PROJECT 3: CLASSIFICATION TYPE PROBLEMS
FMSF90F: STATISTICAL LEARNING AND VISUALIZATION 2023
Hand in: **17.00 on Monday March 6**

Intro

In this project you should use some popular methods used for classification and clustering, and interpret the results.

We encourage students to work in groups of 2, and that as many groups as possible work with data from (one of) the students.

Aim

This project consists of 2 main parts:

1. to try and compare the results of different supervised classification algorithms on a real dataset
2. to perform and interpret clustering methods

Data prerequisites

You can either use data from your own research, or use the datasets supplied with the project (Bigfoot for supervised learning and xxx for unsupervised learning) Note that your data must be in reasonably good shape so that manipulating the data is not too time consuming.

- Supervised learning
  - Reasonable size of data: ≥ 6 variables and > 100 observations, ideally many more.
  - Identify at least one **binary outcome variable** of interest, for which both categories appear in at least 20% of the observations, and at least 5 variables that are interesting to use as explanatory variables . It is OK to create a binary variable by thresholding a continuous or multi-category variable.
- Clustering
  - A dataset with continuous variables that you are interested in clustering
  - Ideally there is labels (i.e. some known categorisation) to compare your resulting clusters with

Report contents

The project consists of producing and presenting the following in a written report:

1. **Supervised classification algorithms**
   The first part of the project concerns classification of a binary outcome variable y from several potential predictors (x). You can use your own data, or the supplied `bigfoot` dataset.

(a) Start by randomly splitting your dataset in a training set (eg. 70% of observations) and a test set (eg. 30%). Please state a seed (`set.seed(nn)`) when you make the split so your results can be reproduced. Use the same split for all models you fit below.

For all analyses below *use the training data only* to fit the model (including variable selection etc). The test data is used for evaluation and comparison between the different models.

(b) Perform classification using

- logistic regression
- Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analyis
- Support Vector Machines (SVM)

(You might also want to try Random Forest or feature selection/LASSO in the logistic regression, but that is optional)

(c) Make confusion matrices for the predictions performed on the test set in 1b, and report the confusion matrices, along with sensitivity and specificity. Explain briefly what sensitivity and specificity means for the bigfoot example.

(d) Present a ROC-curve (with AUC) for the results of each of the methods

(e) Summarize the performance of the classifiers with words, based on the evaluations above. Which one would you choose? Justify briefly.

2. **Clustering**

Under construction!

The second part of the project is about (unsupervised) clustering algorithms. Use either your own data[1], or the data set (soon) supplied with this lab.

General requirements for the report

Report should have:

- Title page, authors, date, page numbers
- Introduction, results and/or conclusions
- been proof read
- Have language and spelling mistakes corrected
- Figures and tables that are
  - Numbered
  - Equipped with suitable captions
  - Referred to in the text
- text divided into paragraphs and well structured with clear and suitable section headings

The R-code and the data (if you use your own) should be submitted as separate files.

[1] If it fullfills the requirements stated above

Bigfoot observations

(This is a suitable data set if your group does not have access to relevant data of your own.)

The Bigfoot Field Researchers Organization (BFRO) is an organization dedicated to investigating the bigfoot mystery, and for years they have been collecting reported sightings in a database. They manually classify their reports into

- Class A: Clear sightings in circumstances where misinterpretation or misidentification of other animals can be ruled out with greater confidence
- Class B: Incidents where a possible bigfoot was observed at a great distance or in poor lighting conditions and incidents in any other circumstance that did not afford a clear view of the subject.

However, they wonder if this can be automated and done by a classification algorithm instead. So in this task, you will set up a few different classification algorithms for this aim, and evaluate their performance.

Feel free to look at the original data[2], `bigfoot_original` below, however in this task we will be using a slightly simplified version of the data set where we have extracted some variables of interest and deleted observations that are missing any of these variables of interest.

Download the data as below, and run through the provided cleaning/preparation steps:

```
# library(tidyverse) needed for the code below!)

bigfoot_original <- readr::read_csv("https://raw.githubusercontent.com/rford

# Prepare the data:
bigfoot <- bigfoot_original %>%
# Select the relevant covariates:
select(classification, observed, longitude, latitude, visibility) %>%
# Remove observations of class C ( second- or third hand accounts):
filter(classification != "Class␣C") %>%
# Turn into 0/1, 1 = Class A, 0 = Class B:
mutate(class = ifelse(classification == "Class␣A", 1, 0)) %>%
# Create new indicator variables for some words from the description:
mutate(fur = grepl("fur", observed),
howl = grepl("howl", observed),
saw = grepl("saw", observed),
heard = grepl("heard", observed)) %>%
# Remove unnecessary variables:
select(-c("classification", "observed")) %>%
# Remove any rows that contain missing values:
drop_na()
```

[2] https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2022/2022-09-13/bigfoot.csv

The data we will use for this task thus includes the following variables:

Variable description

class	the assigned class of the observation, coded as 1 = Class A, 0 = Class B
longitude	longitude of the observation
latitude	latitude of the observation
visibility	estimated visibility at the time and place of observation (higher value means better visibility)
fur	does the report contain the word "fur"? (`TRUE/FALSE`)
howl	does the report contain the word "howl"? (`TRUE/FALSE`)
saw	does the report contain the word "saw"? (`TRUE/FALSE`)
heard	does the report contain the word "heard"? (`TRUE/FALSE`)

End of Project 3