Analyse de Survie
Professeur Abdellatif El Afia

KNN de Survie

KNN

- Les plus proche voisins (nearest neighbour) est la technique la plus ancienne et très simple utilisée pour la classification.
- KNN (K Nearest Neighbour), un nombre entier positif fixe k de voisins les plus proches est autorisé à voter dans le processus de classification d'un nouveau tuple de données.
- KNN est appelé algorithme d'apprentissage paresseux. Il ne commence à fonctionner que lorsqu'il obtient un nouveau tuple pour la classification. En raison de ce principe de fonctionnement, KNN est aussi nommé un apprenant basé sur l'instance.

Algorithme KNN de base :

Entrée : *échantillons d'apprentissage D, échantillon de test d, K*

Sortie *: étiquette de classe de l'échantillon de test*

1 : Calculer la distance entre d et chaque échantillon de D

2 : Choisir les K échantillons de D les plus proches de d ; notons l'ensemble par P (⊂D)

3 : Attribuer à d la classe qui est la plus fréquente (ou la classe majoritaire)

Mesures de distance

1. **Non-négativité** : La distance entre x et y est toujours une valeur supérieure ou égale à zéro.

$$\forall x, y \qquad d(x, y) \geq 0$$

1. **Séparation** : La distance entre x et y est égale à zéro si et seulement si x est égal à y

$$\forall x, y \qquad d(x, y) = 0 \Leftrightarrow x = y$$

1. **Symétrie** : La distance entre x et y est égale à la distance entre y et x

$$\forall x, y \qquad d(x, y) = d(y, x)$$

1. **Inégalité triangulaire** : Compte tenu de la présence d'un troisième point z, la distance entre x et y est toujours inférieure ou égale à la somme de la distance entre x et z et de la distance entre y et z

$$\forall x, y, z \qquad d(x, y) \leq d(x, z) + d(z, y)$$

Mesures de distance

- huit grandes familles de distances qui consistent en cinquante-quatre mesures de distance totales.

1-Mesures de distance Lp Minkowski :

Cette famille de distances comprend trois métriques de distance qui sont des cas particuliers de distance de Minkowski, correspondant à différentes valeurs de p pour cette distance de puissance. La distance de Minkowski, également connue sous le nom de norme $\boldsymbol{L_p}$, est une métrique généralisée. Il est défini comme :

$$D_{mink}(x, y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

Où p est une valeur positive. Lorsque p = 2, la distance devient la distance **euclidienne**. Lorsque p = 1, cela devient la distance de **Manhattan**. La distance de **Chebyshev** est une variante de la distance de Minkowski où p = ∞. x_i est la $i^{ème}$ valeur dans le vecteur x et y_i est la $i^{ème}$ valeur dans le vecteur y.

2-Mesures de distance L1 :

Cette famille de distance dépend principalement de la recherche de la différence absolue, la famille comprend les distances de Lorentzian, Canberra, Sorensen, Soergel, Kulczynski, Mean Character, Non Intersection.

- Distance lorentzienne (LD)
- Distance de Canberra (CanD)
- Distance de Sorensen (SD)
- Distance de Soergel (SoD)
- Distance de Kulczynski (KD)
- Distance moyenne des caractères (MCD)
- Distance de non intersection (NID)

3- Mesures de distance du produit interne :

Les mesures de distance appartenant à cette famille sont calculées par certains produits de valeurs par paires des deux vecteurs, ce type de distances comprend : les distances Jaccard, Cosinus, Dice, Chord.

$$JacD(x,y) = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n} {x_i}^2 + \sum_{i=1}^{n} {y_i}^2 - \sum_{i=1}^{n} x_i y_i}$$

$$CosD(x,y) = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}}$$

$$DicD(x,y) = 1 - \frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2 + \sum_{i=1}^{n} {y_i}^2}$$

$$ChoD(x,y) = \sqrt{2 - \frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2 \sum_{i=1}^{n} {y_i}^2}}$$

4- Mesures de distance Squared Chord :

- Les distances qui appartiennent à cette famille sont obtenues en calculant la somme des géométries.
- La moyenne géométrique de deux valeurs est la racine carrée de leur produit.
- Les distances de cette famille ne peuvent pas être utilisées avec des vecteurs de caractéristiques de valeurs négatives.
- Cette famille comprend les distances Bhattachayya, Squared Chord, Matusita, Hellinger …

