Strathmore
UNIVERSITY

INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN STATISTICAL SCIENCES
END OF SEMESTER EXAMINATION
STA 8303: PREDICTIVE MODELING AND DATA MINING

DATE: Friday 23rd August 2019 Time: 2 Hours

Instructions

1. This examination consists of **FOUR** questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

Question 1. [20 marks]

a) In statistical learning, distinguish between supervised and unsupervised learning. Give appropriate examples of methods that fall into each of these categories.

[**5** Marks]

b) Suppose that we have a training set consisting of a set of points $x_1, \dots, x_n$ and real values y_i associated with each point x_i. We assume that there is a function with noise $y = f(x) + \varepsilon$, where the noise, ε, has zero mean and variance σ^2.

For a function $\hat{f}(x)$, that approximates the true function $f(x)$ as well as possible, by means of some learning algorithm, show that we can decompose its expected error on an unseen sample as follows:

$$E\left[\left(y - \hat{f}(x)\right)^2\right] = Bias\left[\hat{f}(x)\right]^2 + Var\left[\hat{f}(x)\right] + \sigma^2,$$

where $Bias\left[\hat{f}(x)\right] = E\left[\hat{f}(x) - f(x)\right]$ and $Var\left[\hat{f}(x)\right] = E\left[\hat{f}(x)^2\right] - E\left[\hat{f}(x)\right]^2$.

[**5** Marks]

c) Which of the following suggests that the model is overfitting the data? [**2** marks]

A. High accuracy on training data and low accuracy on testing data.

B. High accuracy on training data and high accuracy on testing data.

C. Low accuracy on training data and low accuracy on testing data.

D. Low accuracy on training data and high accuracy on testing data.

E. None of the above.

d) Which of the following tasks would require the use of data mining? [**2** marks]

A. Sorting a customer database by age.

B. Determining which products in a store are likely to be purchased together.

C. Predicting the outcome of rolling two fair dice.

D. Computing the number of products sold over a given time period.

E. All of the above

e) Suppose you have collected data on your customers and you wish to determine the demographics they fall into. Which technique is best suited for this task? [**2** marks]

A. Decision Tree.

B. Neural Network.

C. Linear Regression.

D. Logistic Regression.

E. Clustering

f) How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary least squares regression? (Select one.) [**2** marks]

A. Ridge has larger bias, larger variance

B. Ridge has larger bias, smaller variance

C. Ridge has smaller bias, larger variance

D. Ridge has smaller bias, smaller variance

g) The following list of statements is relevant to the bootstrap methodology Choose from the list which set of statements are true [**2** marks]

I. In practice, it is usual to generate bootstrap samples from the original population

II. When sampling observations from a data set in order to generate bootstrap samples the sampling is done without replacement so that the same observation will not occur twice in a bootstrapped sample

III. If we randomly select n observations for a bootstrap sample from a dataset with N observations it is necessary that n = N

A. I
B. II
C. III
D. None

Question 2. [**20** marks]

a) Describe what is meant by "binary classification". [**2** marks]

b) What is the formula for the accuracy metric? TP = true positive, TN = true negative, FP = false positive, and FN = false negative. [**3** marks]

A. (TP + TN) / (TP + TN + FP + FN).

B. (FP + FN) / (TP + TN + FP + FN).

C. TP / (TP + FP).

D. TP / (TP + FN).

E. TN / (TN + FP)

c) Describe why it is not preferred to fit a binary response variable with a linear regression. [**2** marks]

d) Describe the "odds ratio" and its relationship to logistic regression. [**2** marks]

