

Machine Learning, TD un peu théorique

Master Informatique, parcours I2A

Jean-François COUCHOT

couchot@arobase.femto-st.fr

28 novembre 2024

1 Apprentissages supervisés

Quelques applications directes d'abord.

1.1 Classification bayésienne naïve

Nom	Âge	Revenu	Salarié-e	Trouble cardiaque
Adams	jeune	élevé	Non	Non
Baker	jeune	élevé	Non	Non
Cook	mûr	élevé	Non	Oui
Dodd	sénior	moyen	Non	Oui
Engel	sénior	faible	Oui	Oui
Flynn	sénior	faible	Oui	Non
Grady	mûr	faible	Oui	Oui
Hayes	jeune	moyen	Non	Non
Irons	jeune	faible	Oui	Oui
Jones	sénior	moyen	Oui	Oui
Knapp	jeune	moyen	Oui	Oui
Lord	mûr	moyen	Non	Oui

FIGURE 1 – Données concernant des patients fictifs d'une clinique étudiant les troubles cardiaques.

Exercice 1.1 (Sur des données discrètes). Dans cet exercice, on considère le jeu de données de la FIGURE 1. Cette table contient 12 enregistrements précisant les attributs tels que l'âge, le niveau de revenu, le fait d'être salarié ou non et si la personne a déjà souffert d'un trouble cardiaque.

1. Se présente Gladys, jeune, avec un revenu élevé et qui est salariée. Mettre en place complètement une démarche d'apprentissage bayésien naïf pour estimer si cette personne a davantage de chances de ne pas souffrir d'un trouble cardiaque que d'en souffrir.

2. Discuter de la pertinence de l'estimation précédente en regard des probabilités obtenues en dernière étape.

	date	precipitation	temp_max	temp_min	wind	weather
72	2012-03-13	9.4	5.6	0.6	5.3	snow
170	2012-06-19	1.0	19.4	10.0	3.0	rain
186	2012-07-05	0.0	24.4	10.6	3.1	drizzle
382	2013-01-17	0.0	3.9	-2.8	1.0	drizzle
787	2014-02-26	0.0	13.9	5.6	2.5	sun
862	2014-05-12	0.0	24.4	9.4	2.7	sun
1063	2014-11-29	3.6	4.4	-4.3	5.3	snow
1123	2015-01-28	0.0	12.2	5.0	1.8	fog
1376	2015-10-08	0.0	18.9	13.3	1.1	fog
1440	2015-12-11	0.3	9.4	4.4	2.8	rain

FIGURE 2 – Données météorologiques concernant la ville de Seattle

Exercice 1.2 (Sur des données continues). On considère dans cette exercice des données météorologique de la ville de seattle¹, dont un extrait est donné à la figure 2. Ce tableau donne, pour chaque date entre 2012 et 2015, la quantité de précipitations (en mm), les températures maximum et minimum (en °C), la vitesse du vent (en m/s) et la description du temps qu'il fait. L'objectif est de mettre en place une approche bayésienne naïve gaussienne.

1. Quels calculs devrait-on réaliser pour chaque valeur du temps qu'il fait ?

1. <https://github.com/vega/vega/blob/main/docs/data/seattle-weather.csv>

weather	precipitation		temp_max		temp_min		wind	
	mean	std	mean	std	mean	std	mean	std
drizzle	0.0	0.1	15.9	8.8	7.1	6.2	2.4	0.9
fog	0.0	0.2	16.8	6.8	8.0	5.0	2.5	1.2
rain	6.6	8.6	13.5	5.0	7.6	4.0	3.7	1.6
snow	8.6	7.0	5.6	3.1	0.1	2.2	4.4	1.4
sun	0.0	0.1	19.9	7.7	9.3	5.5	3.0	1.2

FIGURE 3 – Statistiques générales sur le dataset météorologique de Seattle.

