Evaluating Hypotheses (Model)

M.M. Pedram
pedram@tmu.ac.ir
Tarbiat Moallem University of Tehran
(Fall 2009)

1

Terminology

- Statistics
  - Basic terms
  - Sample error, true error
  - Confidence intervals for observed hypothesis error
  - Estimators
  - Distributions, Central Limit Theorem
  - Cost/utility
  - Tests for significance (Paired t-tests)
- Comparing Learning Methods

2

Objectives

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?

- Because limited samples of data might misrepresent the general distribution of data, estimating *true accuracy* from such samples can be misleading. Statistical methods, together with assumptions about the underlying distributions of data, allow one to bound the difference between *observed accuracy* over the sample of available data and the true accuracy over the entire distribution of data.

Data Sets

- Data set: set of examples of a problem
- Feature (attribute, field, variable): one value that defines an instance
  - Categorical (nominal) with a set of possible values versus continuous (qualitative) – numeric range of possible values
  - Input feature (independent variable) versus output feature (dependent variable)
  - Can be missing (value not known)
- Example (instance, case, record, feature vector, tuple): the values of the input (and in some cases output) features of variables
- Skewed data set, or Unbalanced data set: one class occurs far more than others
- Multi-class problem: more than 2 output values
- Regression problem: output value is continuous

Data Sets (continued)

- Training data set: the set of data used to learn (create) a model of a problem.
- Test data set: the set of data used to estimate some value (often accuracy) related to a model.
- Validation set: a set of data used to select parameters for a model, often as follows
  - Divide training data into a "sub" training set and validation set,
  - For each possible set of parameters:
    - Create a model using the "sub" training set,
    - Evaluate the model on the validation set and pick the one that performs the best.

5

Evaluating Models

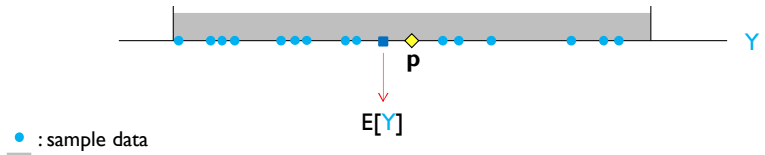
- Need a measure of value: the cost (loss, utility) of a model
- Often use accuracy (or error)
  - *Accuracy*: how many examples we get "right"
  - *Error*: how many examples we get "wrong"
- Can be weighted
  - If examples are not equal, could count the cost (or utility) of misclassified or correct examples
- Building a *confusion matrix* is helpful for result analysis:
  - Perfect prediction has all values down the diagonal
  - Off diagonal entries can often tell us about what is being mis-predicted

6

Some definitions

- *estimator*: a random variable Y used to estimate some parameter **p** of an underlying population.
- The *estimation bias* of Y as an estimator for parameter **p**:

$$(estimation)\ bias = E[Y] - \mathbf{p}$$



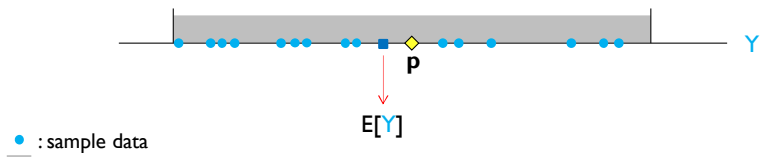
: sample data
: population distribution
: mean of population

- An *unbiased estimator* is one for which the bias is zero.

7

Some definitions

- A N% confidence interval estimate for parameter **p**:
  an interval that includes **p** with probability N%.



: sample data
: population distribution
: mean of population

8

Estimating Hypothesis Accuracy

- X : space of possible instances over which various target functions may be defined.

  e.g. :

  X= the set of all people.

  target concept or target function= people who plan to purchase new skis this year

- Different instances in X may be encountered with different frequencies. A convenient way to model this is to assume there is some unknown probability distribution $\mathcal{D}$ that defines the probability of encountering each instance in X.

  e.g. : $\mathcal{D}$ might assign a higher probability to encountering 19-year-old people than 109-year-old people.

Estimating Hypothesis Accuracy

- Notice: $\mathcal{D}$ says nothing about whether x is a positive or negative example; it only determines the probability that x will be encountered.
- The learning task is to learn the target concept or target function f by considering a space H of possible hypotheses (models).

  e.x.

