Examen de Science des données 2022

MINES Paris - Tronc Commun 1A

Durée : 2h. Tous documents autorisés.

Problème (10 points) : méthode des k plus proches voisins

Étant donné un jeu de données $\mathcal{D} = \{(\vec{x}^{\,i}, y^i)\}_{i=1,\dots,n}$ de n observations étiquetées, $\vec{x}^{\,i} \in \mathbb{R}^p$, $y^i \in \{0,1\}$ (problème de classification binaire), et une distance d sur $\mathbb{R}^p$, on appelle algorithme des k plus proches voisins l'algorithme qui consiste à prédire comme étiquette pour un certain $\vec{x} \in \mathbb{R}^p$ l'étiquette majoritaire des k points du jeu de données les plus proches de $\vec{x}$: si on note $N_k(\vec{x})$ l'ensemble des k observations de $\mathcal{D}$ les plus proches de $\vec{x}$, on a

$$f(\vec{x}) \in \arg\max_{c \in \{0,1\}} \sum_{i : \vec{x}^{\,i} \in N_k(\vec{x})} \mathbb{1}_{y^i = c}.$$

Algorithme du 1 plus proche voisin

Dans un premier temps on se restreint au cas où $k = 1$: l'algorithme prédit l'étiquette du plus proche voisin. On utilise la distance euclidienne. Clémence a le jeu de données de classification binaire suivant, où $p = 2$ et $n = 6$:

x_1	1	2	2	2	3	3
x_2	2	1	2	3	1	2
y	1	1	0	1	0	1

On représente ces données sur le graphique ci-dessous, où les croix correspondent à la classe 1, et les ronds à la classe 0.

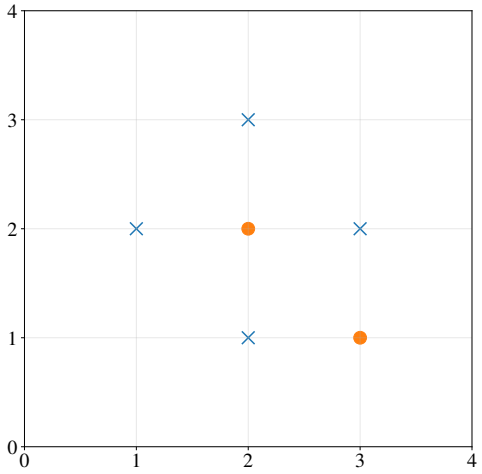

Figure 1

1. *(1 point)* L'algorithme des 1 plus proches voisins est-il paramétrique ou non-paramètrique ? Justifier.

Solution: C'est un algorithme non paramétrique : la fonction de décision s'exprime en fonction des données d'entraînement et non pas comme une fonction analytique fonction des variables. □

2. *(1 point)* On donne une nouvelle observation $\vec{x} = (4, 0.5)$. Quelle est la prédiction de l'algorithme du 1 plus proche voisin entraîné sur les données de Clémence ?

   Solution: Le point le plus proche de $(4, 0.5)$ est $(3, 1)$ qui est de la classe 0, on a donc $f(\vec{x}) = 0$. □

3. *(1 point)* On utilise la perte 0/1 : pour un couple $(\vec{x}, y) \in \mathbb{R}^p \times \{0, 1\}$, $L(y, f(\vec{x})) = \mathbb{1}_{f(\vec{x}) \neq y}$. Quelle est la valeur du risque empirique sur les données d'entraînement de l'algorithme du 1 plus proche voisin ?

   Solution: Sur les données d'entraînement, chaque observation est son propre plus proche voisin : l'algorithme classifie donc parfaitement les données d'entraînement, le risque empirique vaut 0. □

4. *(2 points)* Représenter sur la figure 1 la frontière de décision de l'algorithme du 1 plus proche voisin. (On rappelle que la frontière de décision est une courbe dans $\mathbb{R}^p$ qui sépare les prédictions de chacune des classes : d'un côté de la courbe, l'algorithme prédit la classe 1, de l'autre côté la classe 0).

