Série 1 – Lien des exercices ici

2) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   Ici on est dans le contexte d'un problème de **régression** parce que la variable réponse est le salaire du CEO, qui prend des valeurs numériques. On est intéressés à faire de l'**inférence** parce qu'on veut comprendre l'impact des prédicteurs (profit, nombre d'employés, salaire de l'industrie) sur la variable réponse (salaire du CEO). Par exemple, on pourrait vouloir mesurer l'impact de l'ajout d'un employé sur le salaire du CEO.

   $n = 500$
   $p = 3$ (profit, number of employees et industry salary)

   b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   Ici on est dans un contexte de **classification** parce que la variable réponse (succès ou non) est de nature catégorielle. Plus précisément, c'est une variable binaire. On est intéressés à faire de la **prédiction** parce qu'on veut prédire si chaque produit sera un succès ou non sachant les variables.

   $n = 20$
   $p = 13$ (price charged, marketing budget, competition price et 10 autres variables)

   c) We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

   Ici on est dans un contexte de **régression** parce la variable réponse (le % de changement de valeur du dollar états-unien) est une variable numérique. On est intéressés à faire de la **prédiction.**

   $n = 52$
   $p = 3$ (% change in US market, % change in British market, % change in German market)

4) You will now think of some real-life applications for statistical learning.

   a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction ?

Explain your answer.

1. La classification peut être utile aux banques qui veulent déterminer quels clients feront défaut de paiement sur leur carte de crédit. La variable réponse serait alors la variable indicatrice d'un défaut de paiement (1 s'il y a défaut de paiement et 0 sinon). Les prédicteurs pourraient être le nombre de transactions durant la nuit, la cote de crédit, le sexe, l'âge, la région, etc. Ici on serait plus dans un contexte de prédiction car le but est de prédire quels détenteurs de carte de crédit feront défaut.
2. La classification est utilisée dans les logiciels de messagerie (comme Outlook) pour déterminer quels courriels sont indésirables. La variable réponse est l'indicatrice d'un courriel indésirable (1 si le courriel est indésirable et 0 sinon). Les prédicteurs peuvent inclure la présence d'un symbole de dollar, la présence de la chaine de caractères « Dear friend » ou « business proposal », etc. Dans ce cas, on serait dans un contexte de prédiction, parce qu'on veut classer les courriels dans 2 catégories. On n'est pas intéressés à connaitre l'impact de la présence d'un symbole de dollar sur la probabilité que le courriel soit indésirable, par exemple.
3. La classification peut être utile pour faire de la reconnaissance d'image. On pourrait par exemple vouloir classer de manière automatique des photos de paysages de forêts, de déserts et de plans d'eau. La variable réponse serait la classe de la photo (forêt, désert ou plan d'eau). Les prédicteurs pourraient être les pixels de l'image encodés en RGB. Ici on est encore dans un contexte de prédiction, car on veut classer les images, et donc prédire leur classe.

8) This exercise relates to the College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10% of high school class
- Top25perc : New students from top 25% of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student

- Grad.Rate : Graduation rate

Before reading the data into R, it can be viewed in Excel or a text editor.

c) i) Use the `summary()` function to produce a numerical summary of the variables in the data set.

Code disponible ici : https://github.com/francisduval/demo_act_6100

ii) Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first 10 columns of a matrix A using `A[ , 1:10]`.

Code disponible ici : https://github.com/francisduval/demo_act_6100

iii) Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.

Code disponible ici : https://github.com/francisduval/demo_act_6100

iv) Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

Code disponible ici : https://github.com/francisduval/demo_act_6100

v) Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow = c(2, 2))` useful : it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

Code disponible ici : https://github.com/francisduval/demo_act_6100

Série 2 – Lien des exercices ici

Exercice #1

On attribue à notre nouvelle observation l'indice $i = 0$. Donc notre nouvelle observation est $\boldsymbol{x}_0 = (x_{10}, x_{20}, x_{30}) = (1, 2, 1)$. Il faut calculer la distance euclidienne entre cette nouvelle observation et

chacune des observations de l'ensemble d'entrainement avec la formule suivante :

$$\begin{aligned} d(\boldsymbol{x}_0, \boldsymbol{x}_i) &= \sqrt{\sum_{j=1}^{3} (x_{j0} - x_{ji})^2} \\ &= \sqrt{(x_{10} - x_{1i})^2 + (x_{20} - x_{2i})^2 + (x_{30} - x_{3i})^2}, \quad i = 1, 2, 3, 4, 5, 6, \end{aligned}$$

ce qui nous donne que

$$\begin{aligned} d(\boldsymbol{x}_0, \boldsymbol{x}_1) &= \sqrt{(1-1)^2 + (2-2)^2 + (1-3)^2} = 2 \\ d(\boldsymbol{x}_0, \boldsymbol{x}_2) &= \sqrt{(1-0)^2 + (2-3)^2 + (1-0)^2} = \sqrt{3} \approx 1.73 \\ d(\boldsymbol{x}_0, \boldsymbol{x}_3) &= \sqrt{(1-4)^2 + (2-0)^2 + (1-0)^2} = \sqrt{14} \approx 3.74 \\ d(\boldsymbol{x}_0, \boldsymbol{x}_4) &= \sqrt{(1-1)^2 + (2-1)^2 + (1-1)^2} = 1 \\ d(\boldsymbol{x}_0, \boldsymbol{x}_5) &= \sqrt{(1-2)^2 + (2-2)^2 + (1-1)^2} = 1 \\ d(\boldsymbol{x}_0, \boldsymbol{x}_6) &= \sqrt{(1-2)^2 + (2-1)^2 + (1-1)^2} = \sqrt{2} \approx 1.41. \end{aligned}$$

