TD 3 – Naïve Bayes et k-Plus Proche Voisin

Exercice 1 :

1- Rappeler les principes des méthodes : k-PPV et Bayésien Naïf.
2- Quelles sont leurs ressemblances et leurs différences ?

Exercice 2 : Naive Bayes

Training dataset $\mathcal{D}$

Class	X_1	X_2
+	a	1.0
+	b	1.2
+	a	3.0
-	b	4.4
-	b	4.5

1- Estimer la probabilité d'appartenir à la classe '+', puis celle d'appartenir à la classe '-'.
2- Estimer $P(X_1 = a \mid +)$, $P(X_1 = a \mid -)$, $P(X_1 = b \mid +)$, $P(X_1 = b \mid -)$
3- Estimer $P(X_2 = 1.0 \mid +)$, $P(X_2 = 1.0 \mid -)$, $P(X_2 = 4.5 \mid +)$, $P(X_2 = 4.5 \mid -)$
    a. En discrétisant ;
    b. En faisant l'hypothèse d'une distribution normale. Voir formule en annexes.

Exercice 3 : (exemple question 4 du TD1) -- Naive Bayes

	x_1	x_2	x_3	Classe
1	0	V	N	A
2	1	V	I	A
3	0	F	O	B
4	1	V	N	A
5	1	V	O	A
6	1	F	N	A
7	0	F	O	B
8	0	V	I	A
9	0	F	N	B
10	1	V	I	B
11	1	F	O	A
12	1	F	I	A
13	0	V	O	B

On découpe l'ensemble en 2 : D1 et D2. **D1 contient les 9 premiers objets**, et D2 contient les **3 derniers**.

1- Construire le modèle bayésien naïf en utilisant D1 et en appliquant :
    a. La formule originale d'estimation des probabilités
    b. La formule de Laplace
2- L'ensemble D2 va être utilisé pour tester le modèle bayésien naïf. Déterminer la classe des 3 objets de D2, en appliquant :
    a. La formule originale d'estimation des probabilités
    b. La formule de Laplace
3- Pour chaque formule :
    a. Donner la matrice de confusion sur D2;
    b. Calculer le taux d'erreur apparente avec D2;
    c. Calculer la précision pour chaque classe;
    d. Calculer le rappel pour chaque classe;
4- Les résultats sont-ils identiques ?

Exercice 4 : k-PPV

On souhaite maintenant appliquer la méthode des k plus proches voisins avec k=1, puis k=3, avec les données de l'exercice 3).

1- Déterminer la classe des 3 objets de D2 en utilisant comme mesure, la dissimilarité par les différences entre objets, et comme mode de décision :
   a. le vote majoritaire
   b. le vote majoritaire pondéré par l'inverse du carré de la distance.
2- Reprendre la question 3.a) de l'exercice 3) ci-dessus, avec :
   a. Le vote majoritaire
   b. Le vote majoritaire pondéré.
3- Comparer les résultats avec ceux de l'exercice 3).

Exercices supplémentaires

Exercice S1 :

RID	age	income	student	credit	C_i: buy
1	youth	high	no	fair	C_2: no
2	youth	high	no	excellent	C_2: no
3	middle-aged	high	no	fair	C_1: yes
4	senior	medium	no	fair	C_1: yes
5	senior	low	yes	fair	C_1: yes
6	senior	low	yes	excellent	C_2: no
7	middle-aged	low	yes	excellent	C_1: yes
8	youth	medium	no	fair	C_2: no
9	youth	low	yes	fair	C_1: yes
10	senior	medium	yes	fair	C_1: yes
11	youth	medium	yes	excellent	C_1: yes
12	middle-aged	medium	no	excellent	C_1: yes
13	middle-aged	high	yes	fair	C_1: yes
14	senior	medium	no	excellent	C_2: no

Etant donné l'échantillon d'apprentissage ci-dessus, quelle est la classe de l'objet X suivant ? (utiliser NB, k-PPV et arbre de décision)

$$\mathbf{X} = (age = youth, income = medium, student = yes, credit = fair)$$

Exercice S2 :

sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Déterminer la classe de l'objet suivant :

sex	height (feet)	weight (lbs)	foot size(inches)
?	6	130	8

Annexes :

Naive Bayes : Estimation des probabilités conditionnelles

Ai : une valeur de l'attribut A

Nic : Nombre d'objets ayant la valeur Ai dans la classe c

Nc : Nombre d'objets de la classe c

k : nombre de valeurs de l'attribut A

p : probabilité apriori

Original: $P(A_i \mid C) = \frac{N_{ic}}{N_c}$

Laplace: $P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + k}$

m-estimate: $P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$

Given a training dataset $\mathcal{D}$ of N labeled examples (assuming complete data)

1. Estimate $P(c_j)$ for each class c_j

   $\hat{P}(c_j) = \frac{N_j}{N}$ N_j - the number of examples of the class c_j

2. Estimate $P(X_i = x_k \mid c_j)$ for each value x_k of the attribute X_i and for each class c_j
   - X_i discrete

     $\hat{P}(X_i = x_k \mid c_j) = \frac{N_{ijk}}{N_j}$ N_{ijk} - number of examples of the class c_j having the value x_k for the attribute X_i

   - X_i continuous

     Two options {
     - The attribute is discretized and then treats as a discrete attribute
     - A Normal distribution is usually assumed

     $P(X_i = x_k \mid c_j) = g(x_k; \mu_{ij}, \sigma_{ij})$ onde $g(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

     The mean μ_{ij} e the standard deviation σ_{ij} are estimated from $\mathcal{D}$

---

2. Estimate $P(X_i = x_k \mid c_j)$ for a value of the attribute X_i and for each class c_j
   - A Normal distribution is usually assumed

   $P(X_i = x_k \mid c_j) = g(x_k; \mu_{ij}, \sigma_{ij}) \Rightarrow g(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

   $X_i \mid c_j \sim N(\mu_{ij}, \sigma^2_{ij})$ - the mean μ_{ij} e the standard deviation σ_{ij} are estimated from $\mathcal{D}$

For a variable $X \sim N(74, 36)$, the probability of observing the value 66 is given by:

$f(x) = g(66; 74, 6) = 0.0273$

---

k-PPV: Proximité (Similarité, Dissimilarité), Distances

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Distance de Minkowski :

$$\boxed{dist = \left(\sum_{k=1}^{n} \lvert p_k - q_k \rvert^r \right)^{\frac{1}{r}}}$$

où r est un paramètre, n est le nombre de dimensions (attributs) et p_k et q_k sont, respectivement, les $k^{èmes}$ attributs (composants) des objets p et q.

r = 1 : City block (Manhattan, taxicab, L_1 norm) distance. Aussi appelée distance de Hamming pour des vecteurs binaires.

r = 2 : distance euclidienne

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities
  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1
- Simple Matching and Jaccard Coefficients
  SMC = number of matches / number of attributes
  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values
  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$