#	Age	Revenu	Achete
1	Jeune	Bas	Non
2	Adulte	Moyen	Oui
3	Senior	Haut	Oui
4	Jeune	Moyen	Non
5	Adulte	Bas	Non
6	Senior	Moyen	Oui

TABLE 1 – Achat en fonction du revenu et de l'âge

2. On considère comme donné le tableau donné à la figure 3. De plus, les statistiques suivantes ont été calculées :

<i>weather</i>	<i>rain</i>	<i>sun</i>	<i>fog</i>	<i>drizzle</i>	<i>snow</i>
fréquence en %	43.9	43.8	6.91	3.63	1.78

Sachant qu'il n'y a aucune précipitation, qu'il n'y a pas de vent, que les températures minimum et maximum observées sont respectivement de -2°C et 4°C , quel est le temps le plus probable ? Détailler toute la démarche.

<i>weather</i> <i>y_i</i>	<i>precipitation</i>			<i>temp_max</i>			<i>temp_min</i>			<i>wind</i>		<i>f_{ij}(x_j)</i>
	<i>mean</i>	<i>std</i>		<i>mean</i>	<i>std</i>		<i>mean</i>	<i>std</i>		<i>mean</i>	<i>std</i>	
<i>drizzle</i>	0.0	0.1		15.9	8.8		7.1	6.2		2.4	0.9	
<i>fog</i>	0.0	0.2		16.8	6.8		8.0	5.0		2.5	1.2	
<i>rain</i>	6.6	8.6		13.5	5.0		7.6	4.0		3.7	1.6	
<i>snow</i>	8.6	7.0		5.6	3.1		0.1	2.2		4.4	1.4	
<i>sun</i>	0.0	0.1		19.9	7.7		9.3	5.5		3.0	1.2	

1.2 Arbres de décision

Exercice 1.3. Construire et exploiter un arbre.

On considère le jeu de données présenté au tableau 1. Dans celui-ci, pour chaque personne dont l'identifiant est *ID*, on sait dans quelle tranche d'âge elle est, quel est sa classe de revenu et si elle va effectuer un achat immobilier.

1. Montrer que l'entropie du dataset complet T sans considérer les attribut discriminants est $\log_2(2) = 1$.

2. Pour l'attribut Age :

(a) Montrer que $S_{age}(\{Jeune\}) = \{1, 4\}$ et $S_{age}(\{Adulte, Senior\}) = \{2, 3, 5, 6\}$.

(b) Montrer que $H(S_{age}(\{Jeune\})) = 0$ et $H(S_{age}(\{Adulte, Senior\})) = -(\frac{3}{4} \log_2(\frac{3}{4}) + \frac{1}{4} \log_2(\frac{1}{4})) \approx 0.81$

(c) Montrer que $H(T|(age, Jeune)) = \frac{2}{3} \times H(S_{age}(\{Adulte, Senior\})) \approx 0.54$.

(d) En déduire que $IG(T, (age, Jeune)) \approx 0.46$.

(e) Montrer que $S_{age}(\{Adulte\}) = \{2, 5\}$ et $S_{age}(\{jeune, Senior\}) = \{1, 3, 4, 6\}$.

(f) Montrer que $H(S_{age}(\{Adulte\})) = \log_2(2)$ et $H(S_{age}(\{Jeune, Senior\})) = \log_2(2)$

(g) Montrer que $H(T|(age, Adulte)) = \log_2(2)$.

(h) En déduire que $IG(T, (age, Adulte)) = 0$

(i) Montrer que $S_{age}(\{Senior\}) = \{3, 6\}$ et $S_{age}(\{jeune, Adulte\}) = \{1, 2, 4, 5\}$.

(j) Montrer que $H(S_{age}(\{Senior\})) = 0$ et $H(S_{age}(\{Jeune, Adulte\})) = -(\frac{3}{4} \log_2(\frac{3}{4}) + \frac{1}{4} \log_2(\frac{1}{4})) \approx 0.811$

(k) Montrer que $H(T|(age, Senior)) = 2/3 \times H(S_{age}(\{Jeune, Adulte\}))$.

