Model Assessment and Selection

Prof. Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2021)
http://cs182.sist.shanghaitech.edu.cn

Outline

Outline

Introduction

- We have discussed several learning algorithms and given a certain application, more than one is applicable.
- Two problems:
  - Performance evaluation: assessing the expected error rate of a learning algorithm on a application.
  - Performance comparison: comparing the expected error rates of two or more learning algorithms to conclude which one is better for a given application.
- We cannot look at the training set errors and decide based on those. Training data alone is not sufficient for performance evaluation and comparison.
- A validation set different from the training set is needed, typically requiring multiple runs to average over different sources of randomness.
- One learner tested over a validation set will generate one validation set error.
- Performance evaluation and comparison is done based on the distribution of the validation set errors.
- Any conclusion drawn is conditioned on the given dataset only. No algorithm is the best for all possible datasets.

Test Set

- When after training and validation, we decide on a particular algorithm and want to report its expected error, we should use a separate test set for this purpose.
- This data should have never been used before for training or validation and should be large for the error estimate to be meaningful.
- Given a dataset, we should first leave some part of it aside as the test set and use the rest for training and validation.

Performance Criteria

- We use error rates as the performance criterion here for performance evaluation and comparison, but there exist some other performance criteria depending on the application.
- Examples of other criteria:
  - Risks based on other loss functions more general than 0-1 loss (recall the lecture on Bayesian Decision Theory)
  - Training time and space complexity
  - Testing time and space complexity
  - Interpretability
  - Easy programmability
- The relative importance of these cost criteria changes depending on the specific application.
- Cost-sensitive learning takes other cost criteria into account.

- Methodology well developed in statistics can be used to conduct performance evaluation and comparison.

Outline

Cross-Validation

- For replication purposes, we need to get a number of training and validation set pairs from a dataset $\mathcal{X}$ (after having left out some part as the test set).
- If the sample $\mathcal{X}$ is large enough, we can randomly divide it into K (K is typically 10 or 30) parts, then randomly divide each part into two and use one half for training and the other half for validation.
- Datasets are generally not large enough to do this.
- We need to repeatedly use the same data split differently, which is called cross-validation (or rotation estimation).
- Given a dataset $\mathcal{X}$, we would like to generate K training/validation set pairs, $\{(\mathcal{T}_i, \mathcal{V}_i)\}_{i=1}^{K}$, from this dataset.
- Stratification: the class distributions in different subsets are kept roughly the same.

K-Fold Cross-Validation

- The dataset $\mathcal{X}$ is randomly partitioned into K equal-sized subsets $\mathcal{X}_i$, $i = 1, \ldots, K$, called folds.
- K training/validation set pairs:

$$\begin{aligned} \mathcal{T}_1 &= \mathcal{X}_2 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K & \mathcal{V}_1 &= \mathcal{X}_1 \\ \mathcal{T}_2 &= \mathcal{X}_1 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K & \mathcal{V}_2 &= \mathcal{X}_2 \\ &\vdots \\ \mathcal{T}_K &= \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_{K-1} & \mathcal{V}_K &= \mathcal{X}_K \end{aligned}$$

- Any two training sets $\mathcal{T}_i$ and $\mathcal{T}_j$ $(i \neq j)$ share $K - 2$ folds.

Leave-One-Out

- If N is small, K should be large to allow large enough training sets.
- One extreme case of K-fold cross-validation is leave-one-out where only one instance is left out as the validation set and the remaining $N - 1$ instances for training.
- We get N separate training/validation set pairs by leaving out a different instance at each iteration.
- This is typically used in applications such as medical diagnosis, where labeled data is hard to find.
- Stratification is not possible with leave-one-out.

5×2-Fold Cross-Validation

- For each iteration, the dataset $\mathcal{X}$ is randomly split into two equal-sized parts, $\mathcal{X}_i^{(1)}$ and $\mathcal{X}_i^{(2)}$, which leads to a 2-fold cross-validation.
- With 5 iterations, we get $K = 10$ training/validation set pairs:

$$\begin{aligned}
\mathcal{T}_1 &= \mathcal{X}_1^{(1)} & \mathcal{V}_1 &= \mathcal{X}_1^{(2)} \\
\mathcal{T}_2 &= \mathcal{X}_1^{(2)} & \mathcal{V}_2 &= \mathcal{X}_1^{(1)} \\
\mathcal{T}_3 &= \mathcal{X}_2^{(1)} & \mathcal{V}_3 &= \mathcal{X}_2^{(2)} \\
\mathcal{T}_4 &= \mathcal{X}_2^{(2)} & \mathcal{V}_4 &= \mathcal{X}_2^{(1)} \\
&\vdots \\
\mathcal{T}_9 &= \mathcal{X}_5^{(1)} & \mathcal{V}_9 &= \mathcal{X}_5^{(2)} \\
\mathcal{T}_{10} &= \mathcal{X}_5^{(2)} & \mathcal{V}_{10} &= \mathcal{X}_5^{(1)}
\end{aligned}$$