  - the target function f : "people who plan to purchase new skis this year"

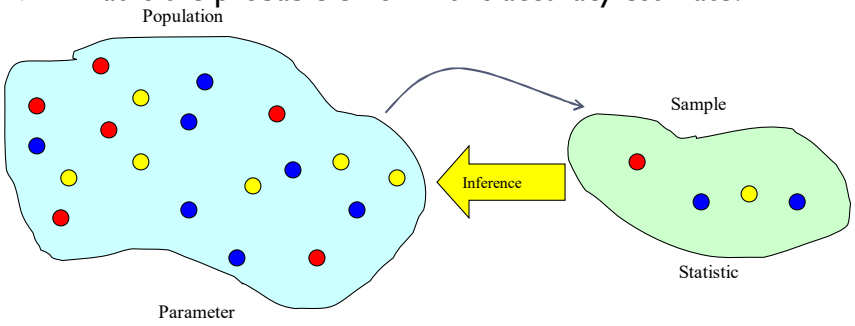
    $f : X \rightarrow \{0,1\}$ classifies each person according to whether or not they plan to purchase skis this year.

  - sample of training data collected by surveying people as they arrive at a ski resort.
  - X : the space of all people, who might be described by attributes such as their *age*, *occupation*, *how many times they skied last year*, etc.
  - distribution $\mathcal{D}$ specifies for each person x the probability that x will be encountered as the next person arriving at the ski resort.

Estimating Hypothesis Accuracy

We are interested in the following questions:

1. Given a hypothesis h and a data sample containing n examples drawn at random from $\mathcal{D}$, what is the best estimate of the accuracy of h over future instances?
2. What is the probable error in this accuracy estimate?



11

Two Definitions of Error

- The *true error* 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 $\mathcal{D}$.

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

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

$$error_S(h) \equiv \frac{1}{|S|}\sum_{x \in S}\delta\left(f(x) \neq h(x)\right)$$

where

$$\delta\left(f(x) \neq h(x)\right) = \begin{cases} 1 & f(x) \neq h(x) \\ 0 & otherwise \end{cases}$$

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

12

Example

Hypothesis h misclassifies 12 of 40 examples in S:

$$error_S(h) = \frac{12}{40} = 0.30$$

What is $error_D(h)$?

Problems Estimating Error

Two key difficulties when only limited data is available:

1. *Bias in the estimate*: the learned hypothesis was derived from these examples, they will typically provide an optimistically biased estimate of hypothesis accuracy over future examples.

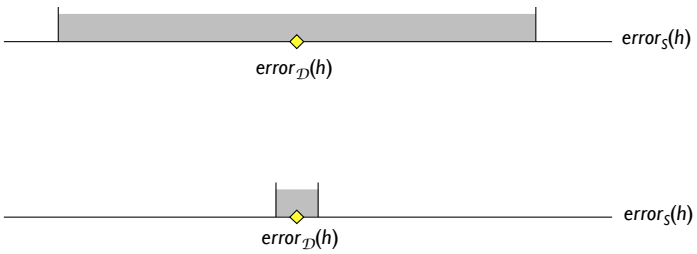
   i.e. if S is training set, $error_S(h)$ is biased

   $$bias \equiv E[error_S(h)] - error_{\mathcal{D}}(h)$$

   - For unbiased estimate, h and S must be chosen independently, i.e. test the hypothesis on some set of test examples chosen independently of the training examples and the hypothesis.

Problems Estimating Error

2. *Variance in the estimate*: Even with unbiased S, $error_S(h)$ may still vary from $error_{\mathcal{D}}(h)$.
   - The smaller the set of test examples, the greater the expected variance.



15

Estimators

Experiment:

1. Choose sample S of size *n* according to distribution $\mathcal{D}$, i.e. independent of *h*.
2. Measure $error_S(h)$

$error_S(h)$ is a random variable (i.e., result of an experiment)

$error_S(h)$ is an unbiased estimator for $error_{\mathcal{D}}(h)$

Given observed $error_S(h)$ what can we conclude about $error_{\mathcal{D}}(h)$?

16

Basics of Sampling Theory

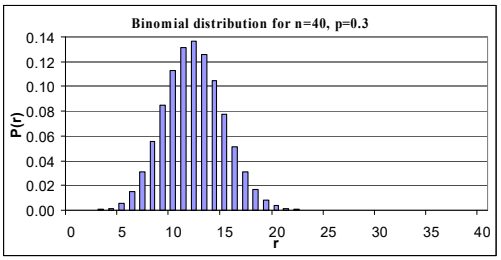
- Error Estimation

  How does the deviation between sample error and true error depend on the size of the data sample?