(l) En déduire que $IG(T, (age, Senior)) = 0.46$

3. Pour l'attribut Revenu :

(a) Montrer que $S_{Revenu}(\{Bas\}) = \{1, 5\}$ et $S_{Revenu}(\{Haut, moyen\}) = \{2, 3, 4, 6\}$.

(b) Montrer que $H(S_{Revenu}(\{Bas\})) = 0$ et $H(S_{Revenu}(\{Haut, Moyen\})) = -(\frac{3}{4} \log_2(\frac{3}{4}) + \frac{1}{4} \log_2(\frac{1}{4})) \approx 0.81$

(c) Montrer que $H(T|(Revenu, Moyen)) = 2/3 \times H(S_{Revenu}(\{Haut, Moyen\})) \approx 0.54$.

(d) En déduire que $IG(T, (Revenu, Bas)) \approx 0.46$

(e) Montrer que $S_{Revenu}(\{Haut\}) = \{3\}$ et $S_{Revenu}(\{Bas, moyen\}) = \{1, 2, 4, 5, 6\}$.

(f) Montrer que $H(S_{Revenu}(\{Haut\})) = 0$ et $H(S_{Revenu}(\{Bas, Moyen\})) = -(\frac{2}{5} \log_2(\frac{2}{5}) + \frac{3}{5} \log_2(\frac{3}{5})) \approx 0.97$

(g) Montrer que $H(T|(Revenu, Haut)) = 5/6 \times H(S_{Revenu}(\{Bas, Moyen\})) \approx 0.81$.

(h) En déduire que $IG(T, (Revenu, Haut)) \approx 0.19$

(i) Montrer que $S_{Revenu}(\{Moyen\}) = \{2, 4, 6\}$ et $S_{Revenu}(\{Haut, Bas\}) = \{1, 3, 5\}$.

(j) Montrer que $H(S_{Revenu}(\{Moyen\})) = -(\frac{1}{3} \log_2(\frac{1}{3}) + \frac{2}{3} \log_2(\frac{2}{3})) \approx 0.92$ et $H(S_{Revenu}(\{Bas, Haut\})) = H(S_{Revenu}(\{Moyen\}))$

(k) Montrer que $H(T|(Revenu, Moyen)) = H(S_{Revenu}(\{Moyen\})) \approx 0.92$.

(l) En déduire que $IG(T, (Revenu, Moyen)) \approx 0.08$

4. En déduire que les découpages qui maximisent le gain d'information sont lorsque discrétise selon que l'âge est jeune, ou bien selon que l'âge soit sénior ou bien que le revenu soit Bas.

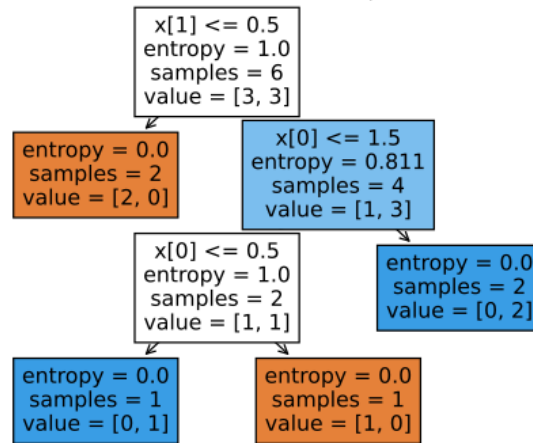


FIGURE 4 – Arbre de décision associé au dataset de la table 1

5. On a fait construire l'arbre de décision correspondant et celui-ci est représenté à la figure 4. Les attributs sont $X=[\text{'Age'}, \text{'Revenu'}]$. Pour l'âge, les valeurs sont classées dans l'ordre [Adulte, Jeune, Senior]. Pour le revenu, les valeurs sont classées dans l'ordre [Bas, Haut, moyen]. Le vecteur `value` représente le nombre Non et de Oui, dans cet ordre. Expliquer toutes les données affichées dans le noeud le plus en haut. Cela est-il cohérent avec la question précédente ?