e) The figure below shows the results of a logistic regression model that predicts the odds of adequate antenatal care (AANC) among women aged 15-49 yearswho had delivered during the 5 year prior to the 2014 Kenya Demographic and Health survey.

```
Call:
glm(formula = Aanc ~ V012 + educ + blast5yrs + decisionm, family = binomial(link = logit),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4668  -0.8133  -0.7749   1.4445   1.7164

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.679114   0.245393   2.767  0.00565 **
V012                  -0.005775   0.006234  -0.926  0.35424
educNo education      -1.448681   0.192388  -7.530 5.07e-14 ***
educPrimary           -1.525955   0.158488  -9.628  < 2e-16 ***
educSecondary         -1.192648   0.171704  -6.946 3.76e-12 ***
blast5yrs2            -0.123283   0.091204  -1.352  0.17646
blast5yrs3+           -0.123790   0.170201  -0.727  0.46703
decisionmlow           0.095119   0.104305   0.912  0.36180
decisionmmid           0.004453   0.107568   0.041  0.96698
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V012	Mother's (age in years)	
Educ	Mother's level of education	Tertiary is the base level
blast5yrs	Births in the last 5 years	1 is the base level
Decisionm	Mother's decision making ability	High is the base level

Compare the odds of adequate ANC for women with no formal education compared to those with tertiary education holding all other variables constant. [**3** marks]

f) Figure 1 presents ROC curves for three models that have been fit to the ADC data. Order these three models from the best fit to the worst fit. [**3** marks]

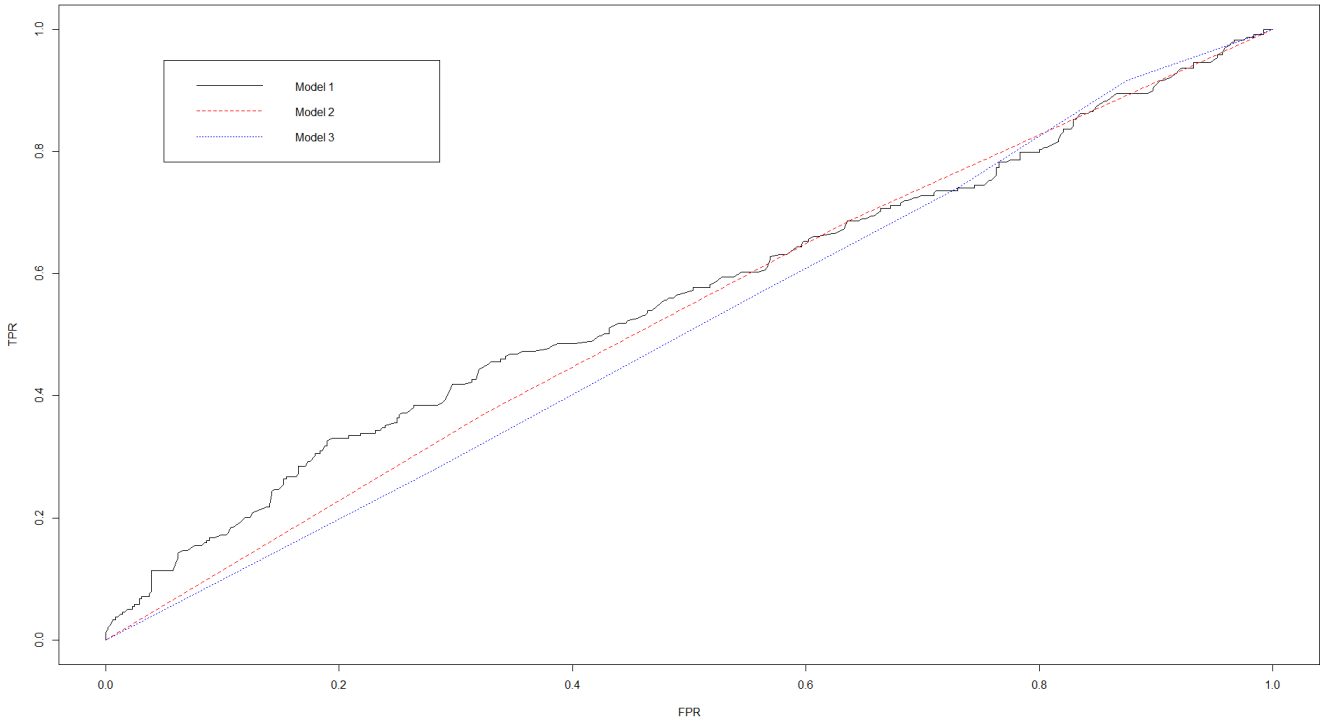

Figure 1 ROC curves for the three candidate models

Question 3 (20 Marks**)**

a) Sequential variable selection and Ridge regression analysis are 2 approaches used in combating *Multicollinearity* in data. Distinguish between them, explaining advantages of each technique.

[**6** Marks]

