Lexi LaMonica
DSC 324

1) **(20 points): The Excel spreadsheet heart.csv contains one sheet named heart. These are data from a sample of 1,988 and consists of four databases from Cleveland, Hungary, Switzerland, and Long Beach with 14 variables for each patient. These are:**

1) Age-age of patient
2) Anaemia-decrease of red blood cells (yes/no)
3) creatinine_phosphokinase-level of CPK enzyme (mcg/L)
4) ejection_fraction-%of blood leaving the heart at each contraction
5) high_blood_pressure – hypertension (Y/N)
6) platelets-in blood (kiloplatelets/mL)
7) serum_creatinine-in blood (mg/dL)
8) serum_sodium-in blood (mEq/L)
9) sex-(woman/man)
10) smoking-(Smoker/Non-smoker)
11) time-follow-up period (days)
12) death_event -death from heart failure (Y/N)

Develop a **Linear Discriminant Analysis** model to classify the death event from the other variables.

Formula: (original data)

age * 2.497 + anaemia * -1.198 + creatinine_phosphokinase * 1.484 + diabetes * 8.345 + ejection_fraction * -4.258 + high_blood_pressure *

-6.1810 + platelets * -3.624 + serum_creatinine * 3.6918 + serum_sodium * -3.291 + sex * -2.757 + smoking * 2.482 + time * -1.183

a) **What is the performance of the classifier using cross-validation?**

	LD1
age	2.409593e-02
anaemia	4.789955e-02
creatinine_phosphokinase	1.734642e-04
diabetes	8.304306e-02
ejection_fraction	-4.587514e-02
high_blood_pressure	5.803614e-02
platelets	-3.927102e-07
serum_creatinine	3.697875e-01
serum_sodium	-6.059682e-02
sex	-2.564251e-01
smoking	1.005217e-01
time	-1.138519e-02

	0	1
0	123	23
1	13	48

From the above figures, the contingency table shows asymmetric binary variables.

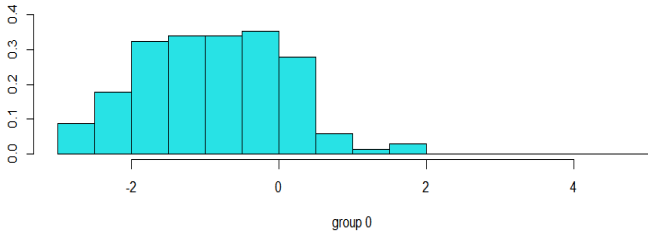
Formula:

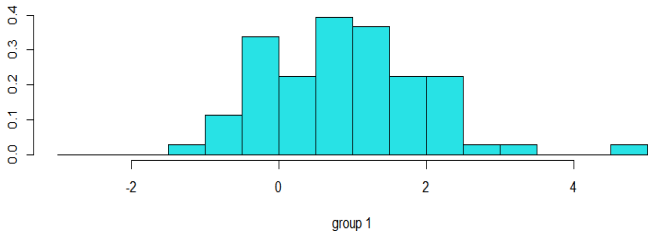
age * 2.4096 + anaemia * 4.78995 + creatinine_phosphokinase * 1.7345 + diabetes * 8.3043+ ejection_fraction * -4.587 + high_blood_pressure * 5.8036 + platelets * -3.9271+ serum_creatinine * 3.698 + serum_sodium * - -6.0597 + sex * -2.5642 + smoking * 1.005 + time * -1.139

The question that should be asked is there a significant relationship between death and the other variables? The contingency table shows that there was 123 no and 48 yes. Down the diagonal it's not accurate since there is room for misclassifications.