- In total, we have done 5 times of 2-fold cross-validations.

Bootstrapping

- When the sample size N is very small, a better alternative to cross-validation is bootstrapping.
- Bootstrapping generates new samples, each of size N, by drawing instances randomly from the original sample with replacement.
- The bootstrap samples usually overlap more than the cross-validation samples and hence their estimates are more dependent, but is considered the best way to do resampling for very small datasets.
- The original dataset is used as the validation set.
- Probability that an instance is not chosen after N random draws:

$$\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^N = e^{-1} \approx 0.368$$

  So each bootstrap sample (the training set) contains only approximately 63.2% of the instances.
- We can repeat the process many times and look at the average behavior. Multiple bootstrap samples are used to maximize the chance that the system is trained on all the instances.

Classifier Performance Measures

- If 0-1 loss is used, error calculations will be based on the confusion matrix (or table of confusion):

	Predicted Class	
True Class	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

- For more general loss functions, the risks should be measured instead.
- For $K > 2$ classes, the class confusion matrix is a $K \times K$ matrix such that its (i, j)-th entry contains the number of instances that belong to C_i but are assigned to C_j.

Classifier Performance Measures (2)

	Predicted Class	
True Class	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

- Error rate:

$$\text{Error rate} = \frac{\#\text{ of errors}}{\#\text{ of instances}} = \frac{|\text{FN}| + |\text{FP}|}{|\text{TP}| + |\text{FP}| + |\text{TN}| + |\text{FN}|} = \frac{|\text{FN}| + |\text{FP}|}{N}$$

  where N is the total number of instances in the validation set.

- Accuracy (ACC):

$$\text{ACC} = \frac{\#\text{ of correctness}}{\#\text{ of instances}} = \frac{|\text{TP}| + |\text{TN}|}{|\text{TP}| + |\text{FP}| + |\text{TN}| + |\text{FN}|} = \frac{|\text{TP}| + |\text{TN}|}{N}$$

Classifier Performance Measures (3)

	Predicted Class	
True Class	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

- TP rate (hit rate, recall, or sensitivity):

$$\text{TP rate} = \frac{\#\text{ of found positives}}{\#\text{ of positives}} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}$$

- FP rate (or false alarm rate):

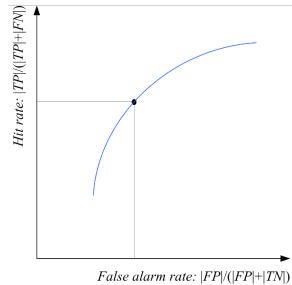
$$\text{FP rate} = \frac{|\text{FP}|}{|\text{FP}| + |\text{TN}|}$$

ROC Curve

- Receiver operating characteristics (ROC) curve:

$$\text{Hit rate} = \text{TP rate} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \quad \text{vs.} \quad \text{False alarm rate} = \text{FP rate} = \frac{|\text{FP}|}{|\text{FP}| + |\text{TN}|}$$

- ROC was original developed in signal processing for operators of radar receivers.
- The curve is obtained by varying a certain parameter (e.g., threshold for making decision) of the classification algorithm.
- Ideally, a classifier has a hit rate of 1 and an false alarm rate of 0, and hence a classifier is better the more its ROC curve gets closer to the upper-left corner.
- On the diagonal, we make as many true decisions as false ones, and this is the worst one can do, i.e., a random classifier (any classifier that is below the diagonal can be improved by flipping its decision).



- The area under curve (AUC) is often used as a performance measure.

