distribution.

Normal Distribution



A Normal distribution (also called a Gaussian distribution) is a bell-shaped distribution defined by the probability density function

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

A Normal distribution is fully determined by two parameters in the above formula: μ and σ.

Normal Distribution

If the random variable X follows a normal distribution, then:

- The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

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

$$E[X] = \mu$$

- The variance of X, $Var(X)$, is

$$Var(X) = \sigma^2$$

- The standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

Confidence level $N\%$:	50%	68%	80%	90%	95%	98%	99%
Constant z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

TABLE 5.1
Values of z_N for two-sided $N\%$ confidence intervals.

To summarize, if a random variable Y obeys a Normal distribution with mean μ and standard deviation σ, then the measured random value y of Y will fall into the following interval $N\%$ of the time

$$\mu \pm z_N \sigma \tag{5.10}$$

Confidence interval

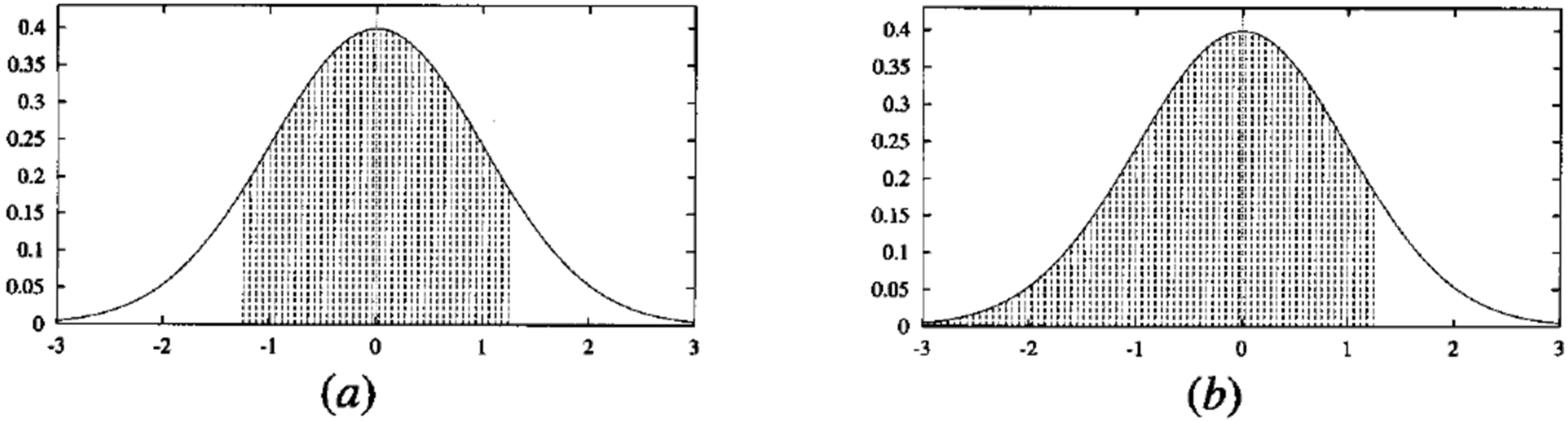


FIGURE 5.1
A Normal distribution with mean 0, standard deviation 1. (*a*) With 80% confidence, the value of the random variable will lie in the two-sided interval [−1.28, 1.28]. Note $z_{.80} = 1.28$. With 10% confidence it will lie to the right of this interval, and with 10% confidence it will lie to the left. (*b*) With 90% confidence, it will lie in the one-sided interval [−∞, 1.28].

$N\%$ confidence intervals for discrete-valued hypotheses

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

Recall that two approximations were involved in deriving this expression, namely:

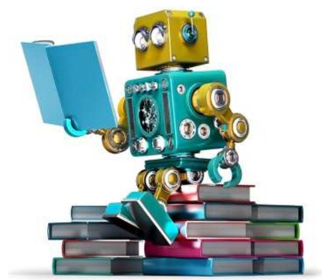
1. in estimating the standard deviation σ of $error_S(h)$, we have approximated $error_{\mathcal{D}}(h)$ by $error_S(h)$ [i.e., in going from Equation (5.8) to (5.9)], and
2. the Binomial distribution has been approximated by the Normal distribution.

The common rule of thumb in statistics is that these two approximations are very good as long as $n \geq 30$, or when $np(1-p) \geq 5$. For smaller values of n it is wise to use a table giving exact values for the Binomial distribution.