   Solution: □

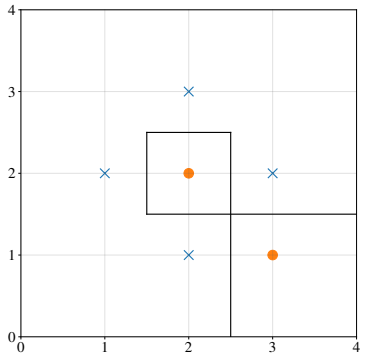

Figure 2

Algorithme des k plus proches voisins

On considère maintenant l'algorithme général des k plus proches voisins.

5. *(1 point)* On prend $k = 3$ et on considère l'algorithme entraîné sur les données de Clémence. Quelle est la prédiction de l'algorithme pour la nouvelle observation $\vec{x} = (4, 0.5)$?

   Solution: Les deux points plus proches de $\vec{x}$ en plus de $(3, 1)$ sont $(3, 2)$ et $(2, 1)$ qui sont tous les deux de la classe 1. On a donc dans ce cas $f(\vec{x}) = 1$. Cette prédiction est différente du cas $k = 1$! □

6. *(2 points)* Quelle est la valeur du risque empirique sur les données d'entraînement de l'algorithme des 3 plus proches voisins ? Que peut-on en déduire sur la performance de l'algorithme des 3 plus proches voisins par rapport à l'algorithme du 1 plus proche voisin ?

x_1	1	2	2	2	3	3
x_2	2	1	2	3	1	2
y	1	1	0	1	0	1
$f(\vec{x})$	1	0	1	1	1	0

Solution: On calcule les prédictions pour chacun des points du jeu d'entraînement, reportées dans le tableau ci-dessous.

L'algorithme des 3 plus proches voisins se trompe pour 4 données parmis les 6, le risque empirique vaut donc

$$R_n(f) = \frac{1}{6}\sum_{i=1}^{6} L(y^i, f(\vec{x}^{\,i})) = \frac{1}{6}\sum_{i=1}^{6} \mathbb{1}_{y^i \neq f(\vec{x}^{\,i})} = \frac{4}{6} = \frac{2}{3}.$$

L'erreur d'entraînement est plus élevée que pour l'algorithme du 1 plus proche voisin. Néanmoins on ne peut pas conclure sur les capacités de généralisation de chacun de ces algorithmes : on est probablement en surapprentissage dans le cas de l'algorithme du 1 plus proche voisin. Pour conclure sur la performance de ces algorithmes, il faudrait calculer l'erreur sur des données de test. □

7. *(1 point)* Quand k augmente, le risque de sur-apprentissage augmente-t-il ou diminue-t-il ?

Solution: Quand k augmente, la frontière de décision devient plus lisse : on moyenne les prédiction sur plus d'observations. Le risque de surapprentissage diminue donc quand k augmente. □

8. *(1 point)* Ayant à disposition un jeu d'entraînement de n observations et p variables, comment choisireriez-vous la valeur de k ?

Solution: k est un hyperparamètre du modèle, on le sélectionnera donc soit en prenant la valeur qui minimise l'erreur sur un jeu de validation, soit en utilisant de la validation croisée. □

Problème (25 points): modèles linéaires

Nous considérons un problème de régression en dimension p : nous disposons d'un jeu d'apprentissage $\mathcal{D} = \{(\vec{x}^{\,i}, y^i)\}_{i=1,\dots,n}$, tel que $\vec{x}^{\,i} \in \mathbb{R}^{p+1}$ et $y^i \in \mathbb{R}$. Nous supposons que chaque vecteur $\vec{x}^{\,i}$ contient un 1 comme premier coefficient. On note $X \in \mathbb{R}^{n\times(p+1)}$ la matrice de design qui a pour lignes les $\vec{x}^{\,i}$ et $\Sigma \in \mathbb{R}^{(p+1)\times(p+1)}$ la matrice

$$\Sigma = \frac{1}{n} X^T X.$$

Dans tout le problème, on fera l'hypothèse que la matrice Σ est inversible, et on définit la norme $\|\cdot\|_\Sigma$ sur $\mathbb{R}^{p+1}$ par, pour tout $\vec{x} \in \mathbb{R}^{p+1}$,