Classifier Performance Measures (4)

	Predicted Class	
True Class	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

- Recall:

$$\text{Recall} = \text{TP rate} = \frac{\#\text{ of found positives}}{\#\text{ of positives}} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}$$

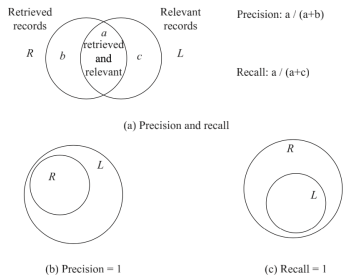
- Precision:

$$\text{Precision} = \frac{\#\text{ of found positives}}{\#\text{ of found}} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|}$$

Recall vs. Precision

- In information retrieval, there is a database of records; we make a query, for example, by using some keywords, and a system (basically a two-class classifier) returns a number of records.
- In the database, there are relevant records and for a query, the system may retrieve some of them (true positives) but probably not all (false negatives); it may also wrongly retrieve records that are not relevant (false positives).

- Precision is the number of retrieved and relevant records divided by the total number of retrieved records
- Recall is the number of retrieved relevant records divided by the total number of relevant records



Classifier Performance Measures (5)

	Predicted Class	
True Class	Yes	No
Yes	TP: True Positive	FN: False Negative
No	FP: False Positive	TN: True Negative

- Sensitivity:

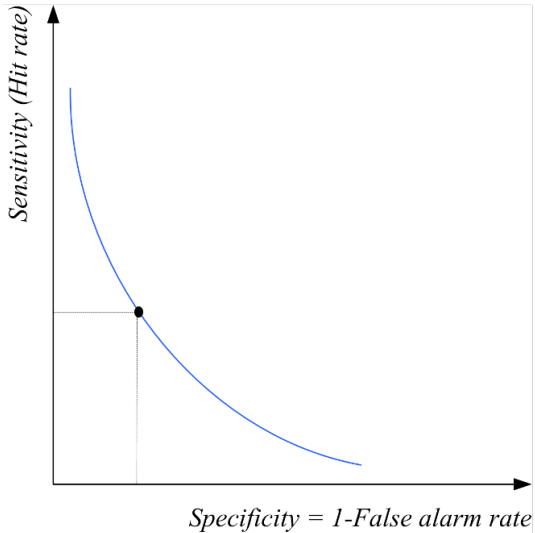
$$\text{Sensitivity} = \text{TP rate} = \frac{\#\text{ of found positives}}{\#\text{ of positives}} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}$$

- Specificity:

$$\text{Specificity} = \frac{\#\text{ of found negatives}}{\#\text{ of negatives}} = \frac{|\text{TN}|}{|\text{FP}| + |\text{TN}|} = 1 - \text{FP rate}$$

measures how well we detect the negatives.

Sensitivity vs. Specificity



Outline

Point Estimation vs. Interval Estimation

- A point estimator (e.g., MLE) specifies a value for a parameter θ.
- An interval estimator specifies an interval within which θ lies with a certain degree of confidence.
- The probability distribution of the point estimator is used to obtain an interval estimator.

Example: Estimation of Mean of Normal Distribution

- Given an i.i.d. sample $\{e_i\}_{i=1}^{N}$ of error rates obtained from N runs, where

$$e_i \sim \mathcal{N}(\mu, \sigma^2)$$

- Sample mean, a point estimator of the mean μ:

$$m = \frac{\sum_i e_i}{N} \sim \mathcal{N}(\mu, \sigma^2/N)$$

since e_i are i.i.d.

- Define a statistic $\mathcal{Z}$ with a unit normal distribution:

$$\mathcal{Z} = \frac{(m - \mu)}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

Two-Sided Confidence Interval

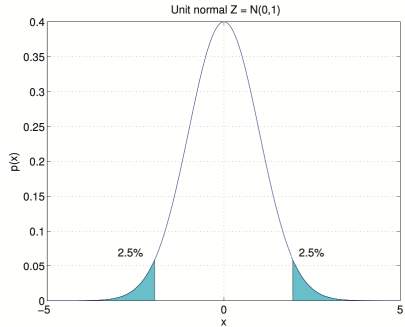
- For a unit normal distribution, about 95% of $\mathcal{Z}$ lies in $(-1.96, 1.96)$:

$$P(-1.96 < \sqrt{N}\frac{(m-\mu)}{\sigma} < 1.96) = 0.95$$

or

$$P(m - 1.96\frac{\sigma}{\sqrt{N}} < \mu < m + 1.96\frac{\sigma}{\sqrt{N}}) = 0.95$$

meaning that with 95% confidence, μ lies within $1.96\sigma/\sqrt{N}$ units of the sample mean m.