5- Mesures de distance L_2 au carré :

Dans la famille de mesures de distance L_2, le carré de la différence à chaque point sur les deux vecteurs est pris en compte pour la distance totale, cette famille comprend Squared Euclidean, Clark, Neyman $\chi 2$, Pearson $\chi 2$, Squared $\chi 2$, Probabilistic Symmetric $\chi 2$, Divergence, Additive Symétrique $\chi 2$, Moyenne, Moyenne Euclidienne censurée et Distances du chi carré au carré.

$$SED(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2 \; ; ClaD(x,y) = \sqrt{\sum_{i=1}^{n} \left(\frac{x_i - y_i}{|x_i| + |y_i|}\right)^2} \; ; NCSD(x,y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i}$$

$$PCSD(x,y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{y_i} \; ; SED(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2 \; ; ClaD(x,y) = \sqrt{\sum_{i=1}^{n} \left(\frac{x_i - y_i}{|x_i| + |y_i|}\right)^2}$$

.....

6- Mesures de distance d'entropie de Shannon :

Les mesures de distance appartenant à cette famille sont liées à l'entropie de Shannon. Ces distances incluent Kullback-Leibler, Jeffreys, la divergence K, Topsoe, JensenShannon, les distances de différence de Jensen.

- Distance de Kullback-Leibler (KLD)
- Distance de Jeffreys (JefD)
- Distance de K divergence (KDD)
- Distance de Topsoe (TopD)
- Distance de Jensen-Shannon (JSD)
- Distance de différence de Jensen (JDD)

7- Mesures de distance de vicissitude :

- La famille de distance de vicissitude se compose de quatre distances, les distances Vicis-Wave Hedges, Vicis Symmetric, Max Symmetric $\chi 2$ et Min Symmetric $\chi 2$.
- Ces distances ont été générées à partir de la relation syntaxique pour les mesures de distance susmentionnées.

❑ Distance Vicis-Wave Hedges (VWHD)

❑ Distance symétrique de Vicis (VSD)

❑ Distance $\chi 2$ symétrique maximale (MSCD)

❑ Distance minimale symétrique $\chi 2$ (MiSCSD)

8- Autres mesures de distance

Ces mesures présentent des mesures de distance utilisant plusieurs idées ou mesures à partir de mesures de distance précédentes, notamment, parmi d'autres, Moyenne (L_1, L_∞), Kumar-Johnson, Taneja, Pearson, Corrélation, Squared Pearson, Hamming, Hausdorff, statistique $\chi 2$, indice d'association de Whittaker, distances de Meehl, Motyka et Hassanat…

Variantes de KNN

Il existe de nombreuses lacunes associées à l'algorithme KNN. En modifiant les facteurs d'influence, les performances de KNN peuvent être améliorées.

Il existe de nombreuses variantes de l'algorithme KNN proposées dans différentes études qui ont tenté de surmonter ces lacunes.

1- KNN localement adaptatif

Dans l'algorithme KNN standard, la valeur globale du paramètre d'entrée k est utilisée. Mais cet algorithme proposé suggérait d'utiliser différentes valeurs du paramètre k pour différentes parties de l'espace d'entrée. A chaque fois pour la classification d'une requête, la valeur de k est déterminée via l'application d'une validation croisée dans son voisinage local.

2- KNN ajusté au poids

Dans l'algorithme KNN standard, tous les attributs ont la même importance. Ils donnent une contribution égale pour la classification des nouveaux tuples. Mais tous les attributs de l'ensemble de données n'ont pas la même importance. Un algorithme KNN ajusté par poids a été proposé, il apprend d'abord des poids pour différents attributs et selon les poids attribués, chaque attribut affecterait le processus de classification uniquement.

3- KNN amélioré pour la catégorisation de texte

- L'utilisation d'une valeur fixe de k pour toutes les classes entraînerait un biais vers la classe qui a le plus grand nombre de tuples.
- Différentes valeurs de k pour différentes classes en fonction de leur distribution de classe. Un plus grand nombre de tuples est utilisé pour classer un nouveau tuple dans une classe qui a des tuples plus nombreux.

