Strathmore
UNIVERSITY

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

DATE: April , 2021 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) Explain how the each of the following resampling techniques is implemented in predictive modeling:
   i) Validation set approach.
   ii) Leave-One-Out cross-validation.
   iii) Bootstrapping.

(**8** marks)

c) 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 $\widehat{\boldsymbol{\beta}}_{RIDGE} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1} X'\boldsymbol{y}$.
Give an expression for the mean square error of this estimator and explain its significance in terms of bias-variance trade-off.

(**7** Marks)

Question 2 (20 Marks)

a) Explain the significance of the concept of *Bias-variance trade-off* in a statistical learning algorithm.

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

[**6** Marks]

c) Sequential variable selection techniques, principal components regression, and Ridge regression analysis are 3 approaches used in combating *Multicollinearity* in data. Distinguish between them, explaining advantages of each technique.

(**9** Marks)

Question 3 (20 Marks)

a) Logistic regression, Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are three classification techniques that are widely used by predictive modelers.

Explain the main similarities and differences that exist between LDA and QDA. Provide a mathematical description of each approach.

[**5** Marks]

b) Consider a data set of 144 observations of household cats. The data contains the cats' gender, body weight and height. The researcher would like to model and accurately predict the gender of a cat based on previously observed values.
To verify and test our model's performance, they split the data into training (60%) and test sets (40%). Two models were entertained:

- Model 1: A logistic regression model with body weight as predictors
- Model 2: A logistic regression model with body weight and height as predictors

The results of these two models are presented in Table 1 and Table 2. The confusion matrices for these two models are also presented in

Table 1 The results of fitting a logistic regression model with body weight as predictor to the training data (Model 1)

```
Call:
glm(formula = Sex.f ~ Bwt, family = binomial, data = training)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.7939     1.8571  -3.658 0.000254 ***
Bwt           2.8989     0.7346   3.946 7.94e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 111.559  on 87  degrees of freedom
Residual deviance:  89.159  on 86  degrees of freedom
AIC: 93.159
```

Table 2 The results of fitting a logistic regression model with body weight and height as predictors to the training data (Model 2)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.8330     1.8334  -3.727 0.000194 ***
Bwt           3.5369     1.1111   3.183 0.001457 **
Hwt          -0.1602     0.2021  -0.792 0.428095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 111.559  on 87  degrees of freedom
Residual deviance:  88.527  on 85  degrees of freedom
AIC: 94.527
```

Table 3 Confusion matrix for Model 1 and 2

	Predicted status				Predicted status	
Actual status	Female	Male		Actual status	Female	Male
Female	12	10		Female	9	15
Male	13	22		Male	13	20

i) From the confusion matrices above, compare the 3 models. Compare your results based on model accuracy.

[**4** Marks]

ii) For the best fitting model, compute the following measures: sensitivity, specificity and the false positive rate.

[**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 4 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 5 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

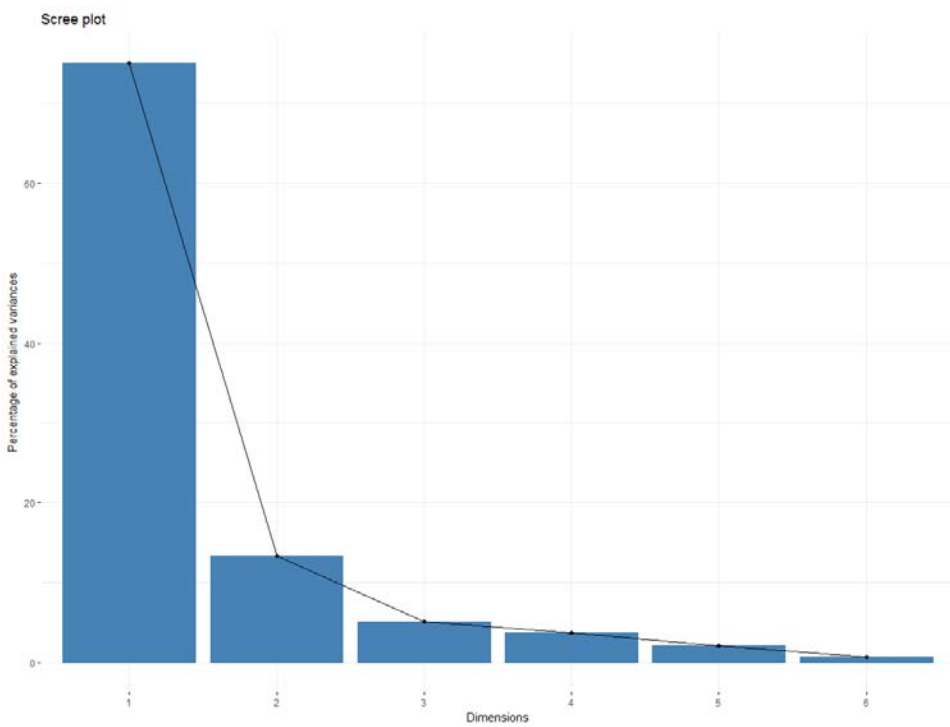

Figure 1 Scree-plot

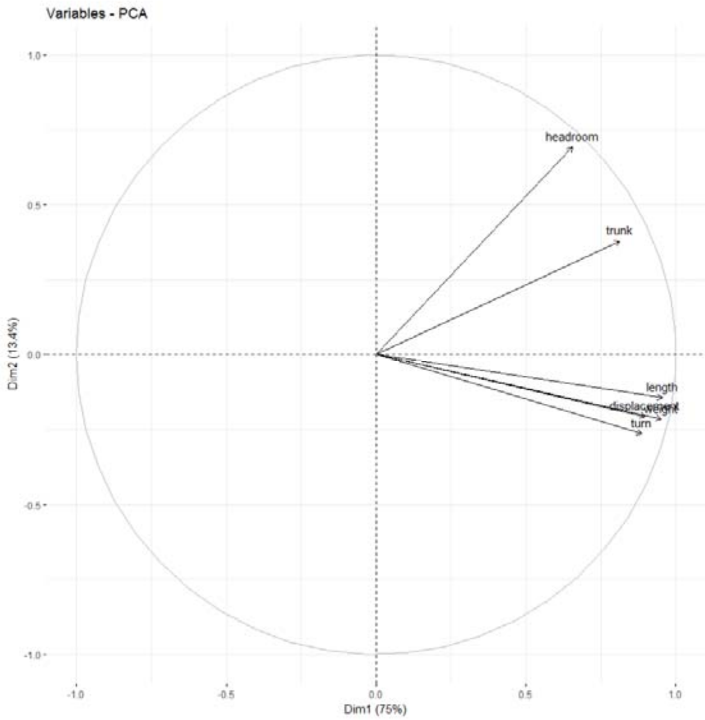

Figure 2 Correlation circle

Table 6 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