- This is a two-sided confidence interval.
- If we want more confidence, the interval gets larger.
- The interval gets smaller as N, the sample size, increases.

Two-Sided Confidence Interval (2)

- This can be generalized for any required confidence.
- Let us denote z_α such that

$$P(\mathcal{Z} > z_\alpha) = \alpha, \qquad 0 < \alpha < 1$$

- Because $\mathcal{Z}$ is symmetric around the mean, $z_{1-\alpha/2} = -z_{\alpha/2}$, and, hence, we have

$$P(\mathcal{Z} < -z_{\alpha/2}) = P(\mathcal{Z} > z_{\alpha/2}) = \alpha/2$$

- Hence, for any specified level of confidence $1 - \alpha$, we have

$$P(-z_{\alpha/2} < \sqrt{N}\frac{(m-\mu)}{\sigma} < z_{\alpha/2}) = 1 - \alpha$$

or

$$P(m - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < m - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}) = 1 - \alpha$$

- Hence, a $100(1-\alpha)$ percent two-sided confidence interval for μ can be computed for any α.

One-Sided Upper Confidence Interval

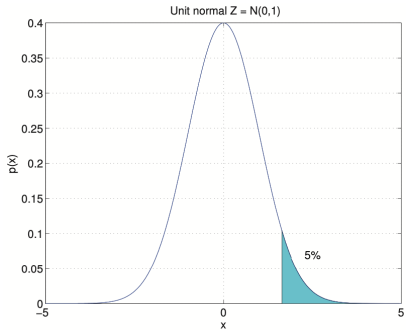
- $P(\mathcal{Z} < 1.64) = 0.95$
- we have

$$P(\sqrt{N}\frac{(m-\mu)}{\sigma} < 1.64) = 0.95$$

or

$$P(m - 1.64\frac{\sigma}{\sqrt{N}} < \mu) = 0.95$$

- $(m - 1.64\frac{\sigma}{\sqrt{N}}, \infty)$ is a 95 percent one-sided upper confidence interval for μ



One-Sided Upper Confidence Interval (2)

- The one-sided upper confidence interval defines a lower bound for μ.
- In general, a $100(1-\alpha)$ percent one-sided upper confidence interval for μ can be computed from

$$P(\sqrt{N}\frac{(m-\mu)}{\sigma} < z_\alpha) = 1-\alpha$$

or

$$P(m - z_\alpha\frac{\sigma}{\sqrt{N}} < \mu) = 1-\alpha$$

- The one-sided lower confidence interval that defines an upper bound can be calculated similarly.

t-Distribution

- In the previous intervals, we used σ; that is, we assumed the variance is known.
- When the variance σ^2 is not known, it can be replaced by the sample variance

$$S^2 = \frac{\sum_i (e_i - m)^2}{N-1}$$

- The statistic $\sqrt{N}(m-\mu)/S$ follows a t-distribution with $N-1$ degrees of freedom: (long proof omitted here ...)

$$\sqrt{N}\frac{(m-\mu)}{S} \sim \tau_{N-1}$$

- For any $\alpha \in (0,1)$ (we have used the symmetry around 0 of the t-distribution, i.e., $t_{1-\alpha/2,N-1} = -t_{\alpha/2,N-1}$),

$$P(-t_{\alpha/2,N-1} < \sqrt{N}\frac{(m-\mu)}{S} < t_{\alpha/2,N-1}) = 1-\alpha$$

or

$$P(m - t_{\alpha/2,N-1}\frac{S}{\sqrt{N}} < \mu < m + t_{\alpha/2,N-1}\frac{S}{\sqrt{N}}) = 1-\alpha$$

t-Distribution (2)

- One-sided confidence intervals can also be defined.
- The t-distribution has larger spread (longer tails) than the unit normal distribution, and generally the interval given by the t-distribution is larger; this should be expected since additional uncertainty exists due to the unknown variance.
- As N increases, the t-distribution will get closer to the normal distribution.

Outline

Hypothesis Testing

- Instead of explicitly estimating some parameters, we may want to use the sample to test some hypothesis concerning the parameters,
- E.g., instead of estimating the mean, we may want to test whether the mean is less than 0.02.
  - If the random sample is consistent with the hypothesis under consideration, we "fail to reject" the hypothesis; otherwise, we say that it is "rejected."
- When we make such a decision, we are not saying that he hypothesis is true or false, but rather that the sample data appears to be consistent with it to a given degree of confidence or not.

