Design & Analysis of Data Science Experiments

Model Inter-comparison

Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

NC STATE UNIVERSITY
Department of Computer Science

Assessing Model Performance

TESTING ERROR RATES

Testing Error Rates: Application

- **Classification error**
- **Squared error in regression**
- **Log likelihood in unsupervised learning**
- **Expected reward in reinforcement learning**

Book Chapter Sections

19.10.1. Binomial Test for Single Classifier, Single Validation Set
19.10.2. Approximate Normal Test for Single Classifier, Single Validation Set
19.10.3. t-Test for Single Classifier with K-fold Cross-Validation
19.11.1. McNemar's Test to Compare Two Classifiers w/ Single Validation Set
19.11.2. K-fold Cross-Validated Paired t-test
19.11.3. 5x2 cv Paired t-test

19.12 Comparing Multiple Algorithms: Analysis of Variance
19.13.1 Nonparametric tests for comparison over multiple data sets: Wilcoxon Signed Rank Test for Two Algorithm Comparison
19.13.2. Nonparametric tests for comparison over multiple data sets: Kruskal-Wallis Test for Multiple Algorithm Comparison
19.14.1 Multi-criterion Assessment: Multivariate Test for Two Algorithms
19.14.2 Multi-criterion Assessment: Multivariate Test for Multiple Algorithms (MANOVA)

Single Classifier Assessment

BINOMIAL TEST FOR SINGLE CLASSIFIER USING SINGLE VALIDATION SET (19.10.1)

help (binom.test)

- http://www.instantr.com/2012/11/06/performing-a-binomial-test/

When and how to use?

- The binomial test is used to assess the error rate of a **single dichotomous (binary) classification algorithm**.
- Assumptions
  - The training is done on the **single training set**
  - The **single validation set** (non-overlapping with the training set) is used for the assessment.

- Hypothesis Testing using **binom.test()**:
  - Let's assume that p is the probability that the classifier makes a misclassification error
  - Estimate or test a hypothesis about the probability of a success in a Bernoulli experiment

Binomial Test

- Let x^t denote the correctness of the classifier decision:
  - it is a 0/1 Bernoulli random variable
  - it is 1 when the classifier commits the error; 0, otherwise
- Let $X = \sum_{t=1}^{N} x^t$ denote the binomial random variable for the total number of errors
- The Null Hypothesis (H_0):
  - The error probability p: $p \leq p_0$ (e.g., $p_0 = 0.5$)
- The Alternative Hypothesis (H_1): $p > p_0$
- If the probability of error is p then
  - $P\,(X = j) = \binom{N}{j} p^j (1-p)^{N-j}$
- Binomial test reject H_0 if for the significance value of α (e.g., $\alpha = 0.05$):
  - $P\,(X \geq e) = \sum_{j=e}^{N} \binom{N}{j} p_0^j (1-p_0)^{N-j} < \alpha$

binom.test()

binom.test (x, n, p = 0.5, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)

- **x - The parameter can be a single value indicating number of successes or a vector with 2 values indicating number of successes and failures**
- **n - The parameter is used to specify the total number of trials. It is unused if the first parameter is a vector.**
- **p - The hypothesis for the probability of success. It has a default value of 0.5.**
- **alternative - This parameter should be one of “two.sided”, “less” or “greater”**
  - **two.sided implies the true probability of success should not be equal to the value of parameter ‘p’**
  - **less implies the true probability of success should be lesser than the value of parameter ‘p’**
  - **greater implies the true probability of success should be greater than the value of parameter ‘p’**
- **conf.level - The confidence level which has a default value of 0.95**

Single Classifier Assessment

APPROXIMATE NORMAL TEST FOR SINGLE CLASSIFIER USING SINGLE VALIDATION SET (19.10.2)

Assumptions & Problem Statement

- **Assumptions:**
  - **The training is done on the single training set**
  - **The single validation set (non-overlapping with the training set) is used for the assessment.**
  - **For large N, $\frac{X}{N} \sim N\left(p_0, \frac{p_0(1-p_0)}{N}\right)$ is approximately normal distribution with the mean p_0 and the variance $\frac{p_0(1-p_0)}{N}$**
    - where X is the total number of errors
- **Hypothesis Testing using qqnorm():**
  - **Let's assume that p is the probability that the classifier makes a misclassification error**
  - **Estimate or test a hypothesis about the probability p**