Module 5 - Outline

Chapter 5: Evaluating Hypothesis

1. Motivation
2. Estimating hypothesis accuracy
3. Basics of sampling theorem
4. **General approach for deriving confidence intervals**
5. Difference in error of two hypothesis
6. Comparing learning algorithms
7. Summary



Deriving Confidence interval

1. Pick parameter p to estimate
   - $error_{\mathcal{D}}(h)$
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

Central limit theorem

Theorem 5.1. Central Limit Theorem. Consider a set of independent, identically distributed random variables $Y_1 \ldots Y_n$ governed by an arbitrary probability distribution with mean μ and finite variance σ^2. Define the sample mean, $\bar{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Then as $n \rightarrow \infty$, the distribution governing

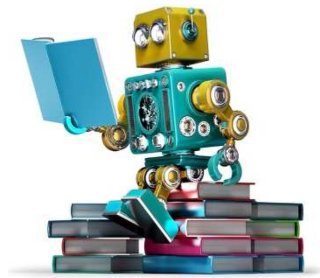
$$\frac{\bar{Y}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

approaches a Normal distribution, with zero mean and standard deviation equal to 1.

Module 5 - Outline

Chapter 5: Evaluating Hypothesis

1. Motivation
2. Estimating hypothesis accuracy
3. Basics of sampling theorem
4. General approach for deriving confidence intervals
5. **Difference in error of two hypothesis**
6. Comparing learning algorithms
7. Summary



Difference in error of two hypothesis

- Difference in true error

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

- Difference between the sample errors

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

- Approximate Variance

$$\sigma_{\hat{d}}^2 \approx \frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}$$

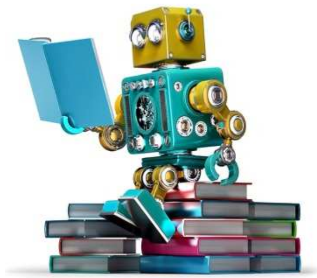
- Approximate ***N%*** confidence interval estimate for d is

$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

Module 5 - Outline

Chapter 5: Evaluating Hypothesis

1. Motivation
2. Estimating hypothesis accuracy
3. Basics of sampling theorem
4. General approach for deriving confidence intervals
5. Difference in error of two hypothesis
6. **Comparing learning algorithms**
7. Summary



Comparing Learning Algorithms

- Often we are interested in comparing the performance of two learning algorithms L_A and L_B, rather than two specific hypotheses.
- To find relative performance of these two algorithms averaged over all the training sets of size n that might be drawn from the underlying instance distribution ***V.***
- i.e. we wish to estimate the expected value of the difference in their errors

$$\underset{S \subset \mathcal{D}}{E}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

where L(S) denotes the hypothesis output

Practical method

- Practically we have limited sample D_0
- So divide D_0 into a training set S_0 and a disjoint test set T_0.
- The training data can be used to train both L_A and L_B,
- Test data can be used to compare the accuracy of the two learned hypotheses.
- In other words, we measure the quantity

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

- To improve above estimator - repeatedly partition the data D_0 into disjoint training and test sets and to take the mean of the test set errors for these different experiments.

A Procedure to estimate the difference in error between two learning methods *LA* and *LB*

1. Partition the available data Do into k disjoint subsets T1, T2, . . . , Tk of equal size, where this size is at least 30.
2. For i from 1 to k, do
   use Ti for the test set, & the remaining data for training set Si

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

3. Return the value $\bar{\delta}$, where

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

The approximate $N\%$ confidence interval for estimating the quantity in Equation (5.16) using $\bar{\delta}$ is given by

$$\bar{\delta} \pm t_{N,k-1}\ s_{\bar{\delta}} \tag{5.17}$$

where $t_{N,k-1}$ is a constant that plays a role analogous to that of z_N in our earlier confidence interval expressions, and where $s_{\bar{\delta}}$ is an estimate of the standard deviation of the distribution governing $\bar{\delta}$. In particular, $s_{\bar{\delta}}$ is defined as

$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)}\sum_{i=1}^{k}(\delta_i - \bar{\delta})^2} \tag{5.18}$$

k-1 is the degree of freedom usally denoted by v

- In $t_{N,k-1}$ N represent confidence interval, k-1

Notice the constant $t_{N,k-1}$ in Equation (5.17) has two subscripts. The first specifies the desired confidence level, as it did for our earlier constant z_N. The second parameter, called the number of *degrees of freedom* and usually denoted by ν, is related to the number of independent random events that go into producing the value for the random variable $\bar{\delta}$. In the current setting, the number of degrees of freedom is $k - 1$. Selected values for the parameter t are given in Table 5.6. Notice that as $k \to \infty$, the value of $t_{N,k-1}$ approaches the constant z_N.

- Note the procedure described here for comparing two learning methods involves testing the two learned hypotheses on identical test sets.
- Tests where the hypotheses are evaluated over identical samples are called paired tests.

	Confidence level N			
	90%	95%	98%	99%
$\nu = 2$	2.92	4.30	6.96	9.92
$\nu = 5$	2.02	2.57	3.36	4.03
$\nu = 10$	1.81	2.23	2.76	3.17
$\nu = 20$	1.72	2.09	2.53	2.84
$\nu = 30$	1.70	2.04	2.46	2.75
$\nu = 120$	1.66	1.98	2.36	2.62
$\nu = \infty$	1.64	1.96	2.33	2.58

TABLE 5.6

Values of $t_{N,\nu}$ for two-sided confidence intervals. As $\nu \to \infty$, $t_{N,\nu}$ approaches z_N.

