Week 1 : Revision

Q1. A data scientist is asked to design an algorithm to predict whether an email is a junk email or not from

- the time it was sent,
- the sender's email address,
- the presence of the keyword "buy", and
- the length of its text.

- To train the algorithm data points need to be represented as pairs, (X, Y), wher X is the model Input and Y is the model output.
- To simplify the algorithm structure, the sender's email address is transformed into a variable by replacing the character string with the number of digits in the address, e.g. nicolo.colombo2021@rhul.ac.uk would be replaced by 4.
- For practical purposes, presence of "buy" and the label are also transformed into variables, which can assum only values.
- Each X is then a vector with N entries, i.e. $X = [X_1, X_2, \ldots, X_n]$, with $N =$, and can be handled in standard ways.

After all transformation, the attributes are all quantitative variables, except for the presence of "buy", which is categorical.

- In particular, the non-categorical variables are time, then # of digits, and the length of the email's text.

Q5. Consider the following text
"my favourite book is elements of statistical learning"
and compute its bag-of-word vector representation associated with the dictionary:
~~D~~ D = ["my", "are", "is", "and", "of", "bad", "good", "favourite",
"hell", "hi", "data", "analysis", "statistical", elements"].
What is the sum of its entries?
$V = [1, 1, 0, 1, 1, 1, 1, 0]$ $\therefore$ sum = 6

Q2. The following text comes from the textbook An introduction to statistical learning.

Since that time, inspired by the advent of machine learning and other disciplines, statistical learning has emerged as a new subfield in statistics, focused on supervised and unsupervised modelling and prediction. In recent years, progress in statistical learning has been marked by the increasing availability of powerful and relatively user-friendly software, such as the popular and freely available R system. This has the potential to continue the transformation of the field from a set of techniques used and developed by statisticians and computer scientists to an essential toolkit for a much broader community.

Q3. Data can be structured or unstructured, but are almost always "dirty". select the correct statements

a. Data with missing values can be repained through statistical imputation techniques

b. A data set of plain-text documents is an example of unstructured data

h. Assum that your data consists of N d-dimensional data objects, e.g. $D = \{X_i\}_{i=1}^{N}$ with $X_i = [X_{i1}, \dots, X_{id}]^T$

Then the data set can be represented by a single $N \times d$ matrix.

Q4. Standardize the attributes of the following data set $D = (x_i, y_i)_{i=1}^{9} = \{(2,0), (5,1), (3,0), (2,1), (1,0), (10,0), (7,1), (2,0), (4,0)\}$ what is the new attribute of object (3,0)?

standardization $= x_i = \frac{x_i - \mu}{\sigma} =$

-0.36 In R; $SX = \frac{X - mean(x)}{sd(x)}$