MINISTERE DE L'ENSEIGNEMENT SUPERIEUR

ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DE SOUSSE

المعهد العالي للإعلامية و تقنيات الاتصال بحمام سوسة

INSTITUT SUPERIEUR D'INFORMATIQUE

ET DES TECHNIQUES DE COMMUNICATION – HAMMAM SOUSSE

DATA MINING :

Recherche Distance MAHALANOBIS

Réalisé par : Gara Chayma

Enseignant : Mr khmais Abdallah

Matière : Data mining

Classe: 3DNI2

Année Universitaire 2021 – 2022

1. Définition distance de mahalanobis

La distance de Mahalanobis est une mesure de distance mathématique introduite par Prasanta Chandra Mahalanobis en 1936. Elle est basée sur la corrélation entre des variables par lesquelles différents modèles peuvent être identifiés et analysés. C'est une manière utile de déterminer la *similarité* entre une série de données connues et inconnues.

2. Difference entre distance Euclidienne et distance mahalanobis

La distance Mahalanobis diffère de la **distance euclidienne** par le fait qu'elle prend en compte la variance et la corrélation de la série de données.

Ainsi, à la différence de la **distance euclidienne** où toutes les composantes des vecteurs sont traitées indépendamment et de la même façon**, la distance de Mahalanobis** accorde un poids moins important aux composantes les plus dispersées. Dans le cas de l'analyse des signaux, et en supposant que chaque composante soit une variable aléatoire de type gaussien, cela revient à minimiser l'influence des composantes les plus bruitées (celles ayant la plus grande variance).

3. Utilisation de distance mahalanobis

La distance de Mahalanobis est souvent utilisée pour :

- La détection de données aberrantes dans un jeu de données en particulier dans le développement de modèles de régression linéaire
- Déterminer la cohérence de données fournies par un capteur
- La distance de Mahalanobis est largement utilisée dans les techniques d' analyse et de classification par grappes
- La distance Mahalanobis est nécessaire du problème d'identification des crânes sur la base des mesures en 1927.

Par exemple : cette distance est calculée entre les données reçues et celles prédites par un modèle.

4. Explication

Considérons le problème d'estimer la probabilité qu'un point à l'étude en espace euclidien N-dimensionnelle appartient à un ensemble, dont certains sont des échantillons de données qui appartiennent certainement à cet ensemble. Intuitivement, plus ce point est près du centre des masses, plus il est probable que fait partie de cet ensemble.

En outre, vous devez savoir si la collection est distribuée sur une petite ou grande distance, afin de décider si un certain rayon du centre est plus ou moins cohérente. L'approche la plus simple est d'estimer la écart-type des échantillons du centre de masse. Si la distance entre le point en cours d'examen et le centre de masse est inférieure à un écart-type, on peut conclure qu'il est fort probable que le point considéré appartient à l'ensemble. Plus cette distance, plus la probabilité que ce point devrait être classé comme appartenant à l'ensemble.

Cette approche intuitive peut être quantitative en définissant la distance normalisée entre le point en question et l'ensemble comme:

L'hypothèse de cette approche est que les points d'échantillonnage sont distribués dans un 'hypersphère autour du centre de masse. Dans le cas où la distribution est non sphérique (par exemple hyperellipsoïdale), Il serait naturel de penser que la probabilité du point en question appartiennent à l'ensemble dépend non seulement de la distance du centre de masse, mais aussi par la direction. Sur les directions dans lesquelles le iperellissoide a un axe plus court, le point à l'examen doit être plus proche afin d'être considérés comme appartenant à l'ensemble, tandis que les directions dans lesquelles l'axe est plus, le point en cours d'examen peut également trouver à des distances plus. Développer tout en termes mathématiques, l'hyper-ellipsoïde qui représente le mieux l'ensemble de la probabilité peut être estimée au moyen de la matrice de covariance de l'échantillon.

- La distance de Mahalanobis est donc simplement la distance du point considéré à partir du centre des masses normalisées ellipsoïde que la largeur dans la direction du point considéré