b) For the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim MVN(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, derive an expression for the mean and variance of ridge regression estimator $\widehat{\boldsymbol{\beta}}_{RIDGE} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}X'\boldsymbol{y}$.

[**6** Marks]

i) Prove that $\widehat{\boldsymbol{\beta}}_{RIDGE}$ is a asymptotically unbiased estimator of $\boldsymbol{\beta}$**;**

[**2** Marks]

ii) Prove that $\text{Var}[\widehat{\boldsymbol{\beta}}_{RIDGE}] \leq \text{Var}[\widehat{\boldsymbol{\beta}}_{MLE}]$, where $\widehat{\boldsymbol{\beta}}_{MLE}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$**.**

[**6** Marks]

Question 4 (20 Marks**)**

a) Describe the purpose and objective of *Principal Components Analysis* (PCA) and give any 3 examples of areas in which its finds application.

(**5** Marks)

b) Cluster analysis is a commonly employed unsupervised learning procedure. Distingush between Agglomerative clustering and Divisive clustering algorithms.

(**3** Marks)

c) Explain how the Partitioning around Medoids (PAM) approach works.

(**4** Marks)

d) A random sample of 74 cars was selected. For each car the following variables were measured: **headroom** [Headroom (in.)], **trunk** [Trunk space (cu. ft.)], **weight** [Weight (lbs.)], **length** [Length (in.)], **turn** [Turn Circle (ft.)], and **displacement** [Displacement (cu. in.)].
Based on the results of the PCA analysis given in the Appendix:

i. Explain how many principal components you would select and why

(**2** Marks)

ii. Explain what each of the selected component(s) describes;

(**2** Marks)

i. Comment on the results of the 10 cars considered on the basis each of the components selected;

(**2** Marks)

ii. Comment on the correlation circle and it's significance.

(**2** Marks)

APPENDIX

Table 1 Correlation Matrix

	headroom	trunk	weight	length	turn	displacement
headroom	1.0000000	0.6620111	0.4834558	0.5162955	0.4244646	0.4744915
trunk	0.6620111	1.0000000	0.6722057	0.7265956	0.6010595	0.6086350
weight	0.4834558	0.6722057	1.0000000	0.9460086	0.8574429	0.8948958
length	0.5162955	0.7265956	0.9460086	1.0000000	0.8642612	0.8351400
