Université J.F. CHAMPOLLION Albi

Méthode des plus proches voisins *kNN*

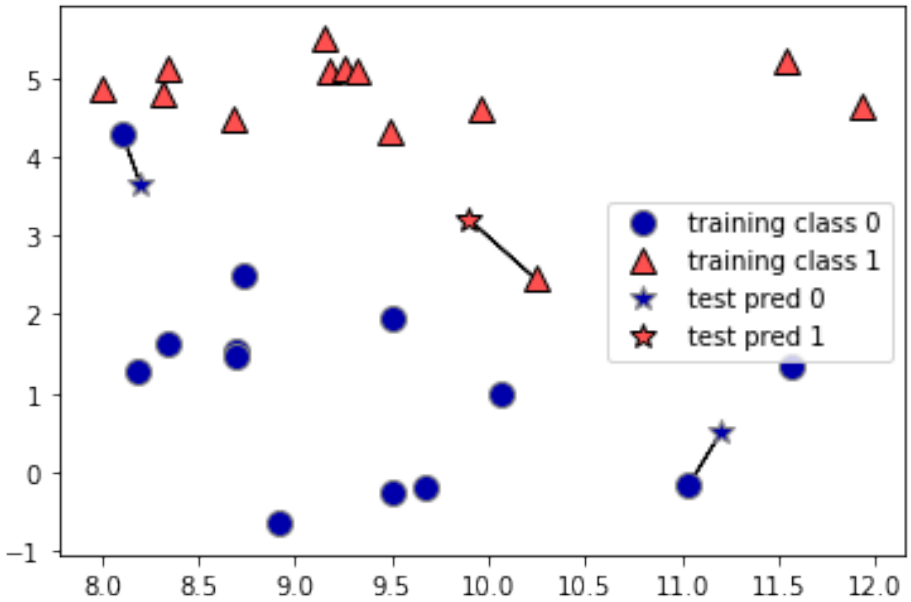
Licence d'informatique S6

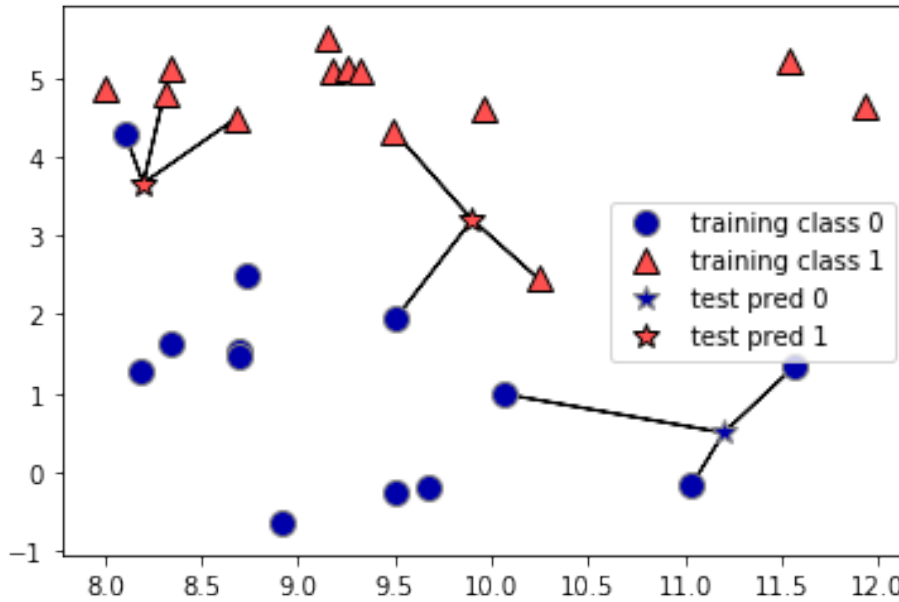
Thierry Montaut

Méthode des plus proches voisins

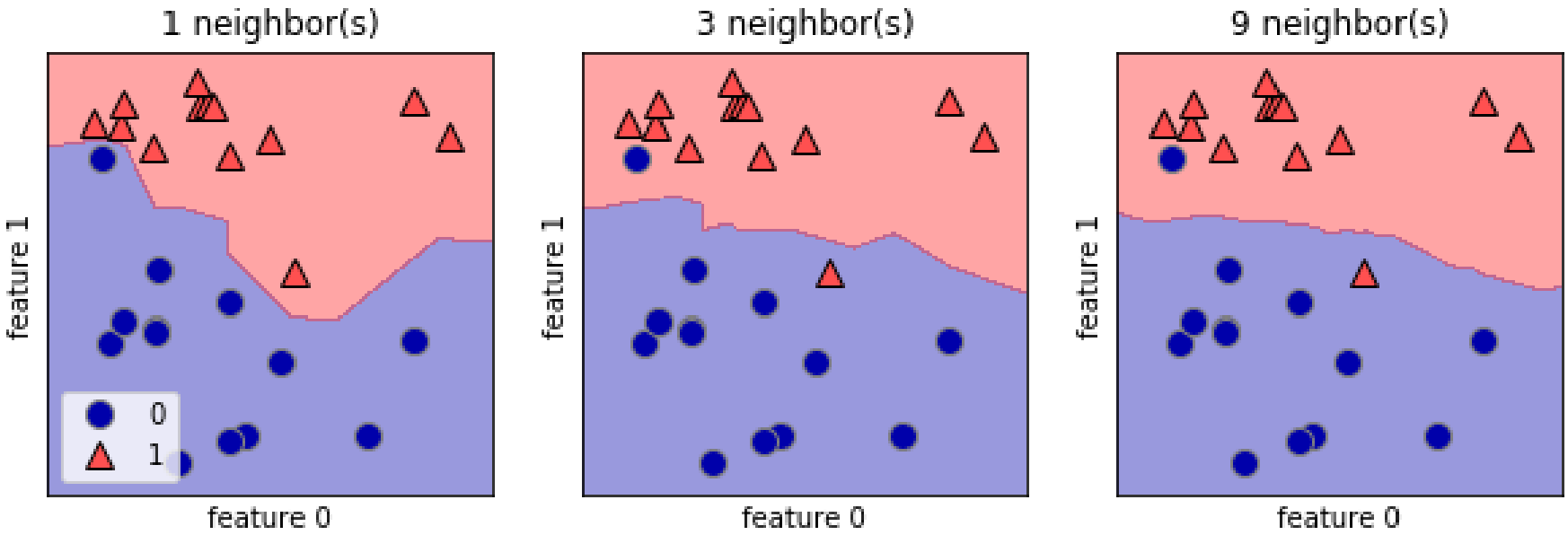
- L'algorithme k-NN des k plus proches voisins est le plus simple des apprentissages automatiques. Il n'y a aucun paramètre à apprendre. L'apprentissage se résume à choisir le jeu de données d'apprentissage A et à choisir une valeur de $k \geq 1$.
- Prédiction : étant donné une nouvelle donnée $x \in X$, on calcule les distances de x à tous les éléments de A et on sélectionne les k plus proches. On attribue à x la valeur majoritaire parmi ses k plus proches voisins.
- Nous mettrons en oeuvre ce modèle en TP à l'aide de la librairie scikit-learn de Python pour la classification des iris et le naufrage du Titanic.

k-NN pour $k = 1$ et $k = 3$.



