Examen : RNC Corrigé

2 février 2021

En bleu les réponses attendues (en gros)
En vert des remarques complémentaires.

Exercice 1 (8 pts)

On dispose de données mesurant les caractéristiques suivantes sur des appartements des Yvelines
— S = superficie en m^2
— E = exposition au soleil (note entre 0 et 10)
— D = distance de Paris (en km)
— P = prix de vente sur le marché (en k€)

On désire prédire le prix de vente P à partir des trois premières caractéristiques S, E et D.

1) **(1 pt)** De quel type de problème s'agit-il ? Quel(s) modèles(s) et algorithme(s) connaissez-vous pour le résoudre ?
Il s'agit d'un problème de régression en apprentissage supervisé. On peut le résoudre grâce à une régression linéaire, par exemple (soit avec l'équation normale ou une descente de gradient).

2) **(1 pt)** Pour cette question on fait comme s'il n'y avait que 4 appartment formant les données (alors qu'il y en a naturellement beaucoup plus). On donne ci dessous le tableau des valeurs ayant été relevées. Les caractéristiques S, E, D ont été mises à l'échelle (scaling).

	apt n°1	apt n°2	apt n°3	apt n°4
S	3	1	2	3
E	2	5	4	7
D	4	2	2	10
P	300	250	280	325

Déterminez la matrice X des données ainsi que le vecteur cible y de sorte à pouvoir effectuer une regression linéaire (affine) en utilisant directement la formule de l'équation normale (on ne demande pas d'effectuer la regression).
Dans la matrice de conception X les colonnes correspondent aux caractéristiques utilisées pour la prédiction (ici S,E,D) et les lignes aux différentes données (ici les appartements). Pour pouvoir appliquer directement l'équation normale il convient d'ajouter une colonne de 1 pour le biais.

$$X = \begin{pmatrix} 1 & 3 & 2 & 4 \\ 1 & 1 & 5 & 2 \\ 1 & 2 & 4 & 2 \\ 1 & 3 & 7 & 10 \end{pmatrix} \text{ et } y = \begin{pmatrix} 300 \\ 250 \\ 280 \\ 325 \end{pmatrix}$$

3) **(1 pt)** On revient maintenant au cas général où la matrice X possède beaucoup plus de lignes. En utilisant les notations du cours, on obtient par l'équation normale le vecteur

$$a = \begin{pmatrix} 181 \\ 36 \\ 6.8 \\ -1.5 \end{pmatrix}$$

Donnez une interprétation du fait que le quatrième coefficient soit négatif.

Le 4ème coefficient, -1.5, correspond à la caractéristique D lorsqu'on fait le produit $X \cdot a$. Le fait qu'il soit négatif implique que la prédiction du prix P sera décroissante quand D augmente ; ce qui est conforme à l'intuition car les prix diminuent en général en s'éloignant de Paris.

4) **(1 pt)** A la suite de la question 3), on considère un nouvel appartement (à la même échelle) avec $S = 4, E = 5, D = 6$. Posez le calcul numérique permettant d'effectuer une prévision pour cet appartement (on ne demande pas d'effectuer le calcul).

$$\hat{y} = (1\ S\ E\ D) \cdot a^T$$

$$\hat{y} = 181 \times 1 + 36 \times 4 + 6.8 \times 5 - 1.5 \times 6$$

5) **(1 pt)** Finalement, il s'avère que ces prédictions ne sont pas très bonnes et on désire effectuer une regression quadratique. Transformez la matrice X des données en une matrice augmentée X_a afin de pouvoir effectuer une regression quadratique en appliquant directement l'équation normale. On ne demande pas de faire la regression mais uniquement de donner la matrice X_a ainsi que le vecteur cible y correspondant. Précisez en haut de chaque colonne de X_a à quoi elle correspond. Remplissez seulement les deux premières lignes de la matrice (à l'aide des deux premières données de la question 2).