Hypothesis Testing (2)

- Statistical hypothesis testing:
  - We define a test statistic that obeys a certain distribution if the hypothesis is correct.
  - If the random sample is inconsistent with the hypothesis under consideration (i.e., if the statistic calculated from the sample has a very low probability of being drawn from the distribution), we reject the hypothesis.
  - Otherwise, we fail to reject it.
- Null hypothesis: the hypothesis to be tested, designated by H_0.
- Alternate hypothesis: describes what you will believe if you reject the null hypothesis, designated by H_1 or H_a. This is the statement that we hope or suspect is true instead of H_0.
- There are parametric and nonparametric tests. Only parametric tests are considered here.

Two-Sided Test

- Given a sample from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where σ is known but μ is unknown.
- We want to test a specific hypothesis about μ, whether it is equal to a specified constant μ_0, which is called the null hypothesis, i.e.,

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_0$$

for some specified constant μ_0.

Two-Sided Test (2)

- Given the sample mean m, a point estimate of μ, we reject H_0 if m is too far from μ_0.
- This is where the interval estimate is used.
- We fail to (or cannot) reject H_0 with level of significance (or significance level) α if μ_0 lies in the $100(1-\alpha)$ percent confidence interval, i.e.,

$$\sqrt{N}\frac{(m-\mu_0)}{\sigma} \in (-z_{\alpha/2}, z_{\alpha/2})$$

- We reject the null hypothesis with level of significance α if it falls outside, on either side. This is called a two-sided test or two-tailed test.
- We call $(-\infty, -z_{\alpha/2})$ as well as $(z_{\alpha/2}, \infty)$ the region of rejection or critical region and $-z_{\alpha/2}, z_{\alpha/2}$ the critical values.

Two-Sided Test (3)

- In hypothesis testing, the *p*-value, or probability value, is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.
- In the two-sided test, the *p*-value is computed by

$$p = P(z > \left|\sqrt{N}\frac{(m-\mu_0)}{\sigma}\right|) + P(z < -\left|\sqrt{N}\frac{(m-\mu_0)}{\sigma}\right|) = 2P(z > \left|\sqrt{N}\frac{(m-\mu_0)}{\sigma}\right|)$$

- The decision rule:
  - we fail to reject the null hypothesis H_0 with level of significance α if $p > \alpha$, and
  - we reject H_0 if $p \leq \alpha$.

Type I and Type II Errors

	Truth about H_0	
Decision about H_0	True	False
Fail to reject	Correct inference	Type II error
Reject	Type I error	Correct inference

- Type I error:
  - We reject the null hypothesis when it is correct, e.g., $\mu = \mu_0$.
  - the predefined significance level α defines how much type I error we can tolerate, i.e., probability of type I error, typical values being $\alpha = 0.1, 0.05, 0.01$.
  - It is equivalent to the false positive (FP) rate.
- Type II error:
  - We fail to reject the null hypothesis when it is incorrect, e.g., $\mu \neq \mu_0$.
  - It is equivalent to the false negative (FN) rate.

One-Sided Test

- We can also define a one-sided test or one-tailed test.
- Null and alternative hypotheses:

$$H_0 : \mu \leq \mu_0$$

$$H_1 : \mu > \mu_0$$

  as opposed to the two-sided test when the alternative hypothesis is $\mu \neq \mu_0$.
- The one-sided test with level of significance α defines a $100(1-\alpha)$ confidence interval bounded on one side in which m should lie for the hypothesis not to be rejected.
- We fail to reject H_0 with level of significance α if

$$\frac{\sqrt{N}(m-\mu_0)}{\sigma} \in (-\infty, z_\alpha)$$

- Otherwise, we reject H_0 with region of rejection (z_α, ∞).

One-Sided Test (2)

- In this one-sided test, the p-value is computed by

$$p = P(z > \sqrt{N}\frac{(m - \mu_0)}{\sigma})$$

- The decision rule:
  - we fail to reject the null hypothesis H_0 with level of significance α if $p > \alpha$, and
  - we reject H_0 if $p \leq \alpha$.
- In this one-sided test, the null hypothesis H_0 can be replaced by equality, i.e., $\mu = \mu_0$, which means that we get ordering information only if the test rejects.
- This tells us which of the two one-sided tests we should use. In the above test, it is a right-tailed test.
- Whatever claim we have should be in H_1 so that rejection of the test would support our claim.

t Tests

- If the variance σ^2 is not known, the sample variance S^2 will be used instead and so

$$\sqrt{N}\frac{(m-\mu_0)}{S} \sim \tau_{N-1}$$

- E.g., we can define a two-sided t test:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

We fail to reject at significance level α if

$$\sqrt{N}\frac{(m-\mu_0)}{S} \in (-t_{\alpha/2,N-1}, t_{\alpha/2,N-1})$$

- One-side t test can be defined similarly.