6. Expliquer le sens des deux cellules en orange aussi.

7. Se présente un jeune aux hauts revenus. Comment serait-il classé par cet arbre ?

1.3 Réseau de neurones

Exercice 1.4. Calcul de la sortie d'un réseau simple

On considère un réseau de neurones dont l'architecture est la suivante : 2 entrée, 1 couche cachée avec 3 neurones et 1 couche de sortie. Les poids, biais et fonction d'activation sont les suivants :

- Couche cachée :
 - Neurone 1 : poids $w_{11} = 0.2$, $w_{12} = 0.3$, biais $b_1 = 0.1$;
 - Neurone 2 : poids $w_{21} = 0.4$, $w_{22} = 0.5$, biais $b_2 = 0.2$;
 - Neurone 3 : poids $w_{31} = 0.6$, $w_{32} = 0.7$, biais $b_3 = 0.3$;
 - Fonction d'activation : ReLU à chaque fois.
- Couche de sortie :
 - Poids $w_{41} = 0.8$, $w_{42} = 0.9$, $w_{43} = 1.0$, biais $b_4 = 0.4$;

— Fonction d'activation : sigmoïde.

1. Calculer la valeur de sortie de ce réseau de neurones lorsque l'entrée est la paire $x = (0.5, 0.2)$.

—
—
—
—
—

2. Développer une fonction qui effectue ce calcul `forward_propagation(X, W1, b1, W2, b2, activation1, activation2)` telle que :

- x est le vecteur des entrées ;
- $W1, W2$ sont les matrices des poids de la couche d'entrée à la couche cachée et de la couche cachée à la couche de sortie respectivement
- $b1$ et $b2$ sont les vecteurs des biais de la couche cachée et de la couche de sortie respectivement ;
- `activation1` et `activation2` sont les fonction d'activation (sigmoid, relu, etc.) de la couche cachée et de la couche de sortie respectivement.

3. Vérifier qu'avec les mêmes paramètre qu'à la question 1, cette fonction retourne la même valeur.

Exercice 1.5. Paramètres d'un réseau de neurones Soit un réseau de neurones dense avec les caractéristiques suivantes :

- Couche d'entrée : 10 neurones ;
- 2 couches cachées :
 - Première couche cachée : 20 neurones ;
 - Deuxième couche cachée : 15 neurones ;
- Couche de sortie : 5 neurones ;
- Fonctions d'activation : ReLU pour toutes les couches cachées et de sortie ;
- Sans biais (pour simplifier le calcul).

1. Combien de paramètres ce réseau possède-t-il ?

—
—
—
—

2. Comment ce nombre de paramètres évolue-t-il si on ajoute une troisième couche cachée de 10 neurones ?

—
—
—

3. Quel est l'impact d'ajouter des biais sur le nombre total de paramètres ?

—
—
—

4. Pourquoi est-il important de contrôler le nombre de paramètres dans un réseau de neurones ?

2 Apprentissages non supervisés

Exercice 2.1. *Compréhension de la fonction de coût de l'algorithme de k -moyenne*

1. *Définissez formellement la fonction de coût utilisée dans l'algorithme k -moyenne. Quelles sont les variables en jeu ?*
2. *Interprétez intuitivement cette fonction de coût. Quel est son objectif ?*
3. *Montrer que l'algorithme converge vers un optimum (local) (difficile).*
4. *Pourquoi l'initialisation des centroïdes est-elle cruciale dans K -means ?*
5. *Comment peut-on améliorer les chances de trouver le minimum global de la fonction de coût ?*