Il faut ajouter des colonnes pour les termes de degré 2, le vecteur y lui reste inchangé.

$$X_a = \left(\begin{array}{cccccccccc} 1 & S & E & D & S^2 & E^2 & D^2 & S \cdot E & S \cdot D & E \cdot D \\ \hline 1 & 3 & 2 & 4 & 9 & 4 & 16 & 6 & 12 & 8 \\ 1 & 1 & 5 & 2 & 1 & 25 & 4 & 5 & 2 & 10 \\ 1 & \dots & & & & & & & & \end{array}\right) \text{ et } y = \begin{pmatrix} 300 \\ 250 \\ 280 \\ 325 \\ \dots \end{pmatrix}$$

6) On fait maintenant la courbe de l'erreur quadratique moyenne minimale obtenue sur les données pour une regression de degré 1,2,3,... jusqu'à 20.

a) **(1 pt)** Quelle allure globale pensez-vous que cette courbe doit avoir ?

Quand le degré de la régression augmente le modèle devient plus complexe, donc permet de mieux approcher les données ; on s'attend donc à ce que la courbe de l'erreur quadratique moyenne minimale sur les données soit décroissante en fonction du degré

b) **(1 pt)** Pourquoi n'est-il pas forcément judicieux d'utiliser la regression semblant la meilleure sur les données d'après la courbe du a) ?

La meilleure approximation sur les données d'apprentissage n'est pas forcément la meilleure sur de nouvelles données test ; si la complexité du modèle est trop grande on aura un phénomène de suradaptation (overfitting).

c) **(1 pt)** Supposons qu'aucune autre donnée ne soit disponible, nous disposons de 200 données en tout et pour tout. Comment procéder pour savoir quelle puissance utiliser afin d'espérer la meilleure prédiction possible sur des données nouvelles ?

On peut séparer 200 les données en deux parties : par exemple 2/3 des données seront dédiées à l'apprentissage et le 1/3 restant sera consacré à des test pour déterminer la qualité de prédiction du modèle sur des données inconnues. En faisant cela pour différents degrés on peut avoir une idée du modèle à choisir (celui qui donnera l'erreur de test la plus faible).

Exercice 2 (6 pts)

On considère les données suivantes pour un problème de classification (classes +1/-1) en dimension 2 :

$$x^1 = (2,1) \text{ et } y^1 = 1$$

$$x^2 = (4,3) \text{ et } y^2 = 1$$

$$x^3 = (0,1) \text{ et } y^3 = -1$$

$$x^4 = (1,2) \text{ et } y^4 = -1$$

1) **(5 pt)** Effectuez l'algorithme du Perceptron sur ces données afin de les séparer linéairement par une droite passant par l'origine (et non pas une droite affine, il est donc inutile ici d'augmenter la dimension des données et d'ajouter une colonne de 1), en respectant les consignes suivantes :

— commencer avec le vecteur initial de poids $w_0 = (0,0)$ (on rappelle que le signe de 0 est considéré comme +1).

— à chaque itération de l'algorithme, considérer le premier vecteur mal classifié (dans l'ordre 1,2,3,4 des données) pour mettre à jour w.

— au fur et à mesure, reportez les vecteurs obtenus dans un tableau comme celui ci-dessous, écrivez les classes prédites pour ce vecteur et entourez le vecteur mal classifié utilisé pour la mise à jour de w (recopiez et complétez le tableau sur votre copie).

ne pas oublier la valeur de y^i lors de la mise à jour

$$w \leftarrow w + y^i \cdot x^i$$

où (x^i, y^i) est le premier point mal classifié (valeur entourée dans le tableau)

i	x^i	y^i	$w = (0,0)$	$w = (0,-1)$	$w = (2,0)$	$w = (2,-1)$	$w = (1,-3)$	$w = (3,-2)$
1	$(2,1)$	1	1	⓵	1	1	⓵	1
2	$(4,3)$	1	1	-1	1	1	-1	1
3	$(0,1)$	-1	①	-1	①	-1	-1	-1
4	$(1,2)$	-1	1	-1	1	①	-1	-1

2) **(1 pt)** Tracez une figure avec les 4 points et la droite de séparation +1/-1 finale, en indiquant la classification de chaque côté de la droite, afin d'illustrer les résultats obtenus.

..

Exercice 3 (7 pts)

On reprend les mêmes données qu'à l'exercice 2 afin d'obtenir une classification linéaire suivant un vecteur w, mais cette fois-ci on ne va pas appliquer l'algorithme du perceptron mais une descente de gradient.

