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

COLAISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

ST6026
Basics of Machine Learning

Workshop 1 - Statistical learning and data science

Question 1 (Warm-up: read files in and write files out)

(a) Download file `Blood Pressure.txt` from Canvas onto your system.

- Read this file in to R using `read.table()`.
- Write this dataset out as a `.csv` file using `write.csv()`.
- Inspect this new file on your system. Notice the extra column with row numbers? Fix it by adjusting argument `row.names` accordingly when using `write.csv()`.

(b) Download file `Blood Pressure CSV.csv` from Canvas onto your system.

- Read this file in to R using `read.csv()`.
- Write this dataset out as a `.txt` file using `write.table`. Adjust argument `row.names` appropriately when using this function.
- Inspect this new file on your system. Notice the column names are now within quotes? Fix it by adjusting argument `quote` accordingly when using `write.table()`.
- Now fix column spacing by setting argument `sep` in your call to `write.table()` so that columns are nicely tabulated in the output file.

Question 2 (Basic steps in regression)

Load the following package into R:

```
library(ISLR)  # contains dataset Wage
```

Note: to install new packages you can simply run e.g. `install.packages('ISLR')` in R.

(a) Inspect dataset `Wage` by running

```
names(Wage)
class(Wage)
class(Wage$wage)
class(Wage$age)
class(Wage$education)
```

(b) Create a basic plot of the dataset. Then,

- Improve it with better looking dot styles (`pch=...`) and colour (`col=...`).
- Change the x- and y-axis labels.

(c) Fit a basic GLM to this dataset.

(d) Fit also a non-parametric regression curve to the data, as follows:

```
np.curve = lowess(Wage$age, Wage$wage)
```

Add this smooth, non-parametric curve to the plot using `lines()`, in navy, and with line width 3.

(e) What do the distributions of each of the features and of the response variable look like? Create a suitable plot to inspect the distributions of variables `wage`, `age` and `education` from this dataset.

(f) Do the features have any relationship with the response variable? Create a plot to visualise relationships between variable wage and each of `year`, `age` and `education`.

Question 3 (Data inspection in a classification scenario)

(a) Plot `Sepal.Width` against `Sepal.Length` from the `iris` dataset. Colour-code the dots with respect to variable `Species` (i.e. use e.g. black, red and blue to display the data points in a way to reflect which species they correspond to).

(b) Increase the size of the data points in your scatterplot, using argument `cex` within your call to `plot()`.

(c) Now set data point size so as to reflect their `Petal.Width` value.

Question 4 (Unsupervised learning)

(a) Perform a simple form of data clustering called *k-means clustering* on the `iris` dataset, but removing the first variable from this dataset, and clustering the data into 3 groups, as follows:

```
x = iris[,c(2:4)]
y = iris[,5]
K = 3
ko = kmeans(x, K)
```

Inspect the distribution of clustered points and create a scatterplot of `Sepal.Width` against `Petal.Width`, showing different clusters in different colours.

(b) Repeat the above analysis on the scaled data, using

```
z = apply(x,2,scale)
```

and discuss.

Question 5 (Performance in regression)

Load the following library into R:

```
library(MASS)  # contains the Boston dataset
```

(a) Prepare 2 datasets, one being the full `Boston` dataset, the other by dropping the first 6 variables in the dataset, as follows:

```
# first shuffle the data!
n = nrow(Boston)
set.seed(6026)
dat = Boston[sample(1:n, n, replace=FALSE),]
sdat = dat[,7:14] # subset eliminating the first 6 features
```

(b) Create a training subset by taking the first 400 observations in `dat`, the remaining observations being used as a test set to evaluate predictive performance on unseen data. Fit a GLM to this first training set.

(c) Create a training subset by taking the first 400 observations in `sdat`, the remaining observations being used as a test set to evaluate predictive performance on unseen data. Fit a GLM to this second training set.

(d) Compare the two model fits in terms of their AIC values.

(e) Generate predictions for `medv` in the test set from both model fits. Calculate the corresponding prediction Mean Squared Errors (MSEs) and discuss.

Question 6 (Performance in classification)

Load the following libraries into R:

```
library(tree)  # contains an implementation of classification and regression trees
library(ISLR)  # contains the Default dataset
library(pROC)  # contains roc()
```

(a) Remove variable `student` from the `Default` dataset in order to make a simpler dataset:

```
dat = Default
dat$student = NULL
```

Then fit a classification tree to explain `default` with respect to the other variables in the dataset:

```
classifier = tree(default~., data=dat)
```

(b) Recover classification probabilities for the fitted values (i.e. the likelihood of predicted classes for this dataset).