$$\|\vec{x}\|_\Sigma = \sqrt{\vec{x}^{\,T}\Sigma\vec{x}}.$$

On admettra que cette norme est bien définie. Pour tout $i = 1, \dots, n$, on suppose que y^i est la réalisation d'une variable aléatoire $Y^i \in \mathbb{R}$ décrite par le modèle probabiliste suivant : il existe $\vec{\beta} \in \mathbb{R}^{p+1}$ tel que

$$Y^i = \langle \vec{\beta}, \vec{x}^{\,i}\rangle + \varepsilon^i = \vec{\beta}^{\,T}\vec{x}^{\,i} + \varepsilon^i = \beta_0 + \beta_1 x_1^i \dots \beta_p x_p^i + \varepsilon^i, \tag{1}$$

où ε^i est une variable aléatoire **gaussienne** d'espérance nulle et de variance σ^2. On suppose les ε^i indépendants et identiquement distribués. On notera que le paramètre $\vec{\beta}$ est **inconnu** et **fixé**: nous chercherons dans l'exercice à en construire un estimateur.

On notera $Y = (Y^1, \dots, Y^n)^T \in \mathbb{R}^n$ le vecteur aléatoire constitué des n réponses Y^i, $\vec{y} = (\vec{y}^{\,1}, \dots, \vec{y}^{\,n})^T \in \mathbb{R}^n$ sa réalisation et de la même manière $\vec{\varepsilon} = (\varepsilon^1, \dots, \varepsilon^n)^T \in \mathbb{R}^n$ le vecteur aléatoire de bruits.

Attention ! Dans tout l'exercice, les $\vec{x}_i$ sont supposés fixés et ne sont pas des vecteurs aléatoires.

Minimisation du risque empirique

1. *(1 point)* Montrer que le modèle (1) revient à supposer que pour tout $i = 1, \dots, n$, Y^i a pour densité

$$g_{Y^i}(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y - \langle\vec{\beta}, \vec{x}^{\,i}\rangle\right)^2}{2\sigma^2}\right), \quad y \in \mathbb{R}.$$

Solution: Pour tout $i = 1, \dots, n$, comme ε^i a pour loi $\mathcal{N}(0, \sigma^2)$, l'équation (1) est équivalente a dire que

$$Y^i \sim \mathcal{N}\big(\langle \vec{\beta}, \vec{x}^{\,i} \rangle, \sigma^2\big).$$

On déduit le résultat en utilisant la formule de la densité d'une loi gaussienne. □

2. *(2 points)* Quel est l'espace des hypothèses $\mathcal{F}$ associé à ce modèle ? Ce modèle est-il paramétrique ? On choisit d'utiliser la fonction de coût quadratique $L(y, f(\vec{x})) = (y - f(\vec{x}))^2$. Écrire le problème de minimisation de risque empirique associé à cet espace d'hypothèses et cette fonction de coût.

Solution: On suppose en (1) que la relation entre $\vec{x}^{\,i}$ et Y^i est linéaire, ce qui corrrespond à un modèle de régression linéaire paramétrique:

$$\mathcal{F} = \big\{ f : \vec{x} \to \langle \vec{b}, \vec{x} \rangle,\ \vec{b} \in \mathbb{R}^{p+1} \big\}.$$

Le problème de minimisation du risque empirique s'écrit alors

$$\vec{\beta}^* \in \operatorname*{arg\,min}_{\vec{b} \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \Big(y^i - \big(b_0 + \sum_{j=1}^{p} b_j x_j \big) \Big)^2.$$

□

3. *(2 points)* Pour tout vecteur $\vec{b} \in \mathbb{R}^{p+1}$, le risque associé à $\vec{b}$ peut s'écrire comme

$$\mathcal{R}(\vec{b}) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (Y^i - \langle \vec{b}, \vec{x}^{\,i} \rangle)^2 \right] = \mathbb{E}\left[\frac{1}{n} \|Y - X\vec{b}\|_2^2 \right].$$

Montrer qu'on a

$$\mathcal{R}(\vec{b}) = \|\vec{b} - \vec{\beta}\|_\Sigma^2 + \sigma^2, \tag{2}$$

où $\vec{\beta}$ est le "vrai" vecteur de coefficients défini dans le modèle (1).

