Royaume du Maroc
Université Abdelmalek Essaâdi
Faculté des Sciences et Techniques - Tanger
Département Génie Informatique
MST: Intelligence Artificielle et Sciences de Données

“ Découverte des Méthodes d'Apprentissage Automatique : Une Étude Approfondie de PCA, GMM et KNN ”

Réaliser par:

- **BENJELLOUN Abdelmajid**
- **DAGHMOUMI Marouan**

Encadrer par:

- **Pr. El Mokhtar EN-NAIMI**

Année universitaire 2023/2024

Table des matières

CHAPITRE 1 :
PRESENTATION GÉNÉRALE

1. INTRODUCTION :

Dans ce chapitre, nous introduisons les concepts généraux liés à notre projet de recherche sur l'application de trois algorithmes d'apprentissage automatique : Principal Component Analysis (PCA), Gaussian Mixture, et K-Nearest Neighbors (KNN). Ces algorithmes, au cœur de nombreuses applications intelligentes, offrent des outils puissants pour l'analyse de données, la classification et la prédiction.

2. OBJECTIFS DE L'ÉTUDE :

L'objectif principal de notre étude est de comprendre en profondeur ces trois algorithmes d'apprentissage automatique et d'explorer leurs applications potentielles dans divers domaines.
Nous cherchons à :

- Décrire en détail chaque algorithme, en mettant en évidence ses caractéristiques fondamentales, ses modes de fonctionnement et ses applications.
- Analyser les avantages et les inconvénients de chaque algorithme pour comprendre leurs limitations et leurs forces respectives.
- Identifier les domaines d'application où chaque algorithme peut être le plus efficace, en mettant en évidence les cas d'utilisation spécifiques et les scénarios où leur utilisation est pertinente.

3. PRÉSENTATION DES ALGORITHMES :

Dans les sections suivantes, nous présentons en détail chacun des trois algorithmes :

- **Principal Component Analysis (PCA) :** Nous explorons les principes fondamentaux de PCA, sa capacité à réduire la dimensionnalité des données tout en préservant leur structure, et ses applications dans la visualisation de données et la compression de données.
- **Gaussian Mixture :** Nous examinons le modèle de Mélange Gaussien, sa capacité à modéliser des données complexes à l'aide de distributions gaussiennes et ses applications dans le clustering et la détection d'anomalies.
- **K-Nearest Neighbors (KNN) :** Nous analysons l'algorithme KNN, son approche intuitive basée sur la similarité des voisins les plus proches, et ses applications dans la classification et la régression.

À travers ces discussions, nous mettons en lumière les forces et les limitations de chaque algorithme, en illustrant leur pertinence dans divers scénarios d'apprentissage automatique.

4. STRUCTURE DU RAPPORT :

Ce rapport est structuré de manière à fournir une analyse approfondie de chaque algorithme, accompagnée d'exemples concrets, de démonstrations et d'implémentations pratiques. Chaque chapitre est dédié à un algorithme spécifique, avec des discussions approfondies sur son fonctionnement, ses applications et ses performances.

Dans les chapitres suivants, nous plongerons dans l'étude détaillée de chaque algorithme, en fournissant des exemples de code, des visualisations et des cas d'utilisation pour illustrer leurs concepts et leur utilisation pratique.

CHAPITRE 2 :
ANALYSE EN COMPOSANTES PRINCIPALES (PCA)

1. INTRODUCTION:

Dans ce chapitre, nous explorons en détail l'Analyse en Composantes Principales (PCA), une technique fondamentale en apprentissage automatique pour la réduction de dimensionnalité et l'extraction de caractéristiques. Nous commençons par une introduction générale à PCA, puis nous examinons ses étapes, ses relations mathématiques, ses caractéristiques, ses modes de fonctionnement, ainsi que ses avantages et inconvénients. Nous terminons en explorant ses domaines d'application.

2. DÉFINITION :

PCA est une méthode statistique qui permet de transformer un ensemble de variables corrélées en un nouvel ensemble de variables non corrélées appelées composantes principales. Elle est largement utilisée pour simplifier la complexité des données tout en préservant leur structure et leurs relations essentielles. PCA est appliquée dans divers domaines, notamment la biologie, la finance, l'imagerie médicale et l'analyse de données.

3. Étapes de l'Analyse en Composantes Principales (PCA) :

L'Analyse en Composantes Principales (PCA) comprend plusieurs étapes cruciales pour réduire la dimensionnalité des données tout en préservant au mieux leur structure et leurs relations essentielles. Dans cette section, nous détaillerons ces étapes :

1. NORMALISATION DES DONNÉES :

Avant d'appliquer PCA, il est essentiel de normaliser les données en les centrant autour de zéro pour avoir une moyenne nulle. Cela garantit que chaque variable contribue également à l'analyse et évite toute domination due à des échelles de mesure différentes entre les variables.

$$x_{ij} = (x_{ij} - \bar{x}_j)/\sigma_j$$

2. CALCUL DE LA MATRICE DE COVARIANCE :

Une fois les données normalisées, nous calculons la matrice de covariance, qui mesure les relations linéaires entre les variables. Cette matrice symétrique indique la force et la direction des relations linéaires entre chaque paire de variables.