(c) Perform an ROC analysis of this classifier using `roc()`.

(d) Plot the corresponding ROC curve.

(e) Quote the corresponding AUC value (AUC = Area Under the ROC Curve) and discuss.

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

COLAISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

ST6026
Basics of Machine Learning

Workshop 2A - Machine Learning paradigm

Eric Wolsztynski
eric.w@ucc.ie

Question 1

Imagine you want to build recognition software that can automatically detect and recognize a dog, a cat or a camel in a photograph. Briefly describe the main tasks that must be performed in order to build such a tool. Also briefly describe some potential errors such software may yield.

Solution:

In reverse order:

(a) A classification model that, for a given photograph of an animal, can label this animal as a dog, a cat or a camel. This model will required training on large datasets comprising of photograph of these animals, with associated labels, and hopefully showing each of these animals in many orientations (angles, positions, etc.).

(b) A detection algorithm that is able to find the animal in the picture (there are several approaches, drawing multiple bounding boxes is an obvious one). In fact, detection could be determined by a high probability that the object in a candidate "extracted sub-image" actually contains a dog, a cat or a camel (i.e. this candidate image would be passed into the classification scheme and detection of the animal would be triggered by a strong response from the classifier to this input).

It is worth noting that each of these steps will likely not perform perfectly. There may be:

- false positives (i.e. calling a tea mug a cat!),
- false negatives (failing to see the animal),
- incorrect classification (calling a dog a camel!).

Question 2

An analyst wants to build a predictive model for cancer patient outcome based on a combination of clinical and radiological variables, such as patient age, weight, gender, tumour grade, and other image-derived variables measuring various tumour characteristics.

(a) To do so the analyst considers training a machine learning model on a clinical cohort of 25 patients with stage 4 lung cancer. Comment on the viability of this training approach.

(b) The analyst's colleague suggests instead to apply deep learning on an open-source clinical dataset comprising of thousands of MRI scans gathered across a large number of European health care centres. The deep learning model will be trained for automatic detection of high-risk tumours. Is this a classification or a regression task? Comment on this approach in terms of the final goal of patient risk modelling (what training data should be used, and whether this is a realistic approach for patient outcome prediction).

Solution:

(a)
- Too small a sample size.
- Too specific a cohort, this will induce bias in the model.

(b)
- DL for lesion detection = classification
- Specific task (detection/classification): model can be trained effectively at detection. Easier to train for overall detection (tumour present/absent) than for detection of high-risk tumours (which is a more specific target requiring elaborate assessment including biopsy, etc.).
- Cancer is a very varied and complex disease to characterize, a detection algorithm will likely not be sufficient at characterizing risk.
- The DL model can be trained only for the specific task of detection. This training would require data comprising of scans with both scenarios (disease present vs absent). A model for high-risk lesion detection may be trained on image sets comprising of three labels: no disease, low-risk disease and high-risk disease, and would be effective if there is a strong difference between low- and high-risk tumours on the images (which is not a given!).
- It is possible to train a DL model for high-risk prediction based on imaging data, but this model would require combining the detection task with more elaborate disease characterization, i.e. include data other than radiological images (e.g. including biopsy etc.). We are thus here talking about a model that performs more than the specific task of image-based detection.

Question 3

A national health service executive body is tendering for innovative solutions for integration, synergization and coordinated exploitation of patient data across all medical services at the national level. One company puts forward a proposal consisting in collating and curating patient records

from handwritten medical notes taken during any consultation by the patient with their GP and/or any health consultant. These patient electronic records will contain free-formatted text for further analysis. Briefly describe a generic machine learning approach that could be considered for analysis of such electronic records.

Solution:

- Natural Language Processing (of free-formatted text).
- This process would require OCR (Optical Character Recognition) for data extraction and sentiment analysis mining records for relevant medical information.

