

Abbes Laghror University of Khenchela
Make up Exam in Machine Learning

Exercice 1 (10points):

1. Donnez deux métriques utilisées pour évaluer les modèles d'apprentissage, et expliquez comment elles sont calculées.
2. Qu'est-ce qu'un classifieur naïf ? Donnez une technique d'apprentissage qui utilise ce concept.
3. Quelle est la différence entre un arbre de décision et le k-NN ?
4. Comment peut-on accélérer la convergence de l'algorithme K-means ?

- Accuracy= Nombre de prédictions correctes/ Nombre total de prédictions

Matrice de confusion :

	Prédit : Positif	Prédit : Négatif
Réel : Positif	Vrai Positif (VP)	Faux Négatif (FN)
Réel : Négatif	Faux Positif (FP)	Vrai Négatif (VN)

- Un classifieur naïf est un modèle de classification qui suppose que les attributs (caractéristiques) sont indépendants conditionnellement à la classe. Cette hypothèse simplifie le calcul des probabilités.

Technique d'apprentissage : le Naïve Bayes, basé sur le théorème de Bayes avec l'hypothèse d'indépendance conditionnelle.

- Arbre de décision : C'est un modèle d'apprentissage supervisé basé sur des règles logiques (sous forme d'arborescence). Il apprend un modèle explicite pendant la phase d'entraînement. Rapide à l'inférence.
K-NN (k-Nearest Neighbors) : C'est un algorithme paresseux (lazy learning), il ne construit pas de modèle pendant l'apprentissage. Pour classer un nouvel exemple, il cherche les k voisins les plus proches dans l'ensemble d'entraînement. Plus coûteux à l'inférence.
- On peut accélérer la convergence de K-means en utilisant K-means++ pour initialiser les centroïdes : Cela choisit de meilleurs points de départ, réduisant les itérations nécessaires.

Exercice 2 (10 points):

On veut classer l'ensemble des points suivants en trois classes : A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).

On suppose initialement que les points A1, B1 et C1 sont choisis comme centres. Utiliser l'algorithme K-means pour déterminer

1. Les trois centres après la première itération en utilisant la distance euclidienne pour déterminer les nouveaux centroïdes.
2. Quelle est la meilleure méthode pour choisir les centres initiaux afin d'assurer une convergence rapide de l'algorithme k-means ?
3. Quand peut-on dire que l'algorithme k-means a convergé ?

Les trois centres calculés après la première itération :

Centres initiaux : A1(2,10), B1(5,8), C1(1,2).

Étapes :

Calcul des distances euclidiennes entre chaque point et les centres initiaux :

La distance euclidienne entre deux points (x_1, y_1) et (x_2, y_2) est donnée par :

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Point	Distance à A1(2,10)	Distance à B1(5,8)	Distance à C1(1,2)	Cluster assigné
A1	0	3.61	8.06	A1
A2	5.0	4.24	3.16	C1
A3	8.49	5.0	7.28	B1
B1	3.61	0	7.21	B1
B2	7.07	3.61	6.08	B1
B3	7.21	4.12	5.39	B1
C1	8.06	7.21	0	C1
C2	2.24	1.41	7.07	B1

Affectation des points aux clusters :

Cluster A1: A1(2,10)

Cluster B1: A3(8,4), B1(5,8), B2(7,5), B3(6,4), C2(4,9)

Cluster C1: A2(2,5), C1(1,2) **(0.25 pts)**

Calcul des nouveaux centroïdes :

Nouveau centroïde A1 : Moyenne des points dans le cluster A1 $\rightarrow (2, 10)$.

Nouveau centroïde B1 : Moyenne des points dans le cluster B1 $\rightarrow$

$$\left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5} \right) = (6.0, 6.0)$$

Nouveau centroïde C1 : Moyenne des points dans le cluster C1 $\rightarrow$

$$\left(\frac{2+1}{2}, \frac{5+2}{2} \right) = (1.5, 3.5)$$

La meilleure méthode pour choisir les centres initiaux afin d'assurer une convergence rapide de l'algorithme k-means est la méthode k-means++. Cette méthode sélectionne les centroïdes initiaux de manière intelligente en maximisant la distance entre les centres choisis.

L'algorithme **k-means** est considéré comme ayant convergé lorsque l'une des deux conditions suivantes est remplie :

- **Stabilisation des centroïdes** :
Les centroïdes des clusters ne changent plus entre deux itérations successives. Cela signifie que la position des centres des clusters est devenue stable, et l'algorithme ne peut plus améliorer la partition des données. **(1 pts)**
- **Stabilisation des clusters** :
Les points ne changent plus de cluster entre deux itérations successives. Cela signifie que chaque point est assigné au cluster le plus proche, et cette affectation ne varie plus.