Paired t Tests

- We discussed procedure for comparing two learning methods given a fixed set of data set.
- Now let us understand statistical justification for the procedure and confidence interval estimate,

$$\bar{\delta} \pm t_{N,k-1}\ s_{\bar{\delta}} \qquad s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

Consider the following estimation problem:

- We are given the observed values of a set of independent, identically distributed random variables $Y_1, Y_2, \ldots, Y_k$.
- We wish to estimate the mean μ of the probability distribution governing these Y_i.
- The estimator we will use is the sample mean $\bar{Y}$

$$\bar{Y} \equiv \frac{1}{k} \sum_{i=1}^{k} Y_i$$

Paired t tests

- How to estimate ***μ?***
  - Assume that instead of having a fixed sample of data Do, we take quest new training examples from underlying distribution.
  - Compute $\bar{\delta}$. This itself is the estimate of ***μ.***

- How good an estimate of ***μ*** is provided by $\bar{\delta}$?
  - To understand this we need standard deviation
- t tests are applied in this situation

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

where $s_{\bar{Y}}$ is the estimated standard deviation of the sample mean

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

Paired t test

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

where $s_{\bar{Y}}$ is the estimated standard deviation of the sample mean

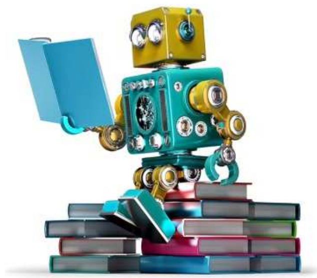
$$s_{\bar{Y}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (Y_i - \bar{Y})^2}$$

and where $t_{N,k-1}$ is a constant analogous to our earlier z_N. In fact, the constant $t_{N,k-1}$ characterizes the area under a probability distribution known as the t distribution, just as the constant z_N characterizes the area under a Normal distribution. The t distribution is a bell-shaped distribution similar to the Normal distribution, but wider and shorter to reflect the greater variance introduced by using $s_{\bar{Y}}$ to approximate the true standard deviation $\sigma_{\bar{Y}}$. The t distribution approaches the Normal distribution (and therefore $t_{N,k-1}$ approaches z_N) as k approaches infinity. This is intuitively satisfying because we expect $s_{\bar{Y}}$ to converge toward the true standard deviation $\sigma_{\bar{Y}}$ as the sample size k grows, and because we can use z_N when the standard deviation is known exactly.

Module 5 - Outline

Chapter 5: Evaluating Hypothesis

1. Motivation
2. Estimating hypothesis accuracy
3. Basics of sampling theorem
4. General approach for deriving confidence intervals
5. Difference in error of two hypothesis
6. Comparing learning algorithms
7. **Summary**



Summary

- Statistical theory provides a basis for estimating the true error ($error_{\mathcal{D}}(h)$) of a hypothesis h, based on its observed error ($error_S(h)$) over a sample S of data. For example, if h is a discrete-valued hypothesis and the data sample S contains $n \geq 30$ examples drawn independently of h and of one another, then the $N\%$ confidence interval for $error_{\mathcal{D}}(h)$ is approximately

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

  where values for z_N are given in Table 5.1.

- In general, the problem of estimating confidence intervals is approached by identifying the parameter to be estimated (e.g., $error_{\mathcal{D}}(h)$) and an estimator (e.g., $error_S(h)$) for this quantity. Because the estimator is a random variable (e.g., $error_S(h)$ depends on the random sample S), it can be characterized by the probability distribution that governs its value. Confidence intervals can then be calculated by determining the interval that contains the desired probability mass under this distribution.

Summary

- One possible cause of errors in estimating hypothesis accuracy is *estimation bias*. If Y is an estimator for some parameter p, the estimation bias of Y is the difference between p and the expected value of Y. For example, if S is the training data used to formulate hypothesis h, then $error_S(h)$ gives an optimistically biased estimate of the true error $error_{\mathcal{D}}(h)$.
- A second cause of estimation error is *variance* in the estimate. Even with an unbiased estimator, the observed value of the estimator is likely to vary from one experiment to another. The variance σ^2 of the distribution governing the estimator characterizes how widely this estimate is likely to vary from the correct value. This variance decreases as the size of the data sample is increased.

Summary

- Comparing the relative effectiveness of two learning algorithms is an estimation problem that is relatively easy when data and time are unlimited, but more difficult when these resources are limited. One possible approach described in this chapter is to run the learning algorithms on different subsets of the available data, testing the learned hypotheses on the remaining data, then averaging the results of these experiments.
- In most cases considered here, deriving confidence intervals involves making a number of assumptions and approximations. For example, the above confidence interval for $error_{\mathcal{D}}(h)$ involved approximating a Binomial distribution by a Normal distribution, approximating the variance of this distribution, and assuming instances are generated by a fixed, unchanging probability distribution. While intervals based on such approximations are only approximate confidence intervals, they nevertheless provide useful guidance for designing and interpreting experimental results in machine learning.

Thank You