4- KNN adaptatif

- Plutôt que d'utiliser une valeur fixe de k, KNN adaptatif utilise un nombre non fixe de plus proches voisins.
- Une grande valeur du paramètre k augmenterait également le coût et le temps de calcul en cas de grands ensembles de données.
- Pour résoudre ce problème, il applique trois heuristiques afin qu'une rupture précoce de l'algorithme soit possible. Ces heuristiques sur la réalisation d'une condition fixe sortiraient de l'algorithme.

5- KNN avec des voisins communs

Cet algorithme propose une autre variante de KNN qui utilise les plus proches voisins partagés pour classer les documents. Pour trouver les voisins d'un nouveau tuple, il utilise la mesure de similarité BM25. Un seuil est défini, seul ce nombre de voisins les plus proches peut voter pour la classification d'un tuple inconnu.

6- KNN avec K-moyennes (K-Means)

- KNN a une grande complexité de calcul. Pour pallier cet inconvénient en a combiné l'algorithme KNN avec l'algorithme de clustering K-Means.
- Les clusters des différentes catégories de l'ensemble de données d'apprentissage sont d'abord formés.
- Les centres de ces clusters nouvellement formés agiront désormais comme de nouveaux échantillons d'apprentissage. Pour classer un tuple inconnu, la distance de celui-ci est calculée avec ces nouveaux tuples d'apprentissage et il sera affecté à la classe du centre avec lequel le tuple a le moins de distance.
- L'avantage de cette variante de KNN est qu'il n'est pas nécessaire de passer le paramètre d'entrée k comme nous devons le faire dans le KNN standard.

7- KNN avec Mahalanobis métrique

- Une nouvelle métrique de distance appelée métrique de distance de Mahalanobis a été introduite.
- Elle transforme tout l'espace d'entrée à l'aide d'une transformation linéaire.
- Dans cet espace d'entrée transformé, la distance euclidienne est la même que la distance de Mahalanobis entre deux points de données.
- La distance euclidienne est la distance entre deux points quelconques, tandis que la distance de Mahalanobis est une distance entre un point et une distribution. Si le point représente la moyenne de la distribution, la distance de Mahalanobis serait nulle.
- Le principal avantage de prendre la métrique de distance de Mahalanobis au lieu de la métrique de distance euclidienne est qu'elle prend également en compte la corrélation entre les tuples de données.

8- KNN généralisé

L'algorithme KNN n'est pas seulement utilisé pour la prédiction des attributs de classe catégorique, mais également des attributs de classe à valeur continue. Dans ce dernier cas, la moyenne des valeurs de l'attribut de classe est affectée à l'attribut de classe du tuple inconnu.

9- KNN informatif

- Cet algorithme introduit une nouvelle mesure appelée informativité.
- Cette mesure tient compte du fait que tous les k plus proches voisins n'ont pas la même importance. L'algorithme prendrait deux paramètres d'entrée au lieu d'un, c'est-à-dire k et I.
- La valeur de I décidera du nombre de tuples de données informatives à prendre en compte pour la classification d'un tuple inconnu.
- L'algorithme trouve d'abord les k tuples de données les plus proches du tuple de test, puis il calcule l'informativité de ces k tuples de données
- La classe majoritaire des I tuples de données les plus informatifs serait la classe du tuple de données inconnues.

10- KNN bayésien

- Le classificateur bayésien est l'un des algorithmes de classification qui donne une assez bonne précision.
- Plutôt que de donner la classe du tuple de test en sortie, le classificateur bayésien donne l'appartenance du tuple aux classes sous forme de probabilités.
- Pour augmenter les performances de l'algorithme KNN dans le classement, une combinaison de classificateur bayésien et de KNN a été proposée.
- Dans cet algorithme, initialement les k voisins les plus proches des tuples de test sont déterminés. Après cela, ces k tuples de données sont utilisés pour former le classifieur.
- Le classificateur donnerait comme résultat les probabilités d'appartenance du tuple de test dans les classes. Ces probabilités sont utilisées pour classer les instances.

