STATS762 Regression for Data Science
Assignment 4
Due date: 11.59pm, 12 June 2019

- Please submit both your R Markdown document and a pdf file containing the document it generates. To create a pdf you should start your R Markdown document with the following lines (having made the appropriate changes):

````
---
title: "STATS 762 Assignment 3"
author: "Your Name, ID 1234567"
date: "Due: 12 June 2019"
output: pdf_document
---

```{r}
#Set the seed of R random number generator
#to obtain the same output when it is reproduced.
set.seed(1e5)
```
````

Data description for Q1 and Q2

The original data contains information for every player registered in the latest edition of FIFA 19 database. The source of data is `https://sofifa.com/`. Firstly we excluded players with any missing entry and only included variables of interest to create the master spreadsheet, Fifa2019.csv. In Fifa2019.csv there are total 18147 registered players and the attributes follow;

Column	Description
1	Overall - overall rate (scale of 100)
2	Position - Position on the pitch
3-36	34 performance scores (scale of 100) follow; Crossing,Finishing, HeadingAccuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle, GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes

For Questions 1 and 2, the study interests are focused on players whose position (2nd column) is either RDM (Right Defensive Midfielder) or RCM (Right Centre Midfielder) or LS (Left Striker).

[2 marks] Create the data `Fifa` only containing players whose position is either RDM or RCM or LS.

Answer the questions 1 and 2 using the data `Fifa`.

1. The first study interest is to predict the player's position (Second column) given 34 performance scores. (i.e., Overall rate is excluded.) We use the default specifications in R unless it is specified.

   (a) Fit the multinomial regression, linear discriminant analysis and quadratic discriminant analysis. We use the entire data for train and test data (no split). Which classification method gives the best overall prediction? Verify your answer. [8 marks]

   (b) Describe performance scores resulting both RDM and RCM using your best model in (a). [3 marks]

   (c) Predict the new player's position using the best model in (a). The performance score of the new player follows;

Overall	Crossing	Finishing	HeadingAccuracy	ShortPassing
69.51655	57.48700	57.71277	58.64657	68.83688
Volleys	Dribbling	Curve	FKAccuracy	LongPassing
54.40426	65.74468	57.09456	53.16312	63.45390
BallControl	Acceleration	SprintSpeed	Agility	Reactions
68.76123	67.00591	66.63475	68.67376	66.62648
Balance	ShotPower	Jumping	Stamina	Strength
67.78369	67.30378	67.24232	73.51773	69.20331
LongShots	Aggression	Interceptions	Positioning	Vision
61.43735	65.65839	55.46690	62.02719	63.98818
Penalties	Composure	Marking	StandingTackle	SlidingTackle
57.40189	65.89835	54.90898	55.46690	51.90544
GKDiving	GKHandling	GKKicking	GKPositioning	GKReflexes
10.69267	10.63357	10.83333	10.65248	10.69031

   [2 marks]

   (d) Plot the regression classification tree minimizing the cross-validation error. We use the entire data for train and test data (no split). Show your working. [3 marks]

   (e) Among the best model in (a) and the model fitted in (d), which model gives a higher overall predictability? Explain how your best model predicts the player's position. [4 marks]

2. The second study interest is predicting the overall score (First column) given 34 performance scores for the players with (i.e., Position is excluded). We use the entire data for both train and test data. We use the default specifications in R unless it is specified.

   (a) Plot an optimal regression tree and show your working. [3 marks]

   (b) A sport scientist wants to fit a gradient boosting regression tree and check if it gives a better prediction. A sport scientist can afford the maximum 4000 trees and the learning rate is 0.04. What is the optimal number of trees? [3 marks]

(c) A scientist wants to compare the predictability of your optimal regression tree in (a), optimal gradient boosting regression tree in (b) and your optimal linear regression model using lasso. Using the mean squared of residuals, pick the best model and verify your answer. [6 marks]

(d) Plot the residual against the overall score for the best model chosen in (c). [2 marks]

(e) Let's compare the relative variable importance of your optimal regression tree in (a) and your optimal gradient boosting regression tree in (b). Which variables are equally important? Describe why all variables are not equally important. [4 marks]

3. Question on spline will be added on the 31th of May. [10 marks]