CS697A – Topic in Computer Science – Machine Learning

Assignment 1

Submitted by: Prachi Panse

NetID: ve3568

In the question, accuracies are given for the 3 classifiers on 10 folds. To calculate the error rates, we will need to compute it with the formula: Error rate = 1- accuracies.

Initial calculations include below table:

Folds	NB	DecTree	NearestNeighbor	Total
1	0.3191	0.2476	0.2836	
2	0.2983	0.1306	0.1117	
3	0.2988	0.3197	0.1590	
4	0.3087	0.0898	0.3175	
5	0.3667	0.2242	0.2401	
6	0.3585	0.1846	0.1521	
7	0.2784	0.3776	0.2988	
8	0.2786	0.2415	0.5041	
9	0.3422	0.0620	0.0721	
10	0.2135	0.2476	0.2545	
Mean	**0.3063**	**0.2125**	**0.2394**	**0.2527**
stddev	**0.04486**	**0.09854**	**0.12484**	
$\sum X$	3.0628	2.1252	2.3935	**7.5815**
$(\sum X)^2$	9.3807	4.5165	5.7288	
$\sum(X^2)$	**0.9562**	**0.5390**	**0.7131**	**2.2084**

Q1 [2.5pts]: Use ANOVA to determine if the three classifiers have equal error rates.

Answer:

Considering, H_0: there is not significant difference between the three classifiers and three classifiers have equal error rates.

H_1: There is significant difference between the three classifiers and three classifiers have different error rates.

Correction Factor (CF) = $(\sum x_{total})^2 / n_{total}$= 57.4791 / 30 = 1.9160

$SS_{Total} = \sum x^2_{total} - CF$=0.2527 - 1.9160 = 0.2924

$SS_{between} = (((\sum x_{NB})^2/n_{NB}) + ((\sum x_{DT})^2/n_{DT})) + ((\sum x_{NN})^2/n_{NN}))) - CF$ = 0.0466

SS_{within} = Sum of $Squares_{Total}$ - Sum of $squares_{between}$ = 0.2924 - 0.0466 = 0.2458

df for $SS_{between}$ = number of groups – 1 = 2

df for SS_{within} = total number of cases – number of groups = 27

df for SS_{total} = df for $SS_{between}$ + df for SS_{within} =29

$S^2_{between} = SS_{between} / df_{between}$ = 0.0466 / 2 = 0.0233

$S^2_{within} = SS_{within} / df_{within}$ = 0.2458 / 27 = 0.0091

F = variance between groups / variance within groups = 0.0233 / 0.0091 = 2.5618

Source of Variation	Sum of squares	Degree of Freedom	Mean Square	F0
Between Groups	0.04663	2.000	0.023	2.562
Within Groups	0.24575	27.000	0.009	
Total	0.292	29.000		

From the f distribution table value of f for df1=2 and df2 = 27 is 2.51061.
Whereas, from the above table f value we got is: 2.562 which is greater than table value.
Hence, we can reject H_0 that is null hypothesis and conclude that there is a significant difference between the three classifiers and three classifiers have different error rates.

Submitted by:
Prachi Panse VE3568

Q2 [2pts]:
Q2a) Use Cross-Validated Paired t-test to determine if NB and DecTree have equal Errors

Answer:

Considering, H_0: There is not significant difference between error rates of NB and DecTree.

H_1: There is significant difference between error rates of NB and DecTree.

We calculated the difference between both the groups and sum up the difference.

∑difference = 0.937600

Mean $_{difference}$ = ∑difference / n = 0.093760

Standard deviation $_{difference}$ = 0.1225287

Standard error of the mean = 0.038747

t score = stddev / std error of the mean = 2.419802

df = 9

Pair 1	Pair Differences			t	df
	Mean	stddev	std error of the mean		
NB and DecTree	0.09376	0.12253	0.03875	2.41980	9

From the t distribution critical values table, value of t, for df=9 and alpha/2 = 0.025 is 2.262.
Whereas, from the above table t value we got is: 2.419 which is greater than the table value.
Hence, we will reject H_0 i.e null hypothesis and conclude that there is significant difference between error rates of NB and DecTree also NB and DecTree have different error rates.
On the other hand, if we change the value of alpha/2 to 0.01, value of t, as the table will become 2.821 which is greater than the t value we calculated. Therefore, in this case we accept H_0 i.e null hypothesis and conclude that there is no significant difference between error rates of NB and DecTree also NB and DecTree have equal error rates.

Submitted by:
Prachi Panse VE3568

Q2b) Use Cross-Validated Paired t-test to determine if DecTree and nearestNeighbor have equal errors

Answer:

Considering, H_0: There is not significant difference between error rates of DecTree and nearestNeighbor.

H_1: There is significant difference between error rates of DecTree and nearestNeighbor.

We calculated the difference between both the groups and sum up the difference.

∑difference = -0.268300

Mean $_{difference}$ = ∑difference / n = -0.026830

Standard deviation $_{difference}$ = 0.1285619

Standard error of the mean = 0.040655

t score = stddev / std error of the mean = -0.659946

df = 9

Pair 1	Pair Differences			t	df
	Mean	stdev	std error of the mean		
DecTree and nearestNeighbor	-0.02683	0.12856	0.04065	-0.65995	9

From the t distribution critical values table, value of t, for df=9 and alpha/2 = 0.025 is 2.262.
Whereas, from the above table t value we got is: -0.65995 which is smaller than the table value.
Hence, we accept H_0 i.e null hypothesis and conclude that there is no significant difference between error rates of DecTree and nearestNeighbor also DecTree and nearestNeighbor have equal error rates.

Submitted by:
Prachi Panse VE3568

Q3) [3pts]: For each classifier (Naive Bayes, Decision Tree, Knearest Neighbor), determine if the error of the classifier less than p0 (=0.1, 0.2, 0.3) with level of significance (alpha) (=0.01 or 0.025)

Answer:

Classifier	Mean	stddev	std error of the mean	p0 value	t	DF= 9, Alpha/2 = 0.01	H0 Status for Alpha = 0.01	DF =9, Alpha/2 = 0.025	H0 Status for Alpha = 0.025
NB	0.30628	0.04486	0.01418	0.1	14.5422	2.821	Reject	2.262	Reject
				0.2	7.4924	2.821	Reject	2.262	Reject
				0.3	0.4427	2.821	Accept	2.262	Accept
DecTree	0.21252	0.09854	0.03116	0.1	3.6111	2.821	Reject	2.262	Reject
				0.2	0.4018	2.821	Accept	2.262	Accept
				0.3	-2.8075	2.821	Accept	2.262	Accept
NearestNeighbour	0.23935	0.12484	0.03948	0.1	3.5299	2.821	Reject	2.262	Reject
				0.2	0.9968	2.821	Accept	2.262	Accept
				0.3	-1.5363	2.821	Accept	2.262	Accept

By checking the student t table, for DF = 9 and Alpha = 0.01, value = 2.821 and for Alpha = 0.025, value = 2.262.

Comparing this value with t score to conclude if we can reject the null hypothesis or we will fail to reject the null hypothesis.

For all the t score values greater than student t table value, will be rejected and for all less than values, null hypothesis is accepted.

Irrespective of the alpha value, results are same for all the records.

Submitted by:
Prachi Panse VE3568