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

**COLÁISTE NA hOLLSCOILE, CORCAIGH**
UNIVERSITY COLLEGE, CORK

# ST4061 - Statistical Methods for Machine Learning II
# ST6041 - Machine Learning and Statistical Analytics II
# EXAM ANSWER DOCUMENT

**NAME AND SURNAME:**
**STUDENT NUMBER:**
**CLASS (FMAS4, MSH4, MScDSA, RAS4, MScMMSLS, MScAS, etc.):**

**Instructions**

- Please provide your answers **in this answer sheet**.
- **Only upload the pdf version of your answer document.**
- **Provide all answers hereafter. Paste relevant figures and R output within this document where appropriate.**
- **Please email any questions directly to eric.w@ucc.ie**

## Answers to Question 1

**Table 1 - Match the models proposed in Column A to the models depicted in Figure 1 of Question 1. Indicate N/A when a model is not represented in Figure 1. Provide a brief explanation for each of your answers in column C.**

| (A) Proposed model | (B) Model of Figure 1 | (C) Explanation for your answer in column B |
|---|---|---|
| A random forest | **Model 3** | **Clear sequence of horizontal and vertical decisions in the branching pattern, typically associated with a tree-based decision process** |
| A logistic regression model | **Model 4** | **Linear boundary (+ associated with poor classification for this nonlinear problem)** |
| A Quadratic Discriminant Analysis with 3 Gaussian components | **N/A** | ***3 components would be used for a 3D dataset, whereas this one is 2D*** |
| A Support Vector Machine using a radial basis function | **Model 2** | **Clear radial kernel pattern** |
| A lasso classifier with an extremely large shrinkage parameter | **N/A** | ***This model would remove one of the two predictors; this would yield the boundary to be either a horizontal or vertical line*** |
| A kNN classifier (with k=5) | **Model 1** | **A non-regular decision region contour resulting from low value of k; the only plausible model for the top-left scenario in Fig 1** |

**Answer to Question 1(b):**

**Euclidean distances between these observations and test point (0,0,0), respectively from obs #1 to obs #5:**

**2.000 3.000 3.162 2.236 1.414**

**(i) If k=1, then obs #5 is closest, so predicted label = Y5 ("LOW")**

**(ii) If k=3, then obs #5, #1 and #4 are closest, so predicted label is determined by a majority between (Y1="high", Y4="high", Y5="low"), i.e. predicted label is "HIGH"**

## Answers to Question 2

**(a)** Regression

**(b)** 12 (ncol(dat)-1)

**(c)** NN analysis:

i.

```
 summary(dat.s$Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.3834  0.4884  0.4887  0.5944  1.0000 
> summary(dat.s$BudgOp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.2197  0.4848  0.4814  0.7071  1.0000 
> summary(dat.s$Training)
   Bad   Good Medium 
    96     85    219 
```

ii.

```
> mean(nn3$residuals^2)
[1] 0.01523216
> mean(nn8$residuals^2)
[1] 0.008778118
```

iii.

```
> mean((p3-ytest)^2)
[1] 0.02901814
> mean((p8-ytest)^2)
[1] 0.04108399
```

iv. Possible overfitting from the more complex FFNN? Needs to be confirmed by CV.

**(d)**

```
> mean((gbmo$fit-dat.train$Sales)^2)
[1] 10.01079
> mean((gbmp-dat.validation$Sales)^2)
[1] 16.47476
```

**(e)**

```
> mean((glmo$fit-dat.train$Sales)^2)
[1] 11.65191
> mean((glmp-dat.validation$Sales)^2)
[1] 15.79001
```

**(f)**

```
> mean((ridge.fit-dat.train$Sales)^2)
[1] 11.77066
> mean((ridgep-dat.validation$Sales)^2)
[1] 15.50165
```

**(g)** >> compare with NN.... WATCH OUT FOR SCALING!!

```
pov = po*(max(dat$Sales)-min(dat$Sales))+min(dat$Sales)
mean((pov-dat.validation$Sales)^2)
```

comparable errors….

**(h)** Cf rfe output

**(i)** Cf rfe output

**(j)** Open question (RMSEs are actually comparable), marks for merit in argumentation

**Paste your R code for Question 2 here:**

**Answers to Question 3**

(a) Mtry=2. Test accuracy = 0.952

(b) Test accuracy = 0.940

(c) Test accuracy = 0.976

(d) None, as CI's around accuracies are comparable.

(e) Cell.shape and Cell.size are definitely the most important predictors for all 3 models. Then Cl.thickness (3rd most important variable for RF) seems to be a strong predictor across models, although Bare.nuclei is ranked 3rd in terms of ROC-based variable importance for the SVMs.

The top 2 features are clear predictors across models, and the 3rd variable depends on which model is used.

Marking will depend on further treatment of the question.

**Paste your R code for Question 3 here:**