- Estimating $error_{\mathcal{D}}(h)$ from testing h on a random sample of n instances is equivalent to estimating the probability p that a bent coin will turn up heads from a random sample of n tosses

  $\Rightarrow$ Binomial Distribution

Binomial Probability Distribution



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

Probability P(r) of r heads(errors)in n coin flips, if p = probability(*heads*)

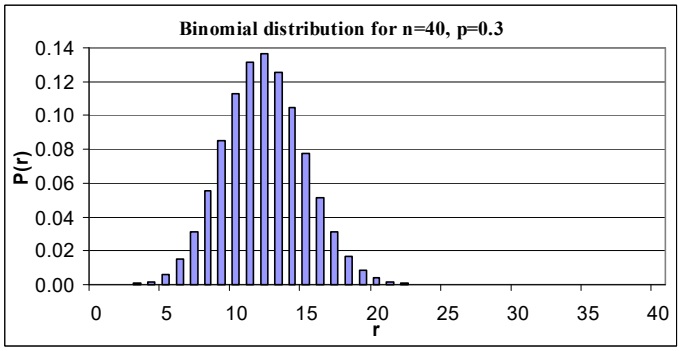
$$E[X]\equiv\sum_{i=0}^{n} i\cdot P(i)=np$$

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

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

$error_S(h)$ is a Random Variable

- Rerun experiment with different randomly drawn S (size n)
- Probability of observing r misclassified examples:



$$P(R=r)=\frac{n!}{r!(n-r)!}error_D(h)^r(1-error_D(h))^{n-r}$$

Binomial Probability Distribution

$error_{\mathcal{D}}(h) = p$

Estimator: $Y = error_S(h) = r/n$

Estimation Bias: $E(Y) - p \ (= 0)$?

Note:

- The variance in this estimate arises completely from the variance in r, because n is a constant. Because r is Binomially distributed, its mean and variance are given by np and $np(1-p)$.

$$E(Y) = 1/n\ E(r)$$
$$= 1/n\ .np = p$$

$\Rightarrow$ Y is an unbiased estimator (with binomial distribution)

Binomial Probability Distribution

Standard deviation: $\sigma_{error_S(h)} = \sigma_r/n = [p(1-p)/n]^{1/2}$

As $error_{\mathcal{D}}(h) = p$, we can substitute $error_{\mathcal{D}}(h)$ for p, i.e.

$$\sigma_{error_S(h)} = \sqrt{\frac{error_{\mathcal{D}}(h)\ (1 - error_{\mathcal{D}}(h))}{n}}$$

- As $error_{\mathcal{D}}(h)$ is unknown, but we will substitute our estimate $error_S(h)=r/n$ for it.

Normal Approximation to Binomial

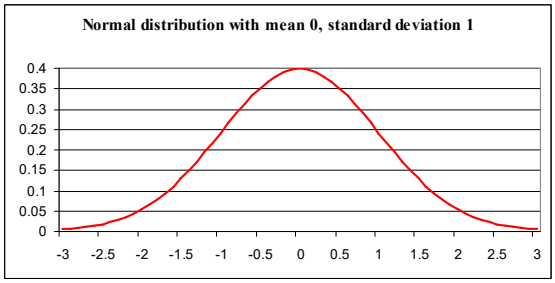
- If n is large enough, then the skew of the distribution is not too great. In this case, if a suitable continuity correction is used, then an excellent approximation to $B(n, p)$ is given by the *normal distribution*:

$$B(n, p) \rightarrow N(\mu,\sigma^2) = N(np, np(1-p))$$

where

$$n>30,\ np>5,\ \text{and}\ n(1-p)>5$$

Normal Probability Distribution



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

- The probability that X will fall into he interval (a, b) is: $\int_a^b p(x)dx$
- For the random variable X:

$$E[X] \equiv \mu$$
$$Var(X) \equiv \sigma^2$$
$$\sigma_X \equiv \sigma$$

Normal Approximation to Binomial

- $Y = error_S(h)$, as our estimator, follows a Binomial distribution, with

$$\mu_{error_S(h)} = error_{\mathcal{D}}(h)$$
$$\sigma_{error_S(h)} = \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

Which we approximate this by a normal distribution:

Note that $error_{\mathcal{D}}(h)$ is unknown, we can substitute our estimate $error_S(h)$ (or r/n) for it in $\sigma_{error_S(h)}$, i.e.

$$\mu_{error_S(h)} = error_{\mathcal{D}}(h)$$
$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