11- SVM-KNN

- La machine à vecteurs de support (SVM) est une technique de classification qui peut être utilisée à la fois sur des données linéaires et non linéaires.
- Cette version hybride de KNN avec SVM est utilisée pour la reconnaissance visuelle des catégories.
- Dans l'algorithme SVM-KNN, les k voisins les plus proches des tuples inconnus sont utilisés pour former SVM. Pour la mise en œuvre de cet algorithme hybride, les k tuples de données les plus proches sont d'abord déterminés.
- Après, les distances par paires entre ces k tuples de données sont calculées. La matrice de noyau est calculée à partir de cette matrice de distances obtenue. Cette matrice de noyau calculée est donnée en entrée au classificateur SVM. La sortie serait la classe du tuple inconnu.

12- Quelques autres extensions pour KNN

- classificateur KNN basé sur la densité.
- le classificateur KNN variable k,
- le classificateur KNN pondéré,
- le classificateur KNN basé sur les classes
- le classificateur KNN de discernabilité.

Une méthode de KNN de survie

- une méthode non paramétrique simple pour prédire la probabilité de survie dans le cadre de données censurées à droite sans risques concurrents.
- Cette méthode construit des courbes de survie de Kaplan–Meier sur la base de la survie observée des K plus proches voisins dans un ensemble d'apprentissage. La distance entre les points de données est mesurée avec une métrique sur les covariables associées aux observations.
- Le choix initial pour la métrique est la distance de Mahalanobis, une métrique largement applicable avec l'avantage de n'imposer aucune hypothèse structurelle sur les données.

Description de la méthode

- Étant donné une nouvelle observation j, cette méthode génère une prédiction pour la courbe de survie de j en créant une courbe de Kaplan-Meier à partir des temps de survie (éventuellement censurés) d'observations "similaires" à partir d'un ensemble de données existant S.
- Chaque observation $i \in S$ est associée à un ensemble de covariables $x_i \in \mathbb{R}^p$, un temps d'événement $t_i > 0$ et un indicateur de censure δ_i, où 0/1 indique une observation censurée/non censurée. L'ensemble de covariables x_j est associé à la nouvelle observation j.
- $d(x_i, x_j)$ une métrique sur x qui mesure à quel point les observations i et j sont similaires. Les K plus proches voisins de j dans S sont choisis selon $d(x_i, x_j)$ pour former l'ensemble $S_j^K \subseteq S$.

Description de la méthode

Un analogue pondéré de la courbe de survie de Kaplan–Meier générée à partir des observations dans S_j^K,

$$\widehat{S}^K(t|x_i;w) = \prod_{i \in S_j^K : t_i < t} \left(1 - \frac{d_i^w}{n_i^w}\right)$$

sert de prédiction pour la courbe de survie de l'observation j, où :

- $n_i^w = \sum_{k \in S_j^K} w\left(d(x_j, x_k)\right).I(t_k \geq t_i)$

est le nombre pondéré d'observations dans S_j^K à risque juste avant t_i et

- $d_i^w = \sum_{k \in S_j^K} w\left(d(x_j, x_k)\right).\delta_k.I(t_k = t_i)$

est le nombre pondéré de décès au temps t_i.

Description de la méthode

- La fonction de pondération $w(.)$ pour les observations en S_j^K est non croissante dans la distance $d(.,.)$ de l'observation à x_j , mettant ainsi davantage l'accent sur des observations d'entraînement qui ressemblent davantage à la nouvelle observation.

- K peut être choisi en séparant S en ensembles d'apprentissage et de validation et en effectuant des tests de validation. Les K plus proches voisins utilisés pour générer la courbe de survie prédite sont tirés de l'ensemble d'apprentissage. L'ensemble de validation est un ensemble de données d'exclusion utilisé pour tester les performances d'une plage de valeurs K et sélectionner le K le plus performant.

Description de la méthode

- Pour le choix de la métrique $d(x_i, x_j)$, il existe de nombreuses options possibles, et le choix doit être orienté par le contexte du problème. Une option polyvalente est la distance de Mahalanobis.

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T \Sigma_S (x_i - x_j)}$$

- Où, Σ_S est la matrice de covariance de x pour les observations appartenant à S.
- Cette pondération inverse dans la distance de Mahalanobis rend la métrique invariante à l'échelle, et donc indépendante des unités dans lesquelles les covariables sont exprimées.

- Notez que cette distance peut également être appliquée aux variables catégorielles en les transformant d'abord en un ensemble de variables indicatrices.

TP

- Implémenter la méthode KNN de survie décrite dans ce chapitre
- Appliquer votre implémentation sur des données de survie pour tirer des résultats