Les $K = 3$ observations les plus proches de $\boldsymbol{x}_0$ sont $i = 4, 5, 6$. Avec $K = 3$, la valeur prédite est

$$\widehat{y}_0 = \frac{y_4 + y_5 + y_6}{3} = \frac{7 + 4 + 5}{3} \approx 5.33.$$

Il n'y a pas de plus proche voisin. En effet, il y en a 2, c'est-à-dire $i = 4, 5$. On calcule donc la valeur prédite en utilisant ces 2 observations. Avec $K = 1$, la valeur prédite est

$$\widehat{y}_0 = \frac{y_4 + y_5}{2} = \frac{7 + 4}{2} = 5.5.$$

Exercice #2

Code disponible ici : https://github.com/francisduval/demo_act_6100

Exercice #3

1. On a que

$$\boldsymbol{y} = \begin{bmatrix} 4 \\ 2 \\ 8 \\ 5 \\ 3 \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \\ 1 & 2 \\ 1 & 9 \\ 1 & 7 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Les paramètres qui minimisent l'erreur quadratique moyenne sont

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{y} \\
&= \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 4 & 2 & 9 & 7 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 4 \\ 1 & 2 \\ 1 & 9 \\ 1 & 7 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 4 & 2 & 9 & 7 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 2 \\ 8 \\ 5 \\ 3 \end{bmatrix} \\
&= \begin{bmatrix} 5 & 23 \\ 23 & 151 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 4 & 2 & 9 & 7 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 2 \\ 8 \\ 5 \\ 3 \end{bmatrix} \\
&= \frac{1}{5 \times 151 - 23 \times 23} \begin{bmatrix} 151 & -23 \\ -23 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 4 & 2 & 9 & 7 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 2 \\ 8 \\ 5 \\ 3 \end{bmatrix} \\
&= \begin{bmatrix} 0.6681416 & -0.10176991 \\ -0.1017699 & 0.02212389 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 4 & 2 & 9 & 7 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 2 \\ 8 \\ 5 \\ 3 \end{bmatrix} \\
&= \begin{bmatrix} 0.56637168 & 0.26106195 & 0.46460177 & -0.24778761 & -0.04424779 \\ -0.07964602 & -0.01327434 & -0.05752212 & 0.09734513 & 0.05309735 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 2 \\ 8 \\ 5 \\ 3 \end{bmatrix} \\
&= \begin{bmatrix} 5.132743 \\ -0.159292 \end{bmatrix}.
\end{aligned}
$$

2. On ne peut pas vraiment ajuster un modèle KNN. En fait, l'algorithme va « mémoriser » tout les jeu de données, et tout le travail est fait lors de la prédiction de la variable réponse d'une nouvelle observation.

3. On prédit premièrement la valeur de la variable réponse de $i = 1$ en utilisant les 4 autres données. Les 2 plus proches voisins de $i = 1$ sont $i = 2, 3$. La prédiction est donc

$$\widehat{y}_1 = \frac{y_2 + y_3}{2} = \frac{2 + 8}{2} = 5.$$

On prédit ensuite la valeur de y_2 en utilisant les 4 autres données. L'observation $i = 2$ a 2 voisins qui sont à égalité ($i = 1$ et $i = 5$), on utilise donc les 3 plus proches voisins. Les 3 plus proches voisins de $i = 2$ sont $i = 1, 3, 5$. La prédiction est donc

$$\widehat{y}_2 = \frac{y_1 + y_3 + y_5}{3} = \frac{4 + 8 + 3}{3} = 5.$$

De la même manière,

$$\begin{aligned}
\widehat{y}_3 &= \frac{y_1 + y_2}{2} = \frac{4+2}{2} = 3 \\
\widehat{y}_4 &= \frac{y_2 + y_5}{2} = \frac{2+3}{2} = 2.5 \\
\widehat{y}_5 &= \frac{y_2 + y_4}{2} = \frac{2+5}{2} = 3.5.
\end{aligned}$$

Finalement,

$$\begin{aligned}
\text{MSE} &= \frac{1}{5}\sum_{i=1}^{5}(y_i - \widehat{y}_i)^2 \\
&= \frac{(4-5)^2 + (2-5)^2 + (8-3)^2 + (5-2.5)^2 + (3-3.5)^2}{5} \\
&= 8.3
\end{aligned}$$