Project Title: Multiple Classifier Algorithm Comparison

Ref:

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/tests-of-means/why-use-paired-t/

Chapter 19 Design and Analysis of Machine Learning Experiments provided by Dr. Samatova

http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric6.html

What is the null hypothesis and the alternative hypothesis in each of the statistical test?

1. **Paired t test**
   Null Hypothesis: H_0: $\mu_d = \mu_0$
   The two populations have the same mean of differences in vector dimensional space.
   Alternative Hypothesis: H_1: $\mu_d \neq \mu_0$
   The population mean of the differences (μ_d) does not equal the hypothesized mean of the differences (μ_0).
2. **ANOVA Test**
   Null Hypothesis: H_o is that all means are equal.
   H_o: $\mu_1 = \mu_2 = \cdots = \mu_L$
   Alternative Hypothesis: H_1: $\mu_r \neq \mu_s$, for at least one pair (r, s).
3. **Wilcoxon Signed Rank Test**
   Null Hypothesis: The median difference is zero.
   Alternative Hypothesis: The median difference is positive.

What did you conclude after performing the tests? State which hypothesis seemed favorable given the evidence.

1. **Paired t test**
   The p-value 0.3434 obtained is more than the significant value. Thus, we fail to reject the null hypothesis that errors have same mean of differences.
2. **ANOVA Test**
   The p-value 2e-16 obtained is less than the significant value. Thus, we reject the null hypothesis that all classifiers have same mean. Thus, there exists atleast one pair of classifiers which have different mean error. Pairs can be found through TukeyHSD test.
3. **Wilcoxon Signed Rank Test**
   The p-value 0.0625 obtained is more than the significant value. Thus, we fail to reject the null hypothesis that the median difference of mean errors is zero.

Briefly explain why signed rank test is more useful than ANOVA test while comparing two classifier algorithms on multiple data sets. (Hint: See Section 19.13 of Chapter 19 Design and Analysis of Machine Learning Experiments)

Suppose, we have two classifier algorithms on several datasets say 10. Out of 10, 9 datasets are of nearly same size. However, one dataset is huge enough. The classifier takes about a day to run on the large dataset while few minutes in the smaller one. Now if we compare training time of the two classifiers in parametric test like ANOVA, the large dataset will dominate the average training time. However, if we use non-parametric test like sign rank test which usually order or rank the effect of two classifiers, thus generating a normalizing effect for each dataset. Therefore, it becomes obvious that a normalizing effect is much more desirable in comparison to avoid any type of bias.