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

COLAISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

ST6026
Basics of Machine Learning

Workshop 2B - Model regularisation

Note: fitting regularised linear models in R is somewhat tedious. The following two functions are more convenient–note their use slightly differs from that of `lm()`(version 12 Feb 2021):

```
regularized.model <- function(x, response, alpha=1, ...){
# wrapping glmnet into a more user-friendly function
        # prepare the data
        if(!(response %in% names(x))){
                stop("The response name provided is not included in
                        the dataset")
        }
        fml = as.formula(paste(response, "~."))
        xm = model.matrix(fml, data=x)[,-1]
        y = x[[response]]

        if(class(y)=="factor"){
                if(length(levels(y))==2){
                        fam = "binomial"
                } else {
                        fam = "multinomial"
                }
        } else {
                fam = "gaussian"
        }

        lam = cv.glmnet(xm, y, alpha=alpha, family=fam, ...)
        mod = glmnet(xm, y, alpha=alpha, lambda=lam$lambda.min,
                family=fam, ...)
```

```
        mod.fit = predict(mod, newx=xm)
        return(list(model=mod, fitted.values=mod.fit, y=response))
}

predict.regularized.model <- function(fit, test.data){
        fml = as.formula(paste(fit$y, "~."))
        newx = model.matrix(fml, data=test.data)[,-1]
        return(predict(fit$model, newx))
}
```

The above functions are available in file `ST6026_package.R`. In order to avail of these functions, you can simply run this code in R prior to performing an analysis.

Question 1

The `Blood Pressure.txt` dataset contains measurements of systolic blood pressure, age, waist circumference, cholesterol and BMI index, for 75 subjects (variable `PatientID` is dropped from the dataset). Four linear regression models have been fit to the dataset in order to describe the variable of interest systolic blood pressure in terms of the other 4 variables. Regularisation parameters have been calibrated so that their corresponding models yield optimal regularisation. Table 1 below presents output from these 4 model fits (where "e-net" stands for "elastic net"). Analyse and comment on these results, for example in terms of:

- The effect of each of the regularisation schemes;
- Increase in model fit error;
- Potential "issues" or "challenges" within the dataset;
- Overall impact of, or necessity for regularisation for this data;
- Limitations of the output presented.

	GLM	ridge	e.net	LASSO
(Intercept)	56.710	64.703	65.854	64.707
Age	0.200	0.220	0.146	0.118
Waist	0.557	0.356	0.482	0.520
Cholesterol	0.003	0.004	0.000	0.000
BMI	0.030	0.378	0.048	0.000
Errors	139.76	141.96	143.26	143.26

Table 1: Table for Question 1.

Question 2

Load datasets `Hitters_train.csv` and `Hitters_test.csv` into R as follows:

```
Hitters.train = read.csv(file="Hitters_train.csv",
        stringsAsFactors=TRUE)
Hitters.test = read.csv(file="Hitters_test.csv",
        stringsAsFactors=TRUE)
```

(a) Fit a GLM to the training dataset `Hitters_train.csv`, and obtain predictions from this model for the test dataset `Hitters_test.csv`.

(b) Compute and compare the training and test Mean Squared Errors (MSEs) for the GLM fit.

(c) Fit a LASSO to the training dataset `Hitters_train.csv`, and obtain predictions from this model for the test dataset `Hitters_test.csv`.

(d) Compute and compare the training and test Mean Squared Errors (MSEs) for the LASSO fit.

(e) Interpret these results.

(f) Bonus: repeat the above process a large number of times (say, 100 times) by encapsulating your code within a for-loop. This framework will yield as many MSE values for GLM and LASSO model fits and test set predictions, and the distributions of these samples of MSEs can be compared to each other.

Question 3

Load datasets `titanic_train.csv` and `titanic_test.csv` (slightly modified versions of the dataset freely available from Kaggle) into R as follows:

```
dat.train = read.csv(file="data/titanic_train.csv",
        stringsAsFactors=TRUE)
dat.test = read.csv(file="data/titanic_test.csv",
        stringsAsFactors=TRUE)
```

Let us also get rid of redundant variables in the datasets:

```
dat.train$PassengerId = NULL
dat.test$PassengerId = NULL
dat.train$Name = NULL
dat.test$Name = NULL
dat.train$Ticket = NULL
dat.test$Ticket = NULL
dat.train$Cabin = NULL
dat.test$Cabin = NULL
```

Finally, let us also clean up a bit and get rid of incomplete observations in the training set:

```
# clean out a couple of tricky observations:
i.rm = which(!(dat.train$Embarked %in% levels(dat.test$Embarked)))
dat.train = dat.train[-i.rm,]
dat.train = droplevels(dat.train)
```

```
str(dat.train)

# remove incomplete observations:
dat.train = na.omit(dat.train)
dat.test = na.omit(dat.test)

# fix problem specification to classification:
dat.train$Survived = as.factor(dat.train$Survived)
dat.test$Survived = as.factor(dat.test$Survived)
```

(a) Fit a GLM to the training dataset `dat_train.csv`, and obtain predictions from this model for the test dataset `dat_test.csv`. You might have to adjust your call to `glm()` and `predict()` because of the fact that this is a classification task.