turn	0.4244646	0.6010595	0.8574429	0.8642612	1.0000000	0.7767647
displacement	0.4744915	0.6086350	0.8948958	0.8351400	0.7767647	1.0000000

Table 2 Eigen-values

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	4.50151930	75.0253217	75.02532
Dim.2	0.80149921	13.3583202	88.38364
Dim.3	0.30817531	5.1362552	93.51990
Dim.4	0.22411069	3.7351781	97.25508
Dim.5	0.12361234	2.0602056	99.31528
Dim.6	0.04108315	0.6847191	100.00000

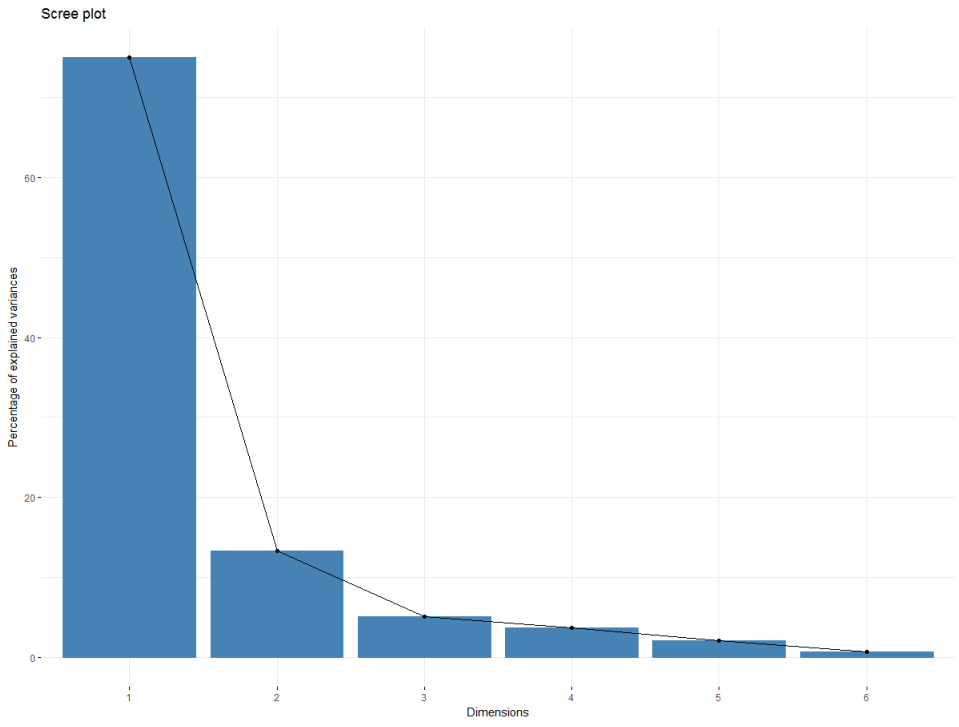

Figure 2 Scree-plot

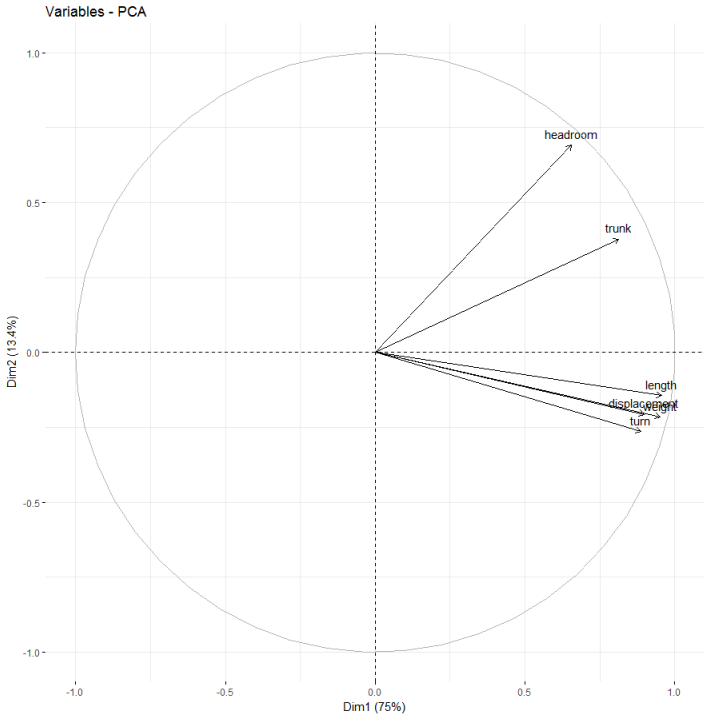

Figure 3 Correlation circle

Table 3 Summary of results

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Variance	4.502	0.801	0.308	0.224	0.124	0.041
% of var.	75.025	13.358	5.136	3.735	2.060	0.685
Cumulative % of var.	75.025	88.384	93.520	97.255	99.315	100.000

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
AMC Concord	1.222	-0.842	0.213	0.475	-0.518	0.452	0.180	-0.085	0.032	0.005
AMC Pacer	1.229	-0.043	0.001	0.001	-0.440	0.326	0.128	0.829	3.014	0.455
AMC Spirit	1.748	-1.581	0.750	0.818	0.600	0.607	0.118	0.083	0.030	0.002
Buick Century	1.930	1.082	0.351	0.314	1.458	3.586	0.571	0.518	1.176	0.072
Buick Electra	3.354	3.272	3.214	0.952	0.359	0.217	0.011	0.001	0.000	0.000
Buick LeSabre	2.761	2.491	1.862	0.813	0.915	1.412	0.110	-0.630	1.741	0.052
Buick Opel	2.351	-1.206	0.436	0.263	0.121	0.025	0.003	1.064	4.961	0.205
Buick Regal	1.542	0.453	0.062	0.086	-1.014	1.735	0.433	-1.059	4.922	0.472
Buick Riviera	1.912	1.844	1.021	0.930	0.071	0.009	0.001	-0.147	0.095	0.006
Buick Skylark	1.167	0.966	0.280	0.685	-0.059	0.006	0.003	0.566	1.402	0.235

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
headroom	0.655	9.536	0.429	0.692	59.741	0.479	0.293	27.901	0.086
trunk	0.813	14.688	0.661	0.379	17.905	0.144	-0.428	59.333	0.183
weight	0.951	20.108	0.905	-0.216	5.807	0.047	0.037	0.435	0.001
length	0.955	20.280	0.913	-0.144	2.577	0.021	-0.060	1.172	0.004
turn	0.887	17.478	0.787	-0.264	8.687	0.070	0.014	0.060	0.000
displacement	0.898	17.911	0.806	-0.206	5.283	0.042	0.185	11.099	0.034