Solution: On vérifie que

$$\begin{aligned}
\mathcal{R}(\vec{b}) &= \mathbb{E}\left[\frac{1}{n} \|X\vec{\beta} - X\vec{b} + \vec{\varepsilon}\|_2^2 \right] \\
&= \frac{1}{n} \mathbb{E}\left[\|X\vec{\beta} - X\vec{b}\|_2^2 + \|\vec{\varepsilon}\|_2^2 + 2(X\vec{\beta} - X\vec{b})^T \vec{\varepsilon} \right] \\
&= (\vec{b} - \vec{\beta})^T \frac{1}{n} X^T X (\vec{\beta} - \vec{b}) + \sigma^2 + \frac{2}{n} (X\vec{\beta} - X\vec{b}) \mathbb{E}[\vec{\varepsilon}] \\
&= \|\vec{b} - \vec{\beta}\|_\Sigma^2 + \sigma^2.
\end{aligned}$$

□

Estimation par maximum de vraisemblance

4. *(2 points)* On suppose maintenant qu'on a un estimateur $\widehat{B}$ du paramètre $\vec{\beta}$: $\widehat{B}$ est un vecteur aléatoire qui dépend de $(Y^1, \dots, Y^n)$. Montrer que l'espérance du risque $\mathbb{E}[\mathcal{R}(\widehat{B})]$ peut s'écrire sous la forme:

$$\mathbb{E}[\mathcal{R}(\widehat{B})] - \sigma^2 = \mathbb{E}\Big[\big\| \widehat{B} - \mathbb{E}[\widehat{B}] \big\|_\Sigma^2 \Big] + \big\| \mathbb{E}[\widehat{B}] - \vec{\beta} \big\|_\Sigma^2. \tag{3}$$

Dans cette expression, comment peut-on interpréter chacun des termes du membre de droite?

Solution: On a:

$$\begin{aligned}\|\widehat{B}-\vec{\beta}\|_\Sigma^2 &= \|\widehat{B}-\mathbb{E}[\widehat{B}]+\mathbb{E}[\widehat{B}]-\vec{\beta}\|_\Sigma^2 \\ &= \|\widehat{B}-\mathbb{E}[\widehat{B}]\|_\Sigma^2+\|\mathbb{E}[\widehat{B}]-\vec{\beta}\|_\Sigma^2+\frac{2}{n}(\widehat{B}-\mathbb{E}[\widehat{B}])X^TX(\mathbb{E}[\widehat{B}]-\vec{\beta}).\end{aligned}$$

En prenant l'espérance, comme $\mathbb{E}\Big[\frac{2}{n}(\widehat{B}-\mathbb{E}[\widehat{B}])X^TX(\mathbb{E}[\widehat{B}]-\vec{\beta})\Big]=\frac{2}{n}\mathbb{E}\Big[(\widehat{B}-\mathbb{E}[\widehat{B}])\Big]X^TX(\mathbb{E}[\widehat{B}]-\vec{\beta})=0$, on en déduit que

$$\mathbb{E}[\|\widehat{B}-\vec{\beta}\|_\Sigma^2]=\underbrace{\mathbb{E}\left[\|\widehat{B}-\mathbb{E}[\widehat{B}]\|_\Sigma^2\right]}_{\mathrm{Var}(\widehat{B})}+\underbrace{\|\mathbb{E}[\widehat{B}]-\vec{\beta}\|_\Sigma^2}_{\mathrm{Bias}^2},$$

d'où le résultat par (2). □

5. *(2 points)* Donner l'expression de la log-vraisemblance de l'échantillon $(y_1,y_2,\dots,y_n)$. Qu'en conclure sur l'estimation par maximum de vraisemblance (on comparera à l'expression obtenue à la question 2.) ?

