KNN

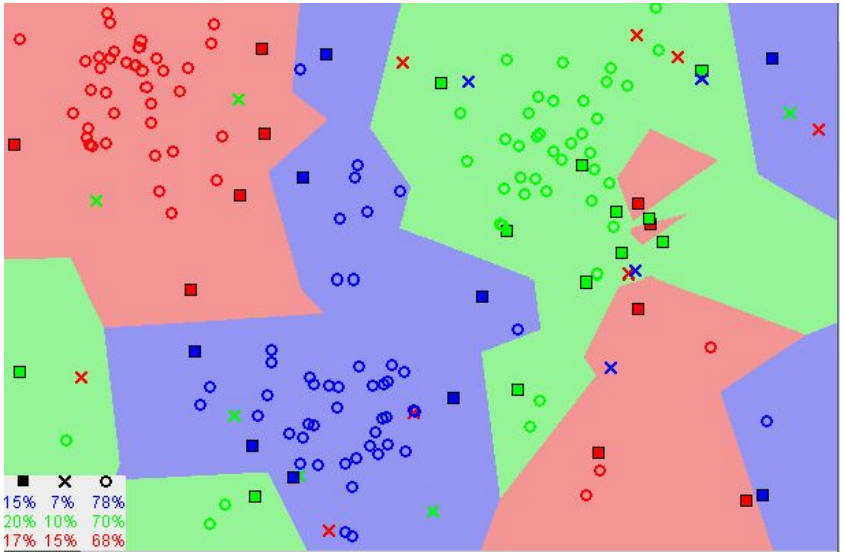
19 Novembre 2019

Sommaire

Défintion

K-Nearest Neighbors (KNN) est un algorithme supervisé qui se base sur la similarité des données quant à leur proximité aux autres données.

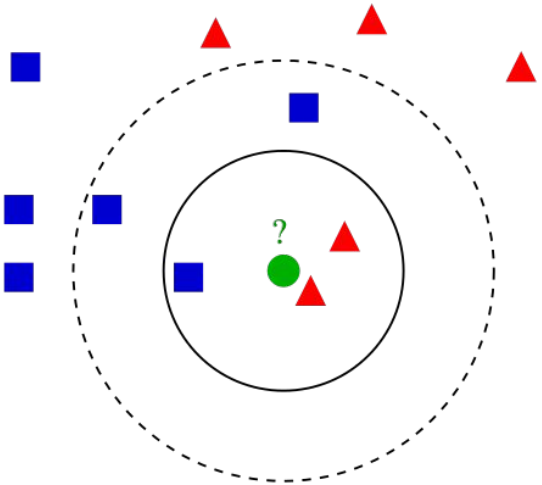
Il peut être utilisé pour la classification et la régression.



KNN Algorithme [1]

L’algorithme KNN

1) Obtenez un point de données non classifié dans l'espace n-dimensionnel.
2) Calculez la mesure de distance du nouveau point de données à tous les autres points de données qui sont déjà classifiés.
3) Obtenir les points de données correspondant aux k plus petites distances.
4) Compter les fois où chaque classe se produit parmi ces points de données.
5) La classe qui apparaissait avec la fréquence la plus élevée serait le choix de l'algorithme.
6) Affectez la classe sélectionnée à l'étape 5 comme classe du point de données.



KNN Example

Caracteristiques

L'algorithme ne fait aucune hypothèse sur la distribution des données sous-jacentes, mais il repose sur la similarité des caractéristiques des éléments. Lorsqu'un KNN fait une prédiction sur un film, il calcule la "distance" entre la cible et chaque autre élément de sa base de données.

Différentes formes de calcul de la distance (à choisir en accord du problème)

- **Distance euclidienne:-** Cette métrique est simplement la distance entre deux point dans un espace vectorielle.
- **Similarité cosinus**
- **coefficient de corrélation**

Cas d’usage

1) Recommandation de films
2) Recommandation de musique
3) … tout ce qui puisse mettre en relation des profils de comportement ou d’usage

Advantages

- Simple, facile à comprendre et facile à mettre en œuvre.
- Le temps de formation est zéro. D'autres algorithmes supervisés utilisent l'ensemble d'apprentissage pour former un modèle (c'est-à-dire ajuster une fonction), puis utilisent ce modèle pour classifier l'ensemble de validation ou l'ensemble de tests. KNN enregistre uniquement l'échantillon et le traite lorsqu'il reçoit les données de test. Le temps d'apprentissage de KNN est donc égal à zéro.
- Particulièrement adapté aux problèmes multimodaux (objets avec étiquettes de catégories multiples).
- KNN peut également traiter des problèmes de régression, c’est-à-dire des prévisions.

Disadvantages

- La quantité de calcul est trop grande, en particulier lorsque le nombre d'entités est très grand. Chaque texte à classer doit calculer sa distance par rapport à tous les échantillons connus afin d'obtenir son Kème voisin le plus proche.

- Lorsque l'échantillon est déséquilibré, la précision de prédiction pour les catégories rares est faible. Lorsque l'échantillon est déséquilibré, si la taille de l'échantillon d'une classe est grande et la taille de l'échantillon des autres classes est très petite, il peut en résulter que lorsqu'un nouvel échantillon est saisi, l'échantillon de la classe de grande capacité parmi les K voisins de l'échantillon est dominant.

- La dépendance à l'égard des données d'entraînement est particulièrement grande et la tolérance aux pannes des données d'entraînement est trop faible. Si le jeu de données d'apprentissage contient une ou deux données incorrectes, juste à côté de la valeur à classer, cela conduira directement à des données inexactes.

Références

[1] https://www.cnblogs.com/lsm-boke/p/11756173.html
[2] ttps://heartbeat.fritz.ai/recommender-systems-with-python-part-ii-collaborative-filtering-k-nearest-neighbors-algorithm-c8dcd5fd89b2