Confidence Intervals

- Thus it can be said about the random variable $Y = error_S(h)$:

$$\left.\begin{aligned}\mu_{error_S(h)} &= error_{\mathcal{D}}(h)\\ \sigma_{error_S(h)} &\approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}\end{aligned}\right\} \xrightarrow[error_S(h)]{\text{normal distribution for}}$$

with approximately N% probability, $error_S(h)$ lies in interval:

$$\mu_{error_S(h)} - z_n\sigma_{error_S(h)} \le \left(Y = error_S(h)\right) \le \mu_{error_S(h)} + z_n\sigma_{error_S(h)}$$

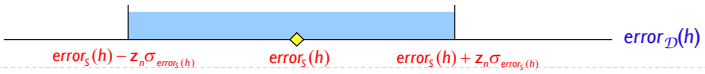
$$error_{\mathcal{D}}(h) - z_n\sigma_{error_S(h)} \le \left(Y = error_S(h)\right) \le error_{\mathcal{D}}(h) + z_n\sigma_{error_S(h)}$$

$$-z_n\sigma_{error_S(h)} \le error_S(h) - error_{\mathcal{D}}(h) \le +z_n\sigma_{error_S(h)}$$

$$-error_S(h) - z_n\sigma_{error_S(h)} \le -error_{\mathcal{D}}(h) \le -error_S(h) + z_n\sigma_{error_S(h)}$$

or, with approximately N% probability $error_D(h)$ lies in interval:

$$error_S(h) - z_n\sigma_{error_S(h)} \le error_D(h) \le error_S(h) + z_n\sigma_{error_S(h)}$$



25

Confidence Intervals

If S contains n examples, drawn independently of h and each other

$$n \ge 30$$

and $error_{\mathcal{D}}(h)$ is not too close to 0 or 1,

Then with approximately N% probability, $error_D(h)$ lies in interval

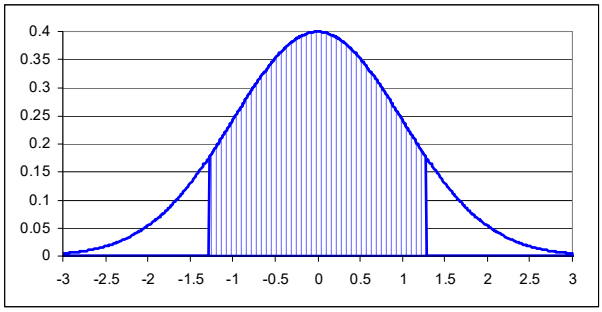
$$error_S(h) \pm z_N\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

Recall that $error_S(h)$ = r/n where r is errors over n examples.

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

26

Normal Probability Distribution



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

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

$$\mu_{error_S(h)} = error_D(h)$$

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

N% :	50%	68%	80%	90%	95%	98%	99%
z_N :	0.67	1.00	1.28	1.64	1.96	2.33	2.53

Confidence Intervals

- A more accurate rule of thumb is that the stated approximation works well when

Rule of Thumb: $n \cdot error_S(h) \cdot [1-error_S(h)] \geq 5$

or: $n \cdot error_S(h) \cdot accuracy(h) \geq 5$

Confidence Intervals, More Correctly

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

$$error_D(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1-error_{\mathcal{D}}(h))}{n}}$$

- equivalently, $error_D(h)$ lies in interval

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

- which is approximately

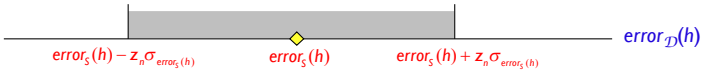
$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

Calculating Confidence Intervals

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

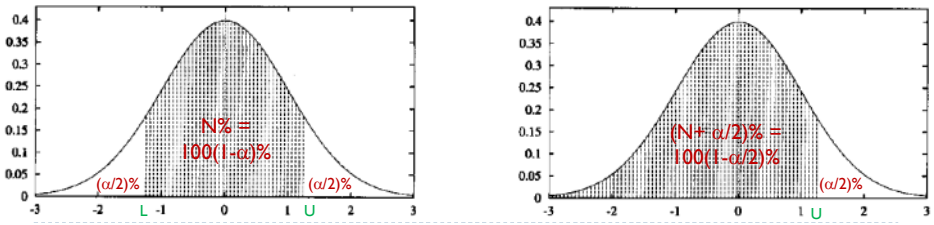
Two-Sided and One-Sided Bounds

Example: What is the probability that $error_D(h)$ is at most E?



answer:

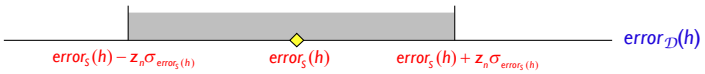
A N% = 100(1-α)% confidence interval with lower bound **L** and upper bound **U**, implies a (N+ α/2)% = 100(1- α/2)% confidence interval with lower bound L and no upper bound, or with upper bound U and no lower bound.