Outline

Performance Evaluation and Comparison

- We have reviewed hypothesis testing, we will discuss how it is used in testing error rates.
- We will discuss the case of classification error, but the same methodology applies for other learning paradigms.
- We now start with error rate assessment for performance evaluation, and, then, we discuss error rate comparison for performance comparison.

t Test

- Multiple training/validation set pairs: if we run the algorithm K times on K training/validation set pairs, we get K error percentages, p_i, $i = 1, \ldots, K$, on the K validation sets.
- Bernoulli variables:

$$x_i^{(\ell)} = \begin{cases} 1 & \text{if classifier trained on } \mathcal{T}_i \text{ makes an error on instance } \ell \text{ of } \mathcal{V}_i \\ 0 & \text{otherwise} \end{cases}$$

- Test statistic:

$$\sqrt{K}\frac{(m - p_0)}{S} \sim \tau_{K-1}$$

where

$$m = \frac{\sum_{i=1}^{K} p_i}{K}, \quad S^2 = \frac{\sum_{i=1}^{K}(p_i - m)^2}{K - 1}, \quad p_i = \frac{\sum_{\ell=1}^{N} x_i^{(\ell)}}{N}$$

t Test (2)

- We can use the t test to determine whether to reject the null hypothesis H_0 that the classifier has error percentage p_0 or less at significance level α.
- We can define a one-sided t test:

$$H_0 : p \leq p_0$$

$$H_1 : p > p_0$$

We fail to reject at significance level α if

$$\sqrt{K}\frac{(m - p_0)}{S} \in (-\infty, t_{\alpha,K-1})$$

Outline

K-Fold Cross-Validated Paired t Test

- K-fold cross-validation is used to obtain K training/validation set pairs, $\{(\mathcal{T}_i, \mathcal{V}_i)\}_{i=1}^{K}$.
- For each training/validation set pair $(\mathcal{T}_i, \mathcal{V}_i)$, the two classification algorithms are trained on $\mathcal{T}_i$ and tested on $\mathcal{V}_i$, with error percentages p_i^1 and p_i^2.
- We define

$$p_i = p_i^1 - p_i^2, \quad i = 1, \ldots, K$$

  The distribution of p_i is used for hypothesis testing.
- This is a paired test; that is, for each i, both algorithms see the same training and validation sets.
- If p_i^1 and p_i^2 are both normally distributed, then p_i should also be normal (with mean μ).

K-Fold Cross-Validated Paired t Test (2)

- If we want to test whether the two classification algorithms have the same error rate, then we expect them to have the same mean, or equivalently, that the difference of their means is 0.
- Null and alternative hypotheses:

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

- Test statistic (under the null hypothesis that $\mu = 0$):

$$\sqrt{K}\frac{(m-0)}{S} = \frac{\sqrt{K}m}{S} \sim \tau_{K-1}$$

where

$$m = \frac{\sum_{i=1}^{K} p_i}{K}, \quad S^2 = \frac{\sum_{i=1}^{K}(p_i - m)^2}{K-1}$$

- K-fold CV paired t test:
  reject H_0 at significance level α if $\sqrt{K}m/S \notin (-t_{\alpha/2,K-1}, t_{\alpha/2,K-1})$;
  fail to reject otherwise.

K-Fold Cross-Validated Paired t Test (3)

- If we want to test whether the first algorithm has less error than the second, we need a one-sided hypothesis and use a one-tailed test.
- Null and alternative hypotheses:

$$H_0 : \mu \geq 0$$

$$H_1 : \mu < 0$$

- K-fold CV paired t test:
  reject H_0 at significance level α if $\sqrt{K}m/S \notin (t_{-\alpha,K-1}, +\infty)$;
  fail to reject otherwise.

Comparing Multiple Algorithms

- When L algorithms are compared based on K training/validation set pairs, we get L groups of K error rates each.
- The problem is to compare the L samples for statistically significant difference.
- One method that is often used is based on analysis of variance (ANOVA).