(b) Fit a LASSO to the training dataset `dat_train.csv`, and obtain predictions from this model for the test dataset `dat_test.csv`.

(c) Using R library `pROC`, perform ROC analyses on the test data for these two models. Compare the AUC values for these fits and interpret these results.

(d) Plot the two ROC curves and comment on this.

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

COLAISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

ST6026
Basics of Machine Learning

Workshop 3 - Statistical validation

Note: Questions 1 and 2 focus on implementing bootstrapping frameworks using for-loops. The exercises aim at developing a practical sense of what bootstrapping strategies boil down to. Some packages are available in R (and other statistical software environments) that provide off-the-shelf functions for bootstrapping, or have this functionality built into other functions that rely on this resampling scheme.

Question 1 (Bootstrap estimation of standard error)

Load the `law` dataset in the `bootstrap` package (containing law school intake data), estimate the correlation between the two variables in this dataset, and evaluate the bootstrap estimate of the standard error associated with this estimation.

Question 2 (Bootstrap linear regression estimates)

Consider R's `cars` dataset.

(a) Obtain relevant regression parameter estimates for this dataset.

(b) Generate $M = 10000$ bootstrap estimates for these coefficients.

(c) Inspect the one-dimensional (i.e. marginal) distributions for all relevant bootstrap parameter estimates, and state your conclusions.

(d) Inspect the joint distribution for these sample parameter estimates, and state your conclusions.

Note: Questions 3 and 4 explore cross-validation frameworks, to build a practical understanding of how they work. Question 3 involves coding for-loops to implement cross-validation step by step. Some simple "coding hacks" are used in order to make the code smaller. Do not be too worried about these coding aspects; again the main goal here is to break CV frameworks down into practical steps.

Question 3 (Cross-validation frameworks)

Load the following data from dataset `Boston` in library `MASS`:

```
x = Boston[, c("crim", "indus", "rm", "tax")]
y = Boston$medv
```

You will probably need to coerce `x` into a matrix before passing into `lm`:

```
x = as.matrix(x)
lmo = lm(y~x)
summary(lmo)
```

This linear model seems alright, but what about its predictive performance?

(a) Perform a 50%-50% train-test split of the dataset. Fit the linear model on the training data, generate predictions from this model for the test data, and calculate the corresponding prediction Root Mean Square Error (RMSE).

(b) Implement Leave-One-Out CV on the data (`x`, `y`) and calculate the LOO-CV test set prediction RMSE estimate.

(c) Implement K-fold CV on the data (`x`, `y`) and calculate the K-fold CV test set prediction RMSE estimate, using K=5.

(d) Implement K-fold CV on the data (`x`, `y`) and calculate the K-fold CV test set prediction RMSE estimate, using K=10.

(e) Compare the prediction error estimates obtained from (a), (b), (c), and (d).

Question 4

Table 1 and Figure 1 capture the output of 10-fold cross-validation of two distinct multilinear models applied to a dataset with $N = 263$ observations of major baseball league players on $P = 19$ variables.

(a) Indicate what errors **A**, **B**, **C** and **D** are likely to quantify in this comparative analysis, and why. Include an indication of which model each of these four quantities could relate to and why.

(b) Provide a possible explanation for the greater variances observed for distributions **C** and **D** compared to those of **A** and **B**.

(c) Name two multilinear models that could yield these results, and why.

(d) Which of the two models would you consider as the more reliable one? Why?

	A	**B**	**C**	**D**
K=1	0.34	0.36	0.57	0.50
K=2	0.33	0.35	0.58	0.53
K=3	0.36	0.38	0.26	0.24
K=4	0.36	0.38	0.28	0.30
K=5	0.33	0.34	0.55	0.60
K=6	0.35	0.37	0.40	0.39
K=7	0.36	0.38	0.27	0.26
K=8	0.34	0.35	0.50	0.53
K=9	0.34	0.36	0.42	0.41
K=10	0.35	0.37	0.39	0.38
Mean	**0.34**	**0.36**	**0.42**	**0.41**

Table 1: Mean squared errors obtained from K-fold cross-validation of the two distinct multilinear models.