$$Cov(X, Y) = (1/(n-1)) \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

3. DÉCOMPOSITION EN VALEURS PROPRES :

Après avoir obtenu la matrice de covariance, nous décomposons cette matrice en ses valeurs propres et vecteurs propres associés. Les valeurs propres représentent la quantité de variance expliquée par chaque composante principale, tandis que les vecteurs propres indiquent la direction de ces composantes.

$$CovMat = V * D * V^{-1}$$

4. SÉLECTION DES COMPOSANTES PRINCIPALES :

Une fois que nous avons les valeurs propres, nous sélectionnons les composantes principales qui captent le plus de variance dans les données. Habituellement, nous choisissons les premières composantes principales qui expliquent la majorité de la variance dans les données.

5. PROJECTION DES DONNÉES :

Enfin, nous projetons les données sur les nouveaux axes principaux définis par les composantes principales sélectionnées. Cela réduit la dimensionnalité des données tout en préservant les informations importantes contenues dans les données originales.

Ces étapes constituent le cœur de l'Analyse en Composantes Principales (PCA) et sont essentielles pour obtenir des résultats significatifs dans la réduction de dimensionnalité et l'extraction de caractéristiques.

4. Relations Mathématiques de PCA :

Dans cette section, nous présentons les relations mathématiques essentielles de l'Analyse en Composantes Principales (PCA) :

1. MATRICE DE COVARIANCE :

La matrice de covariance $\text{cov}(X)$ des données standardisées X est calculée comme suit :

$$\text{cov}(X) = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})$$

où n est le nombre d'échantillons et $\bar{X}$ est le vecteur moyen des caractéristiques.

Objectif : Calculer la matrice de covariance permet de mesurer les relations linéaires entre les variables dans les données, ce qui est essentiel pour l'Analyse en Composantes Principales (PCA).

2. DÉCOMPOSITION EN VALEURS PROPRES :

Après avoir calculé la matrice de covariance, nous effectuons sa décomposition en valeurs propres pour obtenir les vecteurs propres $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ et les valeurs propres correspondantes $\lambda_1, \lambda_2, \ldots, \lambda_p$. Cela permet de réécrire la matrice de covariance sous forme diagonale :

$$\text{cov}(X) = V\Lambda V^{-1}$$

où V est la matrice dont les colonnes sont les vecteurs propres et Λ est une matrice diagonale contenant les valeurs propres.

Objectif : La décomposition en valeurs propres nous permet de trouver les axes principaux qui captent le plus de variance dans les données.

3. PROJECTION DES DONNÉES :

Enfin, pour projeter les données X sur les nouveaux axes principaux définis par les vecteurs propres, nous effectuons la multiplication matricielle suivante ::

$$X_{\text{proj}} = XV$$

où X_{proj} est la matrice des données projetées sur les nouveaux axes principaux.

Objectif : La projection des données dans le nouvel espace de caractéristiques nous permet de réduire la dimensionnalité tout en préservant la structure et les relations sous-jacentes des données.

5. Caractéristiques de PCA :

Les caractéristiques de l'Analyse en Composantes Principales (PCA) comprennent plusieurs aspects qui définissent son fonctionnement et son utilisation. Voici les caractéristiques principales de PCA en détail :

1. RÉDUCTION DE LA DIMENSIONNALITÉ :

PCA permet de réduire la dimensionnalité d'un ensemble de données en transformant les variables d'origine en un nouvel ensemble de variables (composantes principales) qui capturent le maximum de variance dans les données. Cela permet de simplifier la représentation des données tout en préservant leur structure et leurs relations essentielles.

2. ANALYSE DES STRUCTURES DE DONNÉES :

PCA est largement utilisée pour explorer et analyser les structures sous-jacentes des données. En identifiant les axes principaux qui capturent le plus de variance, PCA permet de mettre en évidence les relations linéaires entre les variables et de détecter les tendances ou les motifs significatifs dans les données.

3. DÉCORRÉLATION DES VARIABLES :

L'une des propriétés importantes de PCA est qu'elle produit des composantes principales qui sont orthogonales les unes aux autres, c'est-à-dire qu'elles sont non corrélées. Cela permet de décorréler les variables dans le nouvel espace de caractéristiques, ce qui simplifie l'interprétation des relations entre les variables.

4. COMPRESSION DES DONNÉES :

En réduisant la dimensionnalité des données, PCA permet également de compresser l'information contenue dans les données tout en préservant au mieux leur structure. Cela peut être utile pour stocker et analyser de grandes quantités de données de manière plus efficace.

5. VISUALISATION DES DONNÉES :

PCA est couramment utilisée pour la visualisation des données, en particulier dans le domaine de l'apprentissage automatique et de la science des données. En projetant les données sur un espace de dimension réduite, PCA permet de représenter graphiquement les relations entre les échantillons et de visualiser les clusters ou les groupes dans les données.

6. PRÉTRAITEMENT DES DONNÉES :

PCA peut également être utilisée comme