Approximately Normal Test

Step 1. H_0: $P<=P_0$ vs H_1: $P>P_0$

Step 2. H_0: One-tail test. The critical value is 1.64 for $\alpha=0.05$

Step 3. The test value is

$$z_{\text{stat}} = \frac{X/N - p_0}{\sqrt{p_0(1-p_0)/N}}$$

Step 4. Reject the null hypothesis if $Z_{stat}>1.64$, otherwise fail to reject

- **Assumptions:**
  - **N is not too small and p is not very close to 0 or 1**
  - **as a rule of thumb: Np>=5 and N(1-p)>=5.**

Single Classifier Assessment

T-TEST FOR SINGLE CLASSIFIER USING K-FOLD CROSS-VALIDATION SETS (19.10.3)

help (qqnorm)
help(t.test)

When and how to use?

- The t-test is used to assess the error rate of a single classification algorithm using K-fold cross-validation results.
- Assumptions
  - The training is done on the K-fold training set
  - The K-fold cross-validation result sets (non-overlapping with the corresponding training set for each fold) are used for the assessment.

- Hypothesis Testing using t.test():
  - classification algorithm has p_0 or less error percentage at significance level α using the t-statistic with parameters α and (K-1): $t_{\alpha,K-1}$

t-Test using t-statistic: $t_{\alpha,(K-1)}$

- Let the algorithm run K times, on K training/ validation set pairs:
  - then $p_i, i = 1, 2, \dots, K$ are error percentages on K validation sets
- Let x_i^t denote the correctness of the classifier on the i^{th} training set:
  - it is 1 when the classifier commits the error on the corresponding i^{th} validation set of the pair; 0, otherwise
  - Then $p_i = \frac{1}{N}\sum_{t=1}^{N} x_i^t$ is the error percentage for fold i
- The mean and the variance across all the K folds:
  - $m = \frac{1}{K}\sum_{i=1}^{K} p_i$
  - $S^2 = \frac{1}{K-1}\sum_{i=1}^{K}(p_i - m)^2$
- The t-statistic with (K-1) degrees of freedom: $t_{K-1} \sim \frac{\sqrt{K}(m-p_0)}{S}$
- The Null Hypothesis (H_0): $p < p_0$
- The t-test rejects H_0 at significance level α (e.g., $\alpha = 0.05$):
  - if $t_{K-1} > t_{\alpha,(K-1)}$ (e.g., $t_{0.05,9} = 1.83$ for K=10)

TWO Classifier Assessment

MCNEMAR'S TEST FOR TWO CLASSIFIERS USING SINGLE VALIDATION SET (19.11.1)

help (mcnemar.test)

When and how to use?

- The McNemar's test is used to assess whether the **two classification algorithms** have the same error rate using a **single validation set**.
- Assumptions
  - Two different classification algorithms
  - The training is done on the single training set
  - The single validation set (non-overlapping with the training set) is used for the assessment.

- Hypothesis Testing using mcnemar.test():
  - McNemar's test rejects the hypothesis that the two classification algorithms have the same error rate at significance level α if the proper χ^2-statistic is greater than $\chi^2_{\alpha,1}$

McNemar's Test using χ_1^2-statistic

e_{00}: number of examples misclassified by both	e_{01}: number of examples misclassified by 1 but not 2
e_{10}: number of examples misclassified by 2 but not 1	e_{11}: number of examples correctly classified by both

- **Under the null hypothesis (H_0):**
  - **both classification algorithms have the same error rate**
  - $e_{01} = e_{10} = (e_{01} + e_{10})/2$
- **The chi-squared statistic with one degree of freedom:**

$$\boxed{\chi_1^2 \sim \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}}}$$

- **McNemar's test rejects the hypothesis that the two classification algorithms have the same error rate at significance level α if the proper χ_1^2-statistic is greater than $\chi_{\alpha,1}^2$:**
  - **for $\alpha = 0.05$, $\chi_{0.05,1}^2 = 3.84$**

TWO Classifiers Assessment

PAIRED *T*-TEST FOR TWO CLASSIFIERS USING K-FOLD CROSS-VALIDATION SETS (19.11.2)

t.test (..., paired=TRUE)

When and how to use?