Solution: On a, par indépendance des $(Y^i)_{1\le i\le n}$:

$$P(Y^1,\dots,Y^n;\vec{b})=\prod_{i=1}^n P(Y^i;\vec{b}).$$

On en déduit la log-vraisemblance de l'échantillon:

$$\ell(y_1,\dots,y_n;\vec{b})=\sum_{i=1}^n\left(-\ln(\sigma\sqrt{2\pi})-\frac{(y^i-\langle\vec{b},\vec{x}^{\,i}\rangle)^2}{2\sigma^2}\right).$$

Le premier terme étant indépendant de $\vec{b}$, maximiser la log-vraisemblance revient donc à minimiser la quantité

$$\frac{(y^i-\langle\vec{b},\vec{x}^{\,i}\rangle)^2}{2\sigma^2}.$$

On retrouve exactement la minimisation du risque empirique de la question 2 : pour le modèle gaussien, estimer par maximum de vraisemblance est équivalent à la minimisation du risque empirique avec la fonction de coût quadratique. □

6. *(2 points)* En déduire l'estimation par maximum de vraisemblance $\widehat{\beta}_{MLE}$ de $\vec{\beta}$. L'estimateur par maximum de vraisemblance, qu'on note $\widehat{B}_{MLE}$, est-il biaisé?

Solution: La log-vraisemblance est maximale pour $\widehat{\beta}_{MLE}$ solution de

$$\widehat{\beta}_{MLE}\in\operatorname*{arg\,min}_{\vec{b}\in\mathbb{R}^{p+1}}\frac{1}{2}\sum_{i=1}^n(y^i-\vec{b}^{\,T}\vec{x}^{\,i})^2.$$

Sous forme matricielle, on a:

$$\widehat{\beta}_{MLE}\in\operatorname*{arg\,min}_{\vec{b}\in\mathbb{R}^{p+1}}\frac{1}{2}\|\vec{y}-X\vec{b}\|^2.$$

C'est un problème de minimisation quadratique et convexe, on trouve la solution en annulant le gradient :

$$-2X^\top\left(\vec{y}-X\widehat{\beta}_{MLE}\right)=0\quad\Leftrightarrow\quad\widehat{\beta}_{MLE}=(X^TX)^{-1}X\vec{y}$$

L'estimateur par maximum de vraisemblance est donc

$$\widehat{B}_{MLE}=(X^TX)^{-1}XY.$$

Son biais est

$$\mathbb{E}[\widehat{B}_{MLE}]-\vec{\beta}=\mathbb{E}[(X^TX)^{-1}XY]-\vec{\beta}=(X^TX)^{-1}X^TX\vec{\beta}+(X^TX)^{-1}X\mathbb{E}[\vec{\varepsilon}]-\vec{\beta}=0.$$

L'estimateur du maximum de vraisemblance de $\vec{\beta}$ est non-biaisé. □

7. *(2 points)* Quelle est la variance de cet estimateur $\widehat{B}_{MLE}$? On rappelle que dans le cas de l'estimation d'un vecteur de paramètres, la variance est définie comme

$$\mathbb{V}(\widehat{B}_{MLE}) = \mathbb{E}\Big[\big\|\widehat{B}_{MLE} - \mathbb{E}[\widehat{B}_{MLE}]\big\|_2^2\Big].$$

On remarquera de plus que pour tout vecteur $\vec{x} \in \mathbb{R}^{p+1}$, on a:

$$\|\vec{x}\|_2^2 = \vec{x}^T\vec{x} = \mathrm{Tr}(\vec{x}\vec{x}^T),$$

où pour tout $A \in \mathbb{R}^{(p+1)\times(p+1)}$, $\mathrm{Tr}(A)$ désigne la trace de la matrice A.

Solution: On vérifie que

$$\widehat{B}_{MLE} - \vec{\beta} = (X^TX)^{-1}X^T\vec{\varepsilon}, \tag{4}$$

de sorte que

$$\|\widehat{B}_{MLE} - \vec{\beta}\|_2^2 = \mathrm{Tr}[(X^TX)^{-1}X^T\vec{\varepsilon}\vec{\varepsilon}^TX(X^TX)^{-1}].$$

