OLLSCOIL NA hÉIREANN, CORCAIGH
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

Examination Session and Year	Winter 2023
Module Code	ST4060 ST6040
Module Name	Statistical Methods for Machine Learning I Machine Learning and Statistical Analytics I
Paper Number	Paper Number: 1
External Examiner	Mr. Andrew Maclaren
Head of School	Dr. Kevin Hayes
Internal Examiner(s)	Dr. Eric Wolsztynski
Instructions to Candidates	• Please answer all questions. • Provide all your answers in the Word document provided. • Paste your R code into the Word document at the end of each question. • Submit a pdf version of your final Word document for Canvas submission. Note: if you do not manage to answer a question item, provide the R code you would have used, or a comment on the answer you would expect for that question, as relevant.
Duration of Paper	3 hours

List of required R libraries:

```
MASS
splines
```

List of (possibly) useful R functions:

```
abline()
apply()
approx()
as.numeric()
boxplot()
bs()
cbind()
coef()
colnames()
cut()
density()
fitted()
kmeans()
lines()
lm()
matrix()
mean()
median()
na.omit()
nrow()
numeric()
order()
par()
plot()
points()
prcomp()
predict()
quantile()
rnorm()
sample()
sd()
seq()
set.seed()
smooth.spline()
sqrt()
sum()
summary()
table()
which()
```

Question 1 [25 marks]

Consider the sample mean $\bar{X}$ of a sample of N independent and identically distributed realizations of a random variable $X \sim \mathcal{N}(\theta^\star, \sigma^2)$,defined by

$$\bar{X} = \frac{X_1 + \cdots + X_N}{N}$$

(a) Recall that the 95% confidence interval for $\bar{X}$ is obtained using the sample standard deviation s by

$$\mathcal{C} = \left[\bar{X} - 1.96\frac{s}{\sqrt{N}}, \bar{X} + 1.96\frac{s}{\sqrt{N}}\right]$$

Using $\theta^\star = 3$, $\sigma = 1.5$, and $M = 1,000$ Monte Carlo resamples, each of size $N = 30$, compute a Monte Carlo estimate of the proportion

$$p = \mathbf{I}(\theta^\star \in \mathcal{C})$$

i.e. the number of times that the confidence interval includes the true value $\theta^\star$. Set the random seed to 4060 (`set.seed(4060)`) before running your computation. Quote your Monte Carlo estimate of p. **[10]**

(b) Replicate the computation done in (a), but this time using

$$\mathcal{C} = \left[\bar{X} - 1.645\frac{s}{\sqrt{N}}, \bar{X} + 1.645\frac{s}{\sqrt{N}}\right]$$

Set the random seed to 4060 (`set.seed(4060)`) before running your computation. Quote your Monte Carlo estimate of p for this new confidence interval. **[5]**

(c) Comment on the values you obtained in (a) and (b), and indicate what these estimates correspond to. **[5]**

(d) Describe, in one or two sentences, what you would expect to happen if the sample size of each Monte Carlo sample was increased to $N = 100$ in the experiment described in (a), and why. **[5]**

Question 2 [25 marks]

Load the `Animals` dataset from library `MASS` into your `R` session as follows:

```
library(MASS)
x = Animals
```

This dataset contains average brain and body weights recorded for 28 species of land animals.

Implement a bootstrap analysis of the dataset using $B = 1,000$ bootstrap resamples, and setting the random seed to 4060 (`set.seed(4060)`) before running the analysis.

(a) Quote the bootstrap estimate of the mean brain weight of all land animals (your answer should be one value, calculated from the sample of all body weights for all species). Also quote the bootstrap estimate of the mean body weight of all land animals. **[5]**

(b) Quote the bootstrap estimate of the mean ratio of brain-to-body weight of all land animals. For any given animal species, this ratio is calculated as (brain weight)/(body weight). **[5]**

(c) Compute and quote the bootstrap estimate of the bias of the sample mean body weight estimate of these 28 species of land animals. **[5]**

(d) Compute and quote a bootstrap 95% confidence interval for the mean body weight of these 28 species of land animals. (You may use the *naive* (a.k.a. *quantile*) bootstrap confidence interval for this question.) **[5]**

(e) Except for the sample size, what explains the large width of the bootstrap confidence interval you obtained in (d) for mean body weight? Give your answer in two sentences maximum. Provide appropriate R output to support your answer. **[5]**

Question 3 [25 marks]

Load `R` libraries `splines` and `MASS` along with the `Boston` dataset as follows:

```
library(splines) # contains function bs()
library(MASS)
x = Boston$nox
y = Boston$medv
```

Set the random seed to 4060 (`set.seed(4060)`) before running your analysis.

(a) Fit a B-spline to the data, using knots placed at quantiles `c(0.15,0.40,0.60,0.70,0.85)` of `x`. Quote the B-spline coefficient estimates. **[5]**

(b) Generate predictions for new `x` values `newx = c(0.4,0.5,0.6)` from the B-spline obtained in (a). Quote the predicted values for `y`. **[5]**

(c) Fit a P-spline (i.e. smoothing spline) to the data, setting ordinary leave-one-out in the function arguments for computation of the smoothing parameter. Use `set.seed(4060)` before you run your code for this question.

(i) Quote the P-spline penalized criterion (RSS).

(ii) Provide a plot showing the fitted B-spline in (a) (in red) and the fitted P-spline (in blue) obtained in (c), over the (`x`,`y`) scatterplot (in black).

[5]

(d) Generate predictions for new `x` values `newx = c(0.4,0.5,0.6)` from the P-spline obtained in (c). Quote the predicted values for `y`. Compare those to the values obtained in (b) and explain any difference you may find between these two sets of predictions. **[5]**

(e) Implement 5-fold cross-validation of the P-spline. Quote the estimated prediction RMSE obtained from this analysis. Use `set.seed(4060)` before you run your code for this question.
Note: if you were not able to fit a P-spline, implement 5-fold cross-validation of a linear regression model (with intercept) instead, and quote the corresponding prediction error estimate. **[5]**

Question 4 [25 marks]

Load the `Pima.tr` dataset from the `MASS` package into your `R` session as follows:

```
library(MASS)
x = Pima.tr
x$type = NULL
y = Pima.tr$type
```

(a) Consider an analysis of this data that aims to predict `y` based on the measurements in `x`. Is this a regression or a classification problem? Justify your answer. **[5]**

(b) Perform k-means clustering on `x`, so as to cluster the data into k=2 clusters. Provide:

(i) The confusion matrix between the cluster labels and `y`.

(ii) A scatterplot of `x[,1:2]`, using `pch=20` to draw points as filled circles in the plot, and colour-coding (i.e. painting) the points with respect to their cluster (using black and red points).

[5]

(c) Briefly comment on the spatial distribution of the points, in terms of their cluster membership, in the scatterplot obtained in (b). In particular, explain any particular pattern you may see in this spatial distribution. **[5]**

(d) Perform *scaled* PCA on the feature matrix `x`. Indicate the number of principal components that together capture 90% of the information in `x`. Justify your answer. **[5]**

(e) Perform *unscaled* PCA on the feature matrix `x`. Indicate which features mainly influence the first 2 principal components. Justify your answer. **[5]**