- **The paired t. test is used to assess whether the two classification algorithms have the same error rate using K-fold cross-validation results.**
- **Assumptions**
  - **Two different classification algorithms**
  - **The training is done on the K-fold training set**
  - **The K-fold cross-validation result sets (non-overlapping with the corresponding training set for each fold) are used for the assessment.**
- **Hypothesis Testing using t.test(paired=TRUE):**
  - **Use *paired test*; that is, for each *i*, both algorithms see the same training and validation sets**

Paired t-Test using t-statistic: t_{K-1}

- Let two algorithms run K times, on K training/ validation set pairs:
  - then p_i^1 and p_i^2, $i = 1, 2, \dots, K$ are error percentages on K validation sets for algorithm 1 and 2, resp.
- If 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.
  - The difference in error rates on fold i: $p_i = p_i^1 - p_i^2$
  - Assuming that each p_i^1 and p_i^2 is normal, so is p_i
- The Null Hypothesis (H_0): p_i has the mean $\mu = 0$
  - H_1: $\mu \neq 0$
- The mean and the variance across all the K folds:
  - $m = \frac{1}{K}\sum_{i=1}^{K} p_i$
  - $S^2 = \frac{1}{K-1}\sum_{i=1}^{K}(p_i - m)^2$
- The t-statistic with (K-1) degrees of freedom: $t_{K-1} \sim \frac{\sqrt{K}(m-\mu)}{S} = \frac{\sqrt{K}m}{S}$

K-fold CV Paired t-test: $t_{\alpha,(K-1)}$

- **Two classification algorithms have the same error rate at significance level α if t_{K-1} value is outside the interval:**
  - $\left(-t_{\frac{\alpha}{2},K-1};\ t_{\frac{\alpha}{2},K-1}\right)$
  - **Example:** $K = 10$ **and** $\alpha = 0.05$, $t_{0.025,9} = 2.26$

TWO Classifiers Assessment

5X2 CV PAIRED *T*-TEST FOR TWO CLASSIFIERS (19.11.3)

t.test (..., paired=TRUE)

When and how to use?

- **The paired t. test is used to assess whether the two classification algorithms have the same error rate using 5x2 cross-validation results.**
- **Assumptions**
  - **Two different classification algorithms**
  - **The 5x2 cross-validation result sets are used for assessment**
- **Hypothesis Testing using t.test(paired=TRUE):**
  - **Use *paired test*; that is, for each *i*, both algorithms see the same training and validation sets**

5x2 CV Paired t Test

- **5x2 CV Paired t-Test is a simplified variation of the k-fold cross-validation t-test.**
- **To perform the test, the following procedure is followed:**
  1. **Perform 2-fold cross validation 5 times:**
     - Generating 10 (training-testing) data set pairs
  2. **For each pair:**
     - Train both algorithms on the training data
     - Test on the test data
     - Compute the difference between the accuracies of the two algorithms

5x2 CV Paired t Test

- **Once you have these differences, compute the t-statistic of the first training-test pair (the difference of accuracies between the two classifiers on the first run's first fold)**

$$\frac{\Delta_{1,1}}{\left(\frac{1}{5}\sum_{i=1}^{5}\left(\Delta_{1,i}+\frac{\Delta_{1,i}+\Delta_{2,i}}{2}\right)^{2}+\left(\Delta_{2,i}+\frac{\Delta_{1,i}+\Delta_{2,i}}{2}\right)^{2}\right)^{.5}}$$

Paired t-Test using t-statistic: t_{K-1}

- Let two algorithms run K times, on K training/ validation set pairs:
  - then p_i^1 and p_i^2, $i = 1, 2, \ldots, K$ are error percentages on K validation sets for algorithm 1 and 2, resp.
- If 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.
  - The difference in error rates on fold i: $p_i = p_i^1 - p_i^2$
  - Assuming that each p_i^1 and p_i^2 is normal, so is p_i
- The Null Hypothesis (H_0): p_i has the mean $\mu = 0$
  - H_1: $\mu \neq 0$
- The mean and the variance across all the K folds:
  - $m = \frac{1}{K}\sum_{i=1}^{K} p_i$
  - $S^2 = \frac{1}{K-1}\sum_{i=1}^{K}(p_i - m)^2$
- The t-statistic with (K-1) degrees of freedom: $t_{K-1} \sim \frac{\sqrt{K}(m-\mu)}{S} = \frac{\sqrt{K}m}{S}$