Or, on sait que $\mathbb{E}[\vec{\varepsilon}\vec{\varepsilon}^T] = \sigma^2 I$. Par conséquent, comme $\widehat{B}_{MLE}$ est non-biaisé,

$$\begin{aligned}
\mathbb{V}(\widehat{B}_{MLE}) &= \mathbb{E}\big[\|\widehat{B}_{MLE} - \beta\|_2^2\big] \\
&= \mathbb{E}[\mathrm{Tr}((X^TX)^{-1}X^T\vec{\varepsilon}\vec{\varepsilon}^TX(X^TX)^{-1})] \\
&= \mathrm{Tr}[(X^TX)^{-1}X^T\mathbb{E}[\vec{\varepsilon}\vec{\varepsilon}^T]X(X^TX)^{-1}] \\
&= \frac{\sigma^2\Sigma^{-1}}{n}.
\end{aligned}$$

□

8. *(2 points)* En déduire que l'espérance du risque associé à l'estimateur du maximum de vraisemblance est

$$\mathbb{E}[\mathcal{R}(\widehat{B}_{MLE})] - \sigma^2 = \frac{\sigma^2(p+1)}{n}. \tag{5}$$

Solution: D'après le résultat de la question 2, comme l'estimateur $\widehat{B}_{MLE}$ est non-biaisé, on a:

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(\widehat{B}_{MLE})] &= \mathbb{E}\big[\|\widehat{B}_{MLE} - \vec{\beta}\|_\Sigma^2\big] \\
&= \frac{1}{n}\mathbb{E}[\|X(\widehat{B}_{MLE} - \vec{\beta})\|_2^2] \\
&= \frac{1}{n}\mathbb{E}[\mathrm{Tr}(X(\widehat{B}_{MLE} - \vec{\beta})(\widehat{B}_{MLE} - \vec{\beta})^TX^T)] \\
&= \frac{1}{n}\mathrm{Tr}(X\mathrm{Var}(\widehat{B}_{MLE})X^T) \\
&= \mathrm{Tr}\left(\mathrm{Var}(\widehat{B}_{MLE})\Sigma\right) \\
&= \mathrm{Tr}\left(\frac{\sigma^2 I}{n}\right) \\
&= \frac{\sigma^2(p+1)}{n}.
\end{aligned}$$

□

9. *(1 point)* La quantité $\mathbb{E}[\mathcal{R}(\widehat{B}_{MLE})]$ correspond à l'erreur de généralisation du modèle linéaire. Quelle est sa limite quand $n \to \infty$? Commenter cette limite.

Solution: Quand n devient grand, l'erreur de généralisation diminue et tend vers σ^2 : on a de plus en plus de données pour apprendre donc l'algorithme s'améliore. Néanmoins l'erreur ne devient pas nulle : σ^2 est un terme d'erreur irréductible qui provient du bruit ε. □

10. *(2 points)* Rappeler l'expression de l'erreur d'entraînement après minimisation du risque empirique. Commenter sur la valeur d'une part de cette erreur d'entraînement et d'autre part de l'erreur de généralisation lorsque $p \gg n$. Comment appelle-t-on ce phénomène en pratique ?

Solution: L'erreur d'entraînement est

$$\min_{\vec{\beta}\in\mathbb{R}^{p+1}} \frac{1}{n}\sum_{i=1}^{n}(y^i - \langle\vec{\beta}, \vec{x}^{\,i}\rangle)^2.$$

Quand p augmente, l'erreur d'entraînement diminue : on a plus de paramètres sur lequels minimiser la fonction. À l'inverse, on voit d'après la question précédente que l'erreur de généralisation augmente linéairement avec p. On appelle ce phénomène le sur-apprentissage : on a une erreur d'entraînement faible mais une mauvaise généralisation. □

11. *(1 point)* Ce modèle est-il adapté au cas $p \gg n$? Si non, proposer un modèle qui conviendrait mieux.

Solution: Non ce modèle n'est pas adapté : l'erreur de généralisation (le risque de la question 8) est grand dans ce cas. On peut régulariser le modèle en ajoutant une pénalisation ℓ_1 ou ℓ_2 (modèle Ridge ou Lasso). □

Prédiction de la concentration d'ozone

Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone dans l'atmosphère. On cherche à prédire la concentration maximale d'ozone dans la journée (O3) à partir de la température à midi (T12), la force du vent (Vx) et la nébulosité à midi (fraction de ciel couverte par des nuages, Ne12). On dispose de $n = 1014$ données journalières. On représente ci-dessous la matrice des nuages de points de ces différentes variables.