Pour faire cela on considère la fonction de coût

$$\ell(t,y) = \big(\max(0, 1 - yt)\big)^2$$

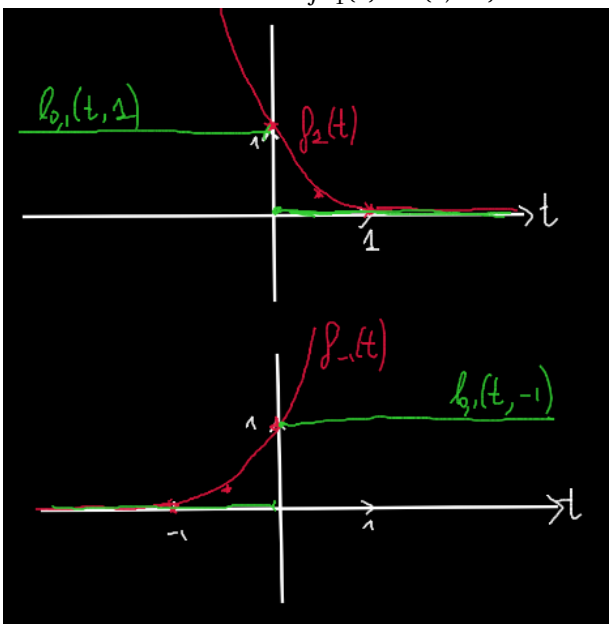
(attention à ne pas oublier le carré) c'est-à-dire que si le vecteur de poids est w, le coût appliqué sur la donnée d'indice i est

$$\ell^i = \ell\Big(\langle w, x^i\rangle, y^i\Big)$$

et le coût total que l'on cherche à minimiser est

$$L = \frac{1}{4}\sum_{i=1}^{4} \ell^i$$

1) **(1 pt)** Tracez la courbe (ou au moins l'allure de la courbe) de la fonction $f_1(t) = \ell(t,1)$ pour $t \in [-3,3]$, puis la courbe de la fonction $f_{-1}(t) = \ell(t,-1)$ sur le même intervalle.



2) **(1 pt)** Rappel : la fonction de coût $\ell_{0,1}$ appliquée sur les données pour le coût 0-1 est

$$\ell_{0,1}(t,y) = \mathbb{1}_{y\cdot t<0}$$

et le coût $L_{0,1}$ est alors

$$L_{0,1} = \frac{1}{4}\sum_{i=1}^{4} \ell_{0,1}\Big(\langle w, x^i\rangle, y^i\Big)$$

Tracez dans chaque cas ($y = \pm 1$) sur chacune des courbes précédentes (d'une autre couleur si possible !) la courbe du coût $t \mapsto \ell_{0,1}(t,y)$. En déduire le coût L est toujours supérieur au coût $L_{0,1}$. En quoi cela justifie-t'il son utilisation ?

D'après les courbes quel que soit t on a toujours

$$f_1(t) = \ell(t,1) \geq \ell_{0,1}(t,1)$$

et

$$f_{-1}(t) = \ell(t,-1) \geq \ell_{0,1}(t,-1)$$

donc dans tous les cas

$$\ell(t,y) \geq \ell_{0,1}(t,y)$$

On en déduit que

$$L = \frac{1}{4}\sum_{i=1}^{4} \ell\left(\langle w, x^i\rangle, y^i\right) \geq \frac{1}{4}\sum_{i=1}^{4} \ell_{0,1}\left(\langle w, x^i\rangle, y^i\right) = L_{0,1}$$

Il nous importe en fait de minimiser le nombre moyen d'erreurs de classification $L_{0,1}$ mais cette fonction a un gradient nul ; en minimisant L grâce à des techniques de descente de gradient on minimisera aussi $L_{0,1}$ en conséquence puisque $L \geq L_{0,1}$.

3) **(1 pt)** Montrez que les fonctions $f_y : t \mapsto \ell(t,y)$ sont dérivables sur $\mathbb{R}$ (distinguer suivant $y = \pm 1$ fixé), et explicitez leurs dérivées, en distinguant suivant les valeurs de t.

