Assignment - 1

Prepared By: Krishna Tank
NetID: zy7886

You are given the accuracies of three classifiers above on each of the 10 folds.

	Accuracies		
Fold	**NB**	**DecTree**	**NearestNeighbor**
1	0.6809	0.7524	0.7164
2	0.7017	0.8694	0.8883
3	0.7012	0.6803	0.841
4	0.6913	0.9102	0.6825
5	0.6333	0.7758	0.7599
6	0.6415	0.8154	0.8479
7	0.7216	0.6224	0.7012
8	0.7214	0.7585	0.4959
9	0.6578	0.938	0.9279
10	0.7865	0.7524	0.7455

Q1: Use ANOVA to determine if the three classifiers have equal error rates.

- Here, One way anova in Spss software.

Now,

- Accuracies are dependent variables.
- Classifiers (NB, DecTree, NearestNeighbor) is an Independent variable.

Data Interpretation:

- Classifiers as nominal measure.
  - 1 = NB
  - 2 = DecTree
  - 3 = NearestNeighbour
- Accuracies as scale measure.

Descriptives

Error Rates (Accuracies)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
NB	10	0.693720	0.0448568	0.0141850	0.661631	0.725809	0.6333	0.7865
DecTree	10	0.787480	0.0985356	0.0311597	0.716992	0.857968	0.6224	0.9380
NearestNeighbour	10	0.760650	0.1248369	0.0394769	0.671347	0.849953	0.4959	0.9279
Total	30	0.747283	0.1004104	0.0183324	0.709789	0.784777	0.4959	0.9380

Now, let's make an Anova table.

<table>
<tr><th colspan="6">ANOVA</th></tr>
<tr><th colspan="6">Error Rates</th></tr>
<tr><th></th><th>Sum of squares</th><th>df</th><th>Mean Square</th><th>F</th><th>p-value</th></tr>
<tr><td>Between Groups</td><td>0.047</td><td>2</td><td>0.023</td><td>2.562</td><td>0.096</td></tr>
<tr><td>Within Groups</td><td>0.246</td><td>27</td><td>0.009</td><td></td><td></td></tr>
<tr><td>Total</td><td>0.292</td><td>29</td><td></td><td></td><td></td></tr>
</table>

- From the anova table mentioned earlier,
    - F - value = 2.562 ≠ 2.51061 (From F-Table for alpha = 0.1)
- So here, we can say that we reject the null hypothesis and conclude that there is a significant difference between the three classifiers.
- All three classifiers have not equal error rates.

Que 2:

Q2a) Use Cross-Validated Paired t-test to determine if NB and DecTree have equal Errors.

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	NB	0.693720	10	0.0448568	0.0141850
	DecTree	0.787480	10	0.0985356	0.0311597

Now, paired test:

Paired Samples Test									
		Paired Differences					t	df	p-value (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	NB - DecTree	-0.0937600	0.1225287	0.0387470	-0.1814118	-0.0061082	-2.420	9	0.039

From the above paired t-test:

- T-value = -2.420 ≠ 2.262 (for alpha = 0.025)
- T-value = -2.420 ≠ 2.821 (for alpha = 0.01)
- So we can reject the null hypothesis and conclude that there is a difference between error rates of NB and DecTree.
- NB and Dectree have different error rates.

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

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	DecTree	0.787480	10	0.0985356	0.0311597
	NearestNeighbour	0.760650	10	0.1248369	0.0394769

Now, paired test:

Paired Samples Test									
		Paired Differences					t	df	P-value (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	DecTree - NearestNeighbour	0.0268300	0.1285619	0.0406548	-0.0651376	0.1187976	0.660	9	0.526

From the above paired t-test

- t-value = 0.660 ≠ 2.262 (for alpha = 0.025)
- t-value = 0.660 ≠ 2.821 (for alpha = 0.01)
- So we reject the null hypothesis and say that there is a difference between error rates of DecTree and NearestNeighbour.
- DecTree and NearestNeighbour have not the same error rates.

Q3): 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)

(i) Naive Bayes Classifier:

- Here we are given the error rate of fold i, pi for i from 1 through 10.
- With m and s as average and standard deviation,
  - Hypothetical mean: 1.000000
  - Actual mean: 0.693720
  - Difference between Hypothetical mean and actual mean: m = 0.30628
  - Standard Deviation: s = 0.044857
  - 95% confidence interval of this difference:
    - From -0.338369 to -0.274191

Now, for p0 = 0.1, 0.2, 0.3 we will calculate the t-value.

- For p0 = 0.1
  - t = ($\sqrt{k}$ (m - p0))/ s
  - So,
    - t = 14.541

  - Now, by taking value of alpha = 0.01,
    - From t-table, the value of t = 2.821
    - Now, t here has less value than t we calculated.
    - So, error of the classifier is not less than p0 = 0.1
  - Now, by taking value of alpha = 0.025,
    - From t-table, the value of t = 2.262
    - Now, t here has less value than t we calculated.
    - So, error of the classifier is not less than p0 = 0.1

- For p0 = 0.2
  - t = ($\sqrt{k}$ (m - p0))/ s
  - So,
    - t = 7.492406

  - Now, by taking value of alpha = 0.01,
    - From t-table, the value of t = 2.821
    - Now, t here has less value than t we calculated.
    - So, error of the classifier is not less than p0 = 0.2
  - Now, by taking value of alpha = 0.025,

        - From t-table, the value of t = 2.262
        - Now, t here has less value than t we calculated.
        - So, error of the classifier is not less than p0 = 0.2

- For p0 = 0.3
    - t = ($\sqrt{k}$ (m - p0))/ s
    - So,
        - t = 0.442720

    - Now, by taking value of alpha = 0.01,
        - From t-table, the value of t = 2.821
        - Now, t here has more value than t we calculated.
        - So, error of the classifier is less than p0 = 0.3
    - Now, by taking value of alpha = 0.025,
        - From t-table, the value of t = 2.262
        - Now, t here has more value than t we calculated.
        - So, error of the classifier is less than p0 = 0.3

(ii) Decision Tree:

- Here we are given the error rate of fold i, pi for i from 1 through 10.
- With m and s as average and standard deviation,
    - Hypothetical mean: 1.000000
    - Actual mean: 0.787480
    - Difference between Hypothetical mean and actual mean: m = -0.21252
    - Standard Deviation: s = 0.0985356
    - 95% confidence interval of this difference:
        - From -0.283008 to -0.142032

Now, for p0 = 0.1, 0.2, 0.3 we will calculate the t-value.

- For p0 = 0.1
    - t = ($\sqrt{k}$ (m - p0))/ s
    - So,
        - t = 3.611075

    - Now, by taking value of alpha = 0.01,
        - From t-table, the value of t = 2.821
        - Now, t here has less value than t we calculated.
        - So, error of the classifier is not less than p0 = 0.1
    - Now, by taking value of alpha = 0.025,
        - From t-table, the value of t = 2.262
        - Now, t here has less value than t we calculated.
        - So, error of the classifier is not less than p0 = 0.1

- For p0 = 0.2
    - t = ($\sqrt{k}$ (m - p0))/ s
    - So,
        - t = 0.4018011

    - Now, by taking value of alpha = 0.01,
        - From t-table, the value of t = 2.821
        - Now, t here has less value than t we calculated.
        - So, error of the classifier is not less than p0 = 0.2
    - Now, by taking value of alpha = 0.025,
        - From t-table, the value of t = 2.262
        - Now, t here has less value than t we calculated.
        - So, error of the classifier is not less than p0 = 0.2

- For p0 = 0.3
    - t = ($\sqrt{k}$ (m - p0))/ s
    - So,
        - t = -2.80747

    - Now, by taking value of alpha = 0.01,
        - From t-table, the value of t = 2.821
        - Now, t here has more value than t we calculated.
        - So, error of the classifier is less than p0 = 0.3
    - Now, by taking value of alpha = 0.025,
        - From t-table, the value of t = 2.262
        - Now, t here has more value than t we calculated.
        - So, error of the classifier is less than p0 = 0.3

(iii) Nearest Neighbour:

- Here we are given the error rate of fold i, pi for i from 1 through 10.
- With m and s as average and standard deviation,
    - Hypothetical mean: 1.000000
    - Actual mean: 0.760650
    - Difference between Hypothetical mean and actual mean: m = -0.23935
    - Standard Deviation: s = 0.1248369
    - 95% confidence interval of this difference:
        - From -0.328653 to -0.150047

Now, for p0 = 0.1, 0.2, 0.3 we will calculate the t-value.

- For p0 = 0.1
    - t = ($\sqrt{k}$ (m - p0))/ s
    - So,
        - t = 3.529912

    - Now, by taking value of alpha = 0.01,
        - From t-table, the value of t = 2.821
        - Now, t here has less value than t we calculated.
        - So, error of the classifier is not less than p0 = 0.1
    - Now, by taking value of alpha = 0.025,
        - From t-table, the value of t = 2.262
        - Now, t here has less value than t we calculated.
        - So, error of the classifier is not less than p0 = 0.1

- For p0 = 0.2
    - t = ($\sqrt{k}$ (m - p0))/ s
    - So,
        - t = 0.9967856

    - Now, by taking value of alpha = 0.01,
        - From t-table, the value of t = 2.821
        - Now, t here has more value than t we calculated.
        - So, error of the classifier is less than p0 = 0.2
    - Now, by taking value of alpha = 0.025,
        - From t-table, the value of t = 2.262
        - Now, t here has more value than t we calculated.
        - So, error of the classifier is less than p0 = 0.2

- For p0 = 0.3
    - t = ($\sqrt{k}$ (m - p0))/ s
    - So,
        - t = -1.536341

    - Now, by taking value of alpha = 0.01,
        - From t-table, the value of t = 2.821
        - Now, t here has more value than t we calculated.
        - So, error of the classifier is less than p0 = 0.3
    - Now, by taking value of alpha = 0.025,
        - From t-table, the value of t = 2.262
        - Now, t here has more value than t we calculated.
        - So, error of the classifier is less than p0 = 0.3