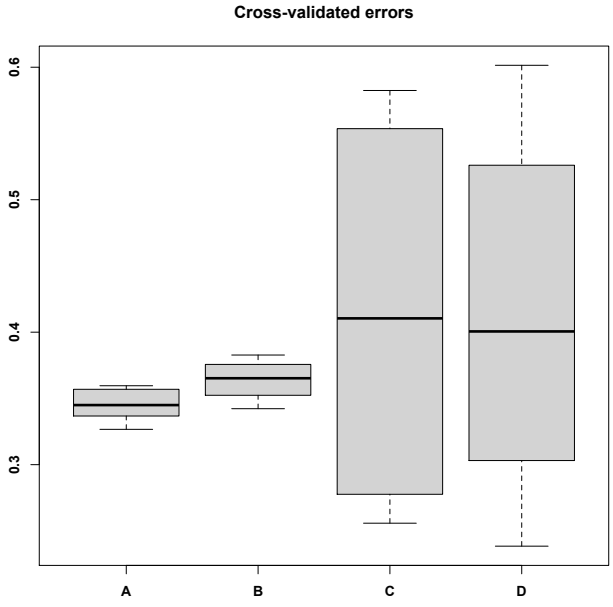

Figure 1: Distributions of cross-validated mean squared errors shown in Table 1.

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

COLAISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

ST6026
Basics of Machine Learning

Workshop 4 - Machine learning models

Eric Wolsztynski
eric.w@ucc.ie

Question 1

Which of the proposed techniques is best suited for the task described?

(a) A model that processes low-dose CT scan images for tumour detection.

  (i) Support Vector Machine
  (ii) Convolutional Neural Network

(b) Benchmarking of a sample of 80 distinct guitar models into 2 categories, on the basis of 60 features.

  (i) Linear logistic regression
  (ii) Support Vector Machine
  (iii) Artificial Neural Network

(c) Benchmarking of a sample of 80 distinct guitar models into 5 categories, on the basis of 60 features.

  (i) Linear logistic regression
  (ii) Support Vector Machine
  (iii) Artificial Neural Network

(d) A machine vision model used in a large warehouse that, given images of tagged items taken by a mobile robot, detects and identifies the tags with serial numbers, applies OCR to extract these numbers, and labels the images according to item category.

  (i) Linear logistic regression
  (ii) Convolutional Neural Network

Solution:

(a) CT scans: a CNN (to address image processing tasks).

(b) 2-class guitar benchmarking: logistic regression may offer better interpretability, even if the model does not perform as well.

(c) 5-class guitar benchmarking: SVM. logistic regression is no longer an option since it only applies to 2-class problems. An alternative classification model like Linear Discriminant Analysis (LDA) may also offer better interpretability, even if the model does not perform as well.

(d) Classification based on imaged tagged: a CNN (to address image processing tasks). Note: OCR – Optical character recognition

Question 2

Table 1 below provides a training dataset containing five observations, three predictors (X_1, X_2 and X_3), and a 2-class categorical response variable Y taking values "HIGH" and "LOW". Assume we use this data to train a kNN classifier and that scaling is not required for this task.

(a) Indicate what is the predicted value of test point (0,0,0) with k=1? Justify your answer.

(b) Indicate what is the predicted value of test point (0,0,0) with k=3? Justify your answer.

Table 1: Table for Question 2.

Observations	X_1	X_2	X_3	Y
1	0	2	0	HIGH
2	3	0	0	HIGH
3	0	1	3	LOW
4	0	1	2	HIGH
5	-1	0	1	LOW

Solution: a) Nearest neighbour is obs. 5, therefore, LOW

b) Nearest neighbours are obs. 5, 1, 4, therefore, HIGH

Question 3

Figure 3 illustrates the decision boundaries of 4 classification models applied to the same 2-class dataset. Indicate which of these scenarios the following models may correspond to:

(a) A random forest;

(b) A logistic regression model;

(c) A linear discriminant analysis with 3 Gaussian components;

(d) A support vector machine using a radial basis function;

(e) A lasso classifier with an extremely large shrinkage parameter;

(f) A kNN classifier (with k=5).