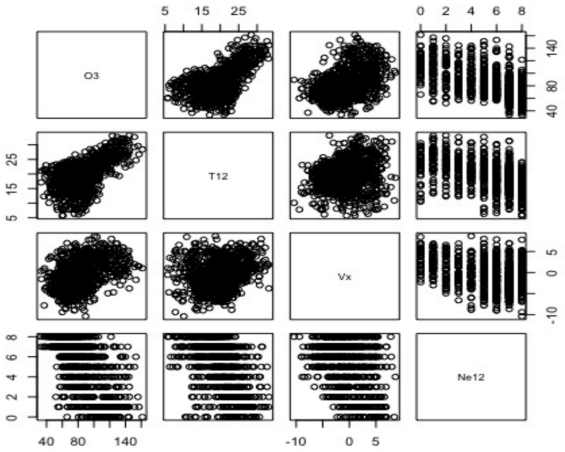

Figure 3

12. *(1 point)* Quelles variables semblent les plus corrélées ? Proposer un premier modèle de régression linéaire tel que $p = 1$.

Solution: L'ozone O3 semble le plus corrélé à la température à midi T12. □

13. *(1 point)* On s'intéresse maintenant au modèle

$$\mathrm{O3}_i = \beta_0 + \beta_1 T_i + \beta_2 V_i + \beta_3 N_i + \varepsilon_i, \tag{6}$$

où $\mathrm{O3}_i$ est la concentration d'ozone au jour i, T_i la température à midi, V_i la force du vent et N_i la nébulosité. Par quoi peut-être causé le terme de bruit ε_i ?

Solution: Le bruit ε_i peut avoir différentes sources : une erreur de mesure de la concentration d'ozone ou le fait que les 3 variables considérées ne suffisent pas a priori à prédire la concentration maximale d'ozone journalière. Par exemple, la température à une autre heure que midi peut influer, ou d'autres paramètres telle que la pression ou les précipitations. □

14. *(2 points)* Formuler sous la forme d'un test d'hypothèse (hypothèse nulle et hypothèse alternative) sur les paramètres du modèle (6) les question suivantes :

   ($Q1$) Est-ce que la concentration d'ozone maximale est influencée par la variable vent V ?

   ($Q2$) Est-ce que la valeur de O3 est influencée par le vent V ou la température T ?

   Solution: La première question correspond au test $\mathcal{H}_0 : \beta_2 = 0$ contre $\mathcal{H}_1 : \beta_2 \neq 0$. La deuxième question correspond au test $\mathcal{H}_0 : \beta_2 = \beta_3 = 0$ contre $\mathcal{H}_1 : \beta_2 \neq 0$ ou $\beta_3 \neq 0$. □

15. *(2 points)* Sachant que l'estimateur par maximum de vraisemblance du paramètre β_2 a pour loi $B_2 \sim \mathcal{N}(\beta_2, \sigma_2^2)$, où $\sigma_2^2 = \frac{\sigma^2}{n}(\Sigma^{-1})_{22}$, et supposant que σ est connu, proposer une statistique de test pour répondre à la question ($Q1$). Donner sa loi sous $\mathcal{H}_0$ ainsi que la zone de rejet pour un niveau de signification $\alpha = 5\%$.

   Solution: On prend comme statistique de test

$$Z = \frac{B_2 - \beta_2}{\sigma_2}.$$

   Sous $\mathcal{H}_0$, on a $Z \sim \mathcal{N}(0,1)$. Comme on fait un test bilatéral, on prend comme zone de rejet l'intervalle $]-\infty, z_0[\cup]z_0, +\infty[$, où z_0 est la valeur critique telle que

$$\mathbb{P}_0(|Z| > z_0) = \alpha.$$

   Par symmétrie, on prend z_0 tel que $\mathbb{P}_0(Z < -z_0) = \alpha/2$. Pour une gaussienne centrée réduite et $\alpha = 5\%$, on a $z_0 = 1.96$. □