— Pour $y = 1$: Si $t \geq 1$ alors $f_1(t) = 0^2 = 0$ donc f_1 est dérivable sur $[1, +\infty[$ et $f'-1(t) = 0$.
Si $t \geq 1$ alors $f_1(t) = (1-t)^2$ donc f_1 est dérivable sur $]-\infty; 1]$ et $f'-1(t) = 2(t-1)$.
En $t = 1$ les deux expressions coincident et donnent la valeur 0 donc en fait f_1 est dérivable sur $\mathbb{R}$
On peut éventuellement écrire (ce n'est pas demandé)

$$f_1'(t) = \min(0, 2(t-1))$$

— une analyse similaire pour f_{-1} donne
Si $t \geq -1$ alors $f'_{-1}(t) = 2(1+t)$
Si $t \leq -1$ alors $f'_{-1}(t) = 0$
On a alors (pareil on pouvait se contenter des deux cas ci-dessus)

$$f'_{-1}(t) = \max(0, 2(t+1))$$

4) **(2 pt)** En déduire le gradient de la fonction $w \mapsto \ell^i(w)$ (x^i et y^i fixés, distinguer les cas $y = \pm 1$ et suivant la valeur de $\left(\langle w, x^i\rangle, y^i\right)$.

Posons

$$F(w) = \ell^i(w) = f_y(\langle w, x^i\rangle) = f_y(w_1 x_1^i + w_2 x_2^i)$$

pour $y = \pm 1$ et $w \in \mathbb{R}^2$.

— pour $y = 1$ et $j \in 1,2$ alors

$$\frac{\partial F}{\partial x_j} = w_j \times f_1'(\langle w, x^i\rangle)$$

soit

$$\frac{\partial F}{\partial x_j} = x_j \times \min(0, 2(\langle w, x^i\rangle - 1))$$

Et donc

$$\nabla_w F = (\min(0, 2(\langle w, x^i\rangle - 1)) \cdot x$$

(on peut aussi distinguer les cas $\langle w, x^i\rangle > 1$ ou ≤ 1 plutôt que d'écrire avec un min)

— pour $y = -1$ et $j \in 1,2$ alors de façon semblable

$$\nabla_w F = (\max(0, 2(\langle w, x^i\rangle + 1)) \cdot x$$

5) **(2 pts)** Effectuez une descente de gradient stochastique sur les données : initialisez le vecteur des poids à $w = (0,0)$ et utilisez le pas de gradient $\alpha = 1$. On peut s'arrêter après avoir fait une mise à jour de gradient stochastique pour chacune des quatre données (x^i, y^i).

(remarque : le pas de gradient $\alpha = 1$ est probablement trop grand pour mener à une convergence et est uniquement utilisé pour simplifier les calculs).

1. pour $w = (0,0)$, on fait un pas de gradient sur la donnée $x^1 = (2,1)$ et $y^1 = 1$.
   On a $\langle w, x^1 \rangle = 0$ donc
   $$\nabla_w F(x^1) = (\min(0,-2)) \cdot x^1 = -2x^1$$
   On met à jour w par
   $$w \leftarrow w - \alpha \nabla_w F$$
   soit
   $$w \leftarrow w - 1 \times (-2x^1)$$
   ce qui donne un nouveau w
   $$w = (4,2)$$
2. pour $w = (4,2)$, on fait un pas de gradient sur la donnée $x^2 = (4,3)$ et $y^2 = 1$.
   On a $\langle w, x^2 \rangle = 16 + 6 = 22$ donc
   $$\nabla_w F(x^2) = (\min(0, 2 * (22-1))) \cdot x^2 = 0$$
   On met à jour w par
   $$w \leftarrow w - \alpha \nabla_w F$$
   soit
   $$w \leftarrow w$$
   donc w reste inchangé
   $$w = (4,2)$$
3. pour $w = (4,2)$, on fait un pas de gradient sur la donnée $x^3 = (0,1)$ et $y^1 = -1$.
   On a $\langle w, x^3 \rangle = 2$ donc
   $$\nabla_w F(x^3) = (\max(0,6)) \cdot x^3 = 6x^3$$
   On met à jour w par
   $$w \leftarrow w - \alpha \nabla_w F$$
   soit
   $$w \leftarrow w - 6x^3)$$
   ce qui donne un nouveau w
   $$w = (4,-4)$$
4. pour $w = (4,-4)$, on fait un pas de gradient sur la donnée $x^4 = (1,2)$ et $y^4 = -1$.
   On a $\langle w, x^1 \rangle = -4$ donc
   $$\nabla_w F(x^4) = (\max(0,-6)) \cdot x^1 = 0$$
   On met à jour w par
   $$w \leftarrow w - \alpha \nabla_w F$$
   donc $w = (4,-4)$ reste inchangé.