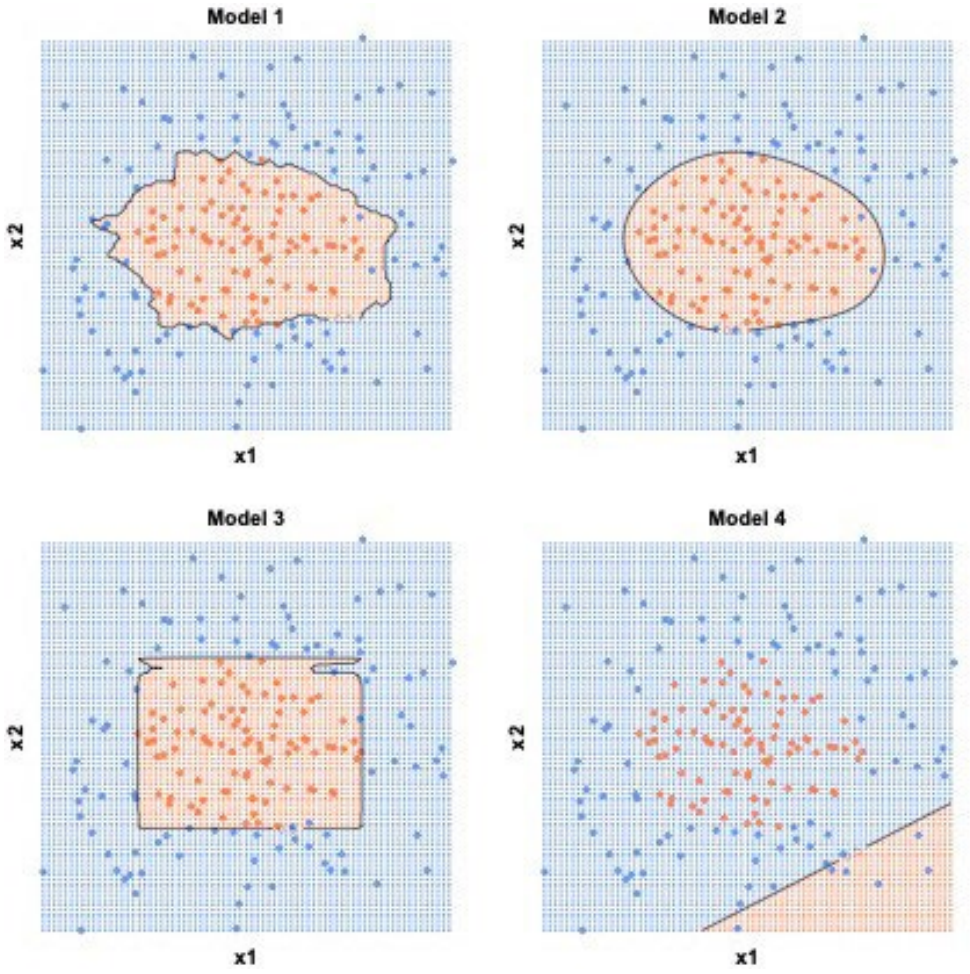

Figure 1: Decision boundary of 4 classification models applied to the same 2-class dataset.

Solution:	Model 1: (f)	A kNN classifier (with k=5)
	Model 2: (d)	A support vector machine using a radial basis function
	Model 3: (a)	A random forest
	Model 4: (b)	A logistic regression model

Question 4

Figure 4 below shows the output of a particular model fit to a sample of data points. This dataset comprises of 5 variables: Length, Width, Leaf, Curve and Age.

(a) Name the model represented by this diagram.

(b) Is this a regression or a classification problem? Explain why.

(c) Provide the predicted values for the following 5 test points according to the model of Fig. 4:

Table 2: Dataset for the model depicted in Figure 4.

	Length	Width	Leaf	Curve
Obs1	4.5	2.3	1.3	0.3
Obs2	5.0	3.5	4.3	0.3
Obs3	6.1	3.0	4.9	1.8
Obs4	7.2	3.0	5.8	1.6
Obs5	5.1	3.8	1.9	0.4

(d) Calculate the misclassification rate from the model, assuming the true test values were as follows:

Table 3: Test set for the model depicted in Figure 4 and Table 2.

Test point	Obs1	Obs2	Obs3	Obs4	Obs5
True value	Young	Young	Intermediate	Mature	Young

(e) Name the ensemble model defined by bootstrapping the model in (a), where all output predictions are averaged to generate final predictions, and where only 2 variables are considered randomly at each split.