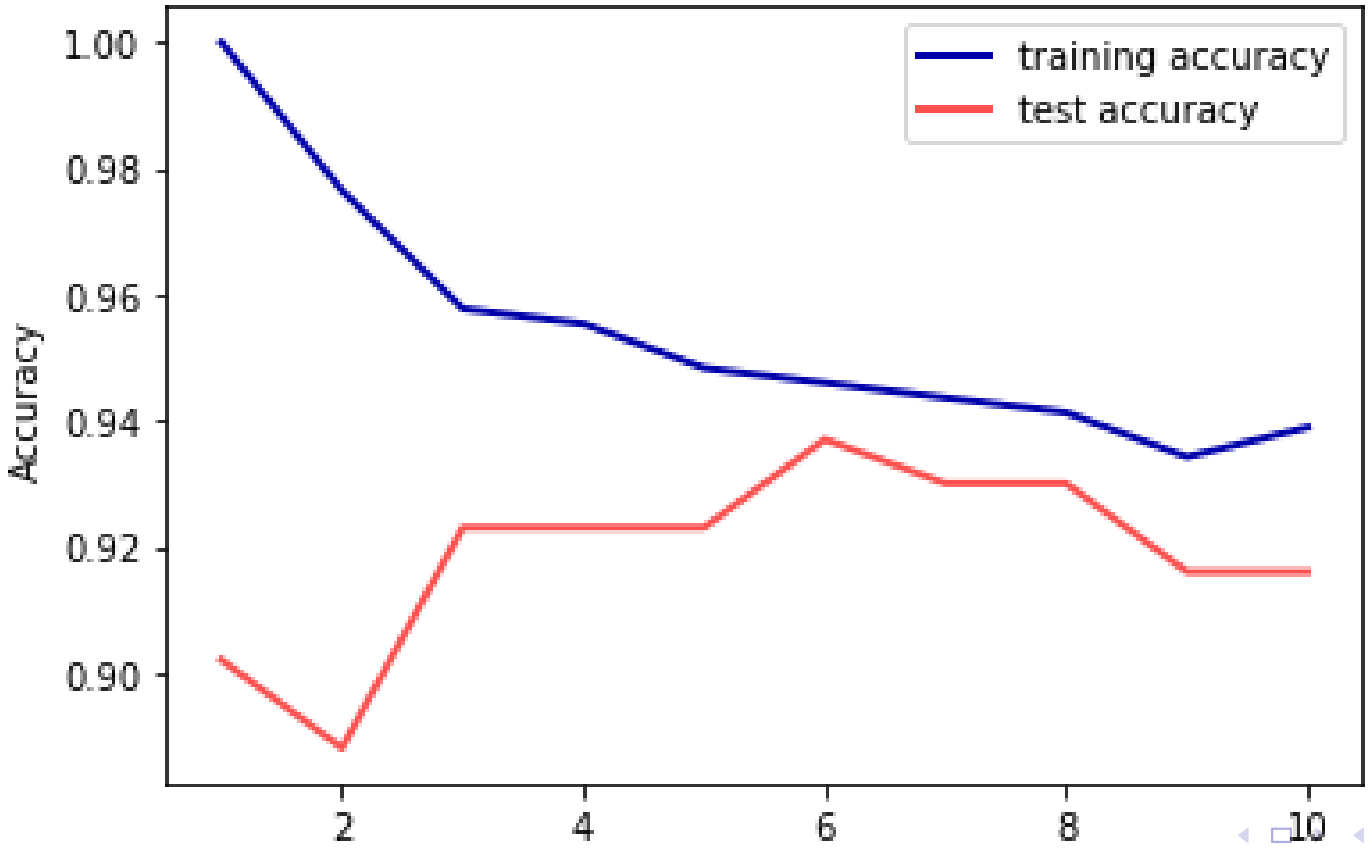


Frontières de décision



Paramètres

1. Le modèle k-NN possède deux paramètres importants : le nombre k de voisins et la distance choisie.
2. Il est bon de tester cet algorithme avec différentes valeurs de k de manière à choisir, pour un jeu de données fixé, la valeur de k qui optimise la prédiction et limite le sur-apprentissage.



Distances

1. Si les données sont toutes numériques, on utilise comme distance, une des trois distances usuelles de $\mathbb{R}^n$.
2. Nous n'utiliserons cette année que la distance euclidienne, mais vous découvrirez en master des distances dans $\mathbb{R}^n$ pouvant donner de meilleures performances.

Distance de Jaccard entre ensembles

Si les données sont des ensembles, on peut utiliser des variantes de la distance de Jaccard entre deux ensembles finis.

1. Etant donnés deux ensembles A et B, cette distance est définie comme le taux d'éléments communs aux deux ensembles

$$d(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

2. La distance entre deux mails peut être le taux de mots significatifs communs
3. La distance entre deux spectateurs est le taux de mêmes films visionnés. De même pour deux auditeurs avec le taux de mêmes musiques écoutées.

Encodage des données qualitatives

Afin de définir des distances, les données qualitatives nécessitent un encodage préalable.

1. Encodage cyclique pour la distance entre deux jours de la semaine ou deux mois de l'année.
2. Encodage spécifiques pour les catégories socio-professionnelles, le thème d'un article de journal etc.

Variantes : Les ε-voisins

- Plutôt que de considérer les k plus proches voisins, on peut considérer tous les voisins à une distance inférieure à ε donné.
- On évite le calcul de minimum et on utilise mieux les données dans les zones denses.

Retenir

- La méthode des plus proches voisins a un apprentissage paresseux compensé par une complexité algorithmique élevée pour les prédictions
- La qualité de la prédiction dépend du nombre de voisins et de la qualité de la distance utilisée
- On le limite en général au cas de faibles dimensions (quitte à utiliser une méthode de réduction des dimensions telles que l'analyse en composantes principales).

Exercices

On considère le jeu de données suivant :

x1	1	2	2	2	3	3
x2	2	1	2	3	1	2
y	+	+	-	+	-	+

1. Faire le choix d'une distance et représenter la frontière de décision pour la méthode du plus proche voisin. Combien y a-t-il d'erreurs sur l'ensemble d'apprentissage ?
2. Même question avec les trois plus proches voisins.

Exercices

On considère le jeu de données suivant :

volume (ml)	250	100	125	250
caféine (g)	0,025	0,010	0,050	0,100
Boisson	Thé	Thé	Café	Café

1. On utilise la distance euclidienne, quelle est la prédiction pour une boisson de 125ml contenant 0,015g de caféine ?
2. Cette classification ne semble pas correcte vu le taux de caféine. Quel est le problème, comment y remédier ?