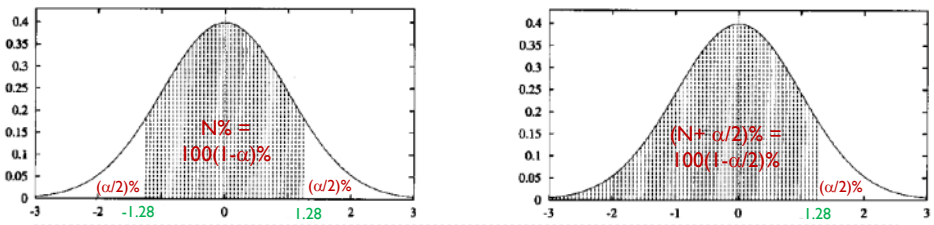


31

Two-Sided and One-Sided Bounds

- With 80% confidence: $error_D(h) \in [L,U]=[-1.28, 1.28]$
- or, with 10% confidence: $error_D(h) \in [U, \infty)=[1.28, \infty)$
- or, with 10% confidence: $error_D(h) \in (-\infty,U]=(-\infty, -1.28)$





32

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

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

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

Central Limit Theorem

- The Central Limit Theorem is a very useful fact:
  - Considering Y_i as error of model for sample i, then the mean error will be:
    $$\bar{Y} = error_S(h)$$
  - the distribution governing $error_S(h)$ can be approximated by a Normal distribution for sufficiently large ***n*** **(*n* > *30*)**. If we also know the variance for this (approximately) Normal distribution, then we can use compute confidence intervals.
  - Recall that we used such a Normal distribution to approximate the Binomial distribution that more precisely describes $error_S(h)$.

Difference Between Hypotheses

Test h_1 on sample S_1 containing n_1 randomly drawn examples, test h_2 on S_2 containing n_2 examples drawn from the same distribution :

1. Pick parameter to estimate

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

2. Choose an estimator

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

3. Determine probability distribution that governs estimator

$$\sigma_d \approx \sqrt{\frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2}}$$

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

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

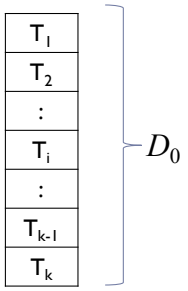
Paired Test

- Tests where the hypotheses are evaluated over identical samples are called ***paired tests***.
- Paired tests typically produce tighter (more concise) confidence intervals because any differences in observed errors in a paired test are due to differences between the hypotheses.
- In contrast, when the hypotheses are tested on separate data samples, differences in the two sample errors might be partially attributable to differences in the makeup of the two samples.

Paired t test to Compare h_A, h_B

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

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



N% confidence interval estimate for δ:

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

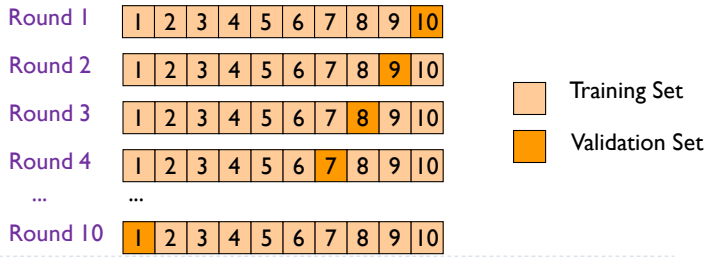
$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)}\sum_{i=1}^{k}(\delta_i - \bar{\delta})^2}$$

Note δ_i approximately Normally distributed

37

N-Fold Cross Validation

- N-Fold Cross Validation is a popular testing methodology.
- The method:
  - Divide data into N even-sized random folds
  - For n = 1 to N
    - Train set = all folds except n
    - Validation set = fold n
    - Create learner with train set
    - Count number of errors on validation set
  - Accumulate number of errors across N validation sets and divide by N (result is error rate)



38

N-Fold Cross Validation

- There are two possible goals in cross-validation:
  - To estimate performance of the learned model from available data using one algorithm. In other words, to gauge the generalizability of an algorithm.
  - To compare the performance of two or more different algorithms and find out the best algorithm for the available data, or alternatively to compare the performance of two or more variants of a parameterized model.