Solution: a) Decision tree

b) Classification (categorized into Young, Intermediate or Mature)

c) Young, Intermediate, Mature, Mature, Young

d) 2/5 = 0.4

e) Random Forest

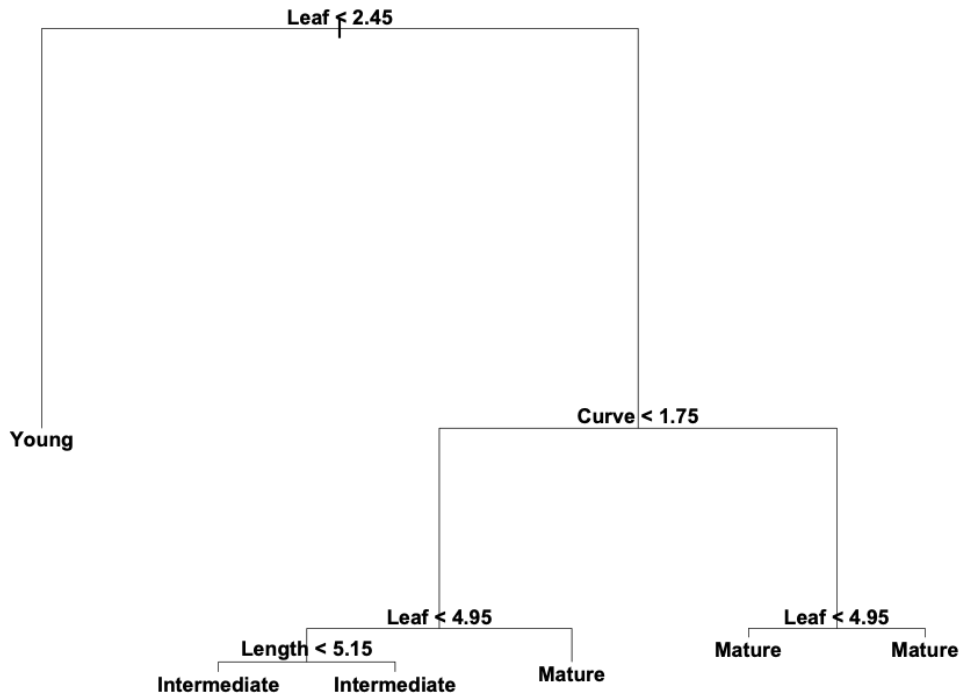

Figure 2: How do trees go online? They log in.

Question 5

Figure 5 below shows the output of some statistical analysis carried out on some dataset of predictors X and observations Y.

(a) Name the model illustrated by this figure.

(b) Is this a regression or classification model?

(c) Indicate how many predictors have been used in fitting this model.

(d) Indicate the number of hidden layers used in this model.

(e) Indicate the size of each hidden layer used in this model.