From the coefficients, it shows that performance when using cross-validation was that smoking and creatinine phosphokinase had most positive significant relationships along death. While diabetes was less significant. Serum sodium had the least negative relationship on death. Time had the most negatively significant relationship.

b) What is the performance of the classifier using training and testing?






```
               Accuracy : 0.8454
                 95% CI : (0.7888, 0.8918)
    No Information Rate : 0.715
    P-Value [Acc > NIR] : 8.03e-06

                  Kappa : 0.6426

 Mcnemar's Test P-Value : 0.05183

            Sensitivity : 0.8514
            Specificity : 0.8305
         Pos Pred Value : 0.9265
         Neg Pred Value : 0.6901
             Prevalence : 0.7150
         Detection Rate : 0.6087
   Detection Prevalence : 0.6570
      Balanced Accuracy : 0.8409

       'Positive' Class : 0
```

From the above figures, the spread shows that misclassification is happening since there is overlay in the numbers. The accuracy of the testing and training set is 0.845. While the mean was the same (0.845). The confidence interval was 95% with it being 0.7888, 0.8918. The confidence interval is for the accuracy of the data.

The p-value is 8.03 which means a p-value that is > 0.05 is not statistically significant and indicates strong evidence for the null hypothesis. The kappa shows to be 0.6426 which means the rows are less related to each other in terms of its accuracy between the predicted and the actual.

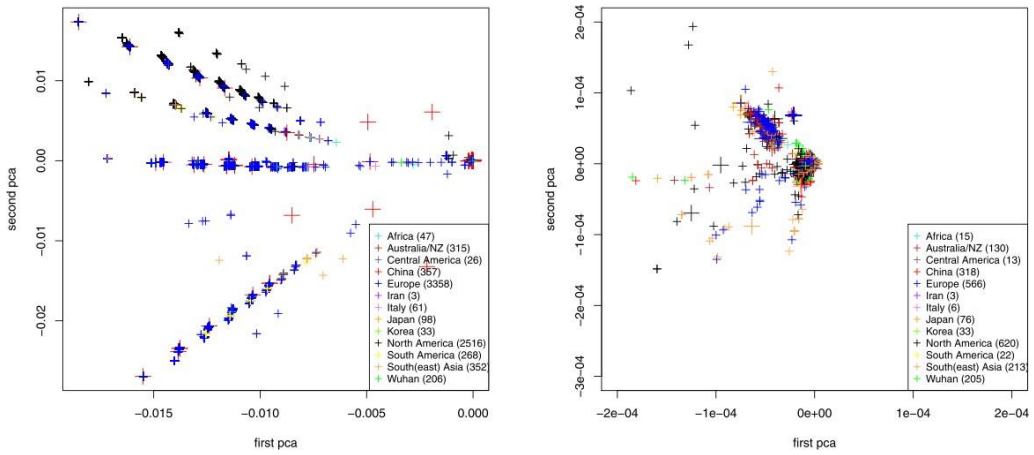
c) Would certain misclassification errors be worse than others? If so, how would you suggest measuring this?

Misclassification can occur when a percentage of classifications that were incorrect. Certain misclassification can be worse than others. An example of how we might go

about measuring this is to look for overlap in the data. From the above training and testing plot, there is overlay around -1.5 through 2

2) **(10 points-Cluster Analysis): Using Google Scholar, locate a journal article, which uses cluster analysis in your field of interest. Write a summary of the journal article and how it utilizes the cluster analysis in two to three paragraphs. Cite the paper in APA format.**

The outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been a developmental stage of advancement for the world of science and human beings. The journal article considers a model-free cluster analysis that compares the viruses at the genome stage They used samples of data collected from patients.



Hahn, G., Lee, S., Weiss, S. T., & Lange, C. (2020).

The authors utilize the cluster analysis within four figures. The figure above relates to principal component analysis. They used the Jaccard index to find similarities in 7640 SARS-CoV-2 genomes by region/country. First PCA is compared with the second, located on the top left, presents the entire data set. While on the right, first PCA is compared with the second again but this time the analysis is a zoomed-in region (Hahn, G., et al).

In sum, the analysis utilized a phylogenic analysis for some of their methods. The analysis comprised of PCA, a computed

tree for sequence alignment, radial representation, and a plot based of the x and y axis.

Hahn, G., Lee, S., Weiss, S. T., & Lange, C. (2020). Unsupervised cluster analysis of SARS-CoV-2 genomes reflects its geographic progression and identifies distinct genetic subgroups of SARS-CoV-2 virus. https://doi.org/10.1101/2020.05.05.079061.