Strathmore
UNIVERSITY

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

DATE: Wednesday 15th August, 2018 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 $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 \boldsymbol{I})$, 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 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) 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.

(**10** Marks)

Question 3 (20 Marks**)**

a) Distinguish between the following concepts in classification algoritms: *Sensitivity and specificity*.

(**4** Marks)

b) A Receiver Operating characteristic function is a useful tool in predictive modeling. Explain how an ROC functions can be developed for a given statistical learning procedure and describe its application in predictive modeling.

(**6** Marks)

c) A team of researcher at **CDC-Kenya** would like to develop *a predictive model* for TB in Kenya using two explanatory variables: HIV status; and smoking status.

Table 1 Confusion matrices for the three models considered

Fit 1: HIV status as the only predictor

	Predicted status	
Actual status	Negative	Positive
Negative	4000	1000
Positive	2000	3000

Fit 2: smoking status as the only predictor

	Predicted status	
Actual status	Negative	Positive
Negative	2000	3000
Positive	3200	1800

Fit 3: HIV and smoking predictors

	Predicted status	
Actual status	Negative	Positive
Negative	3300	1700
Positive	1000	4000

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.

(**6** Marks)

b) Describe any other unsupervised learning procedure, apart from, you are aware of.

(**6** Marks)

c) 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 2 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 3 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

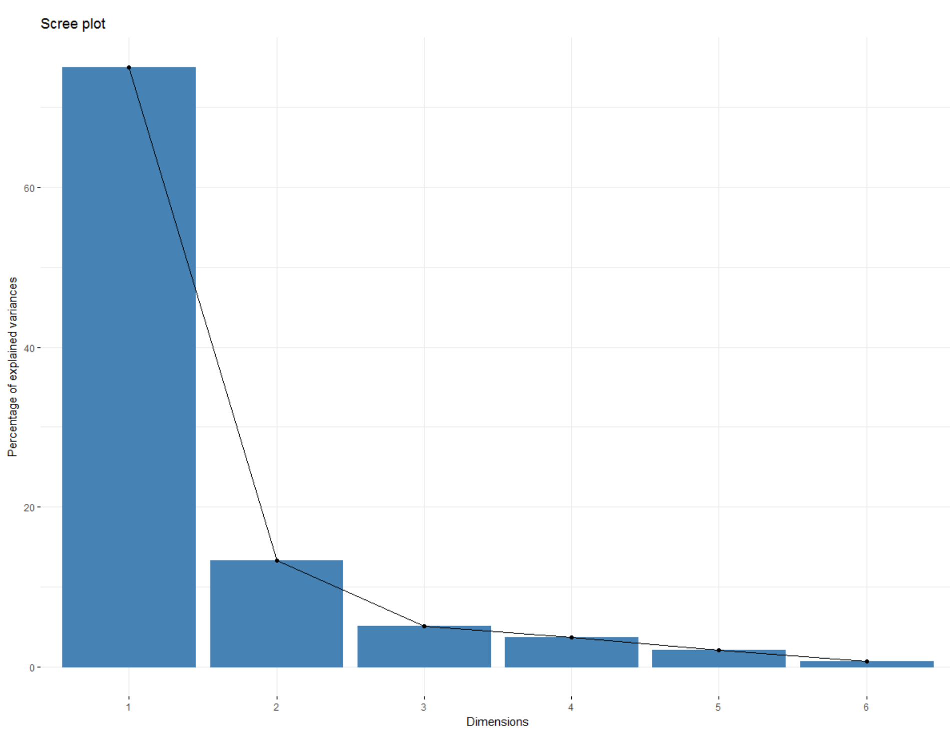

Figure 1 Scree-plot

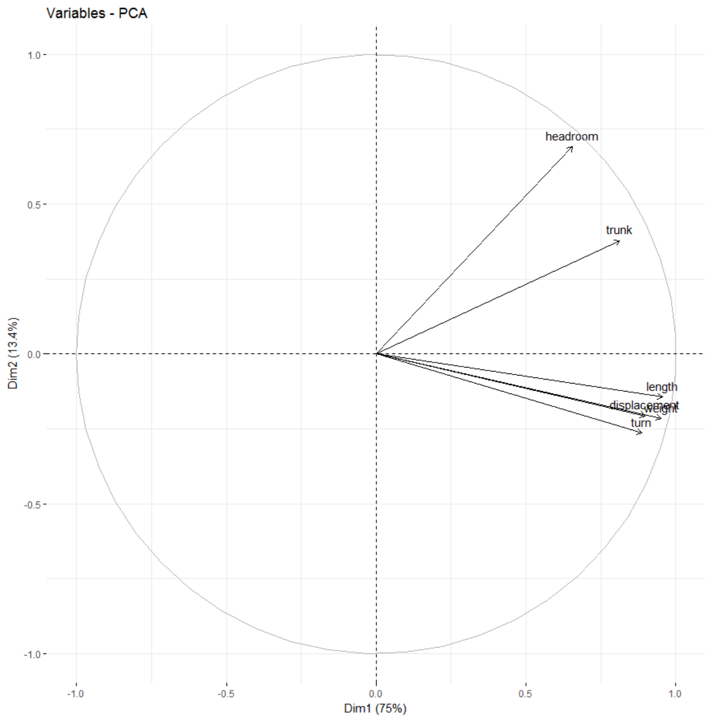

Figure 2 Correlation circle

Table 4 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