Solution: a) Neural Network

b) Regression

c) 6

d) 3

e) Layer 1: 4, Layer 2: 5, Layer 3: 2

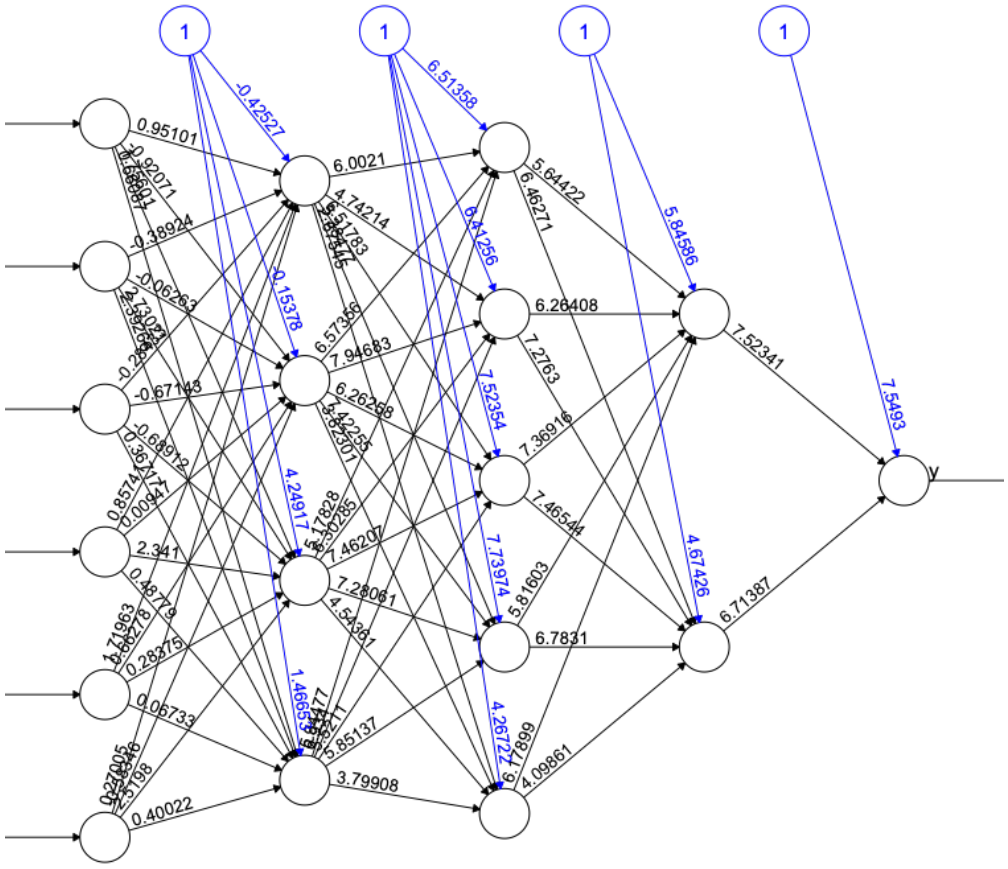

Figure 3: This model looks brainy...

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

COLAISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

ST6026
Basics of Machine Learning

Workshop 5 - Machine learning models and feature selection

For this workshop, we need the following:

```
source("ST6026_package.R")
library(glmnet)
library(caret)
library(ISLR)
library(MASS)
library(randomForest)
```

Question 1

Load the Hitters dataset (removing observations with missing values):

```
dat = na.omit(Hitters)
```

Create a training subset comprising of 70% of this dataset, and the corresponding test set, as follows:

```
i.cv = sample(1:nrow(dat), round(nrow(dat)*.70), replace=FALSE)
dat.cv = dat[i.cv,]
dat.cv.y = dat.cv$Salary
dat.cv$Salary = NULL
dat.test = dat[-i.cv,]
dat.test.y = dat.test$Salary
dat.test$Salary = NULL
```

Train a random forest to the training set via 5-fold CV as follows:

```
trc = trainControl(method="cv",number=5)
caret.kcv = train(dat.cv, dat.cv.y, method="rf", trControl=trc)
```

Train a random forest to the training set via 50 bootstrap resamples, by adapting the above code.

(a) Compare the outputs.

(b) Generate predictions for the test sets using each of these 2 outputs, and compare test set performances.

Question 2

Using again the data from Question 1 (`dat = na.omit(Hitters)`),

(a) Inspect and comment on the correlation matrix of the feature set (i.e. the set of predictors, excluding response `Salary`).

(b) Create 2 subsets of the original feature set, respectively:

- by eliminating highly correlated features (using `caret::findCorrelation()`);
- by pretending you're an MLB expert and knowing that you should remove the following variables: League, Division, Assists.

Using `caret::train()`, fit random forests to the full dataset and each of the two reduced datasets, via 5-fold CV, and:

- compare variable importance summaries (and comment);
- compare RMSEs (and comment).

(c) Carry out univariate t-tests between each of the input features and the response variable, and comment on the distribution of p-values from these.

(d) Perform RFE on the initial model (fit using the whole feature set) using `caret::rfe()`, and:

- identify the best feature subset;
- inspect variable importance and compare with rankings obtained from the random forest using the full set;
- compare RMSEs obtained using the full set and using RFE;
- repeat analysis on random forests fit to feature subset based on correlation pre-filtering.

(e) What was done wrong in this question?

Question 3

Load the Blood Pressure data from Workshop 2:

```
dat = read.table(file="data/Blood Pressure.txt", header=TRUE)
```

Fit GLM and ridge regression models to the whole feature set, as well as to a reduced set obtained by filtering out highly correlated features. Compare and comment on the estimated effects.

Question 4

Load the `cancer_data.csv` dataset:

```
dat = read.csv(file="data/cancer_data.csv")
```

(a) Inspect the correlation structure of the data, and think about its consequences on model building and prediction.

(b) Fit a logistic regression model the full set and comment on the output model.

(c) Adapting code from Workshop 3 Question 3, fit random forests to the full set and a reduced set via 5-fold CV, storing training set model accuracies and test fold prediction accuracies from each model. Use successively multiple t-tests and correlation filtering to generate the reduced feature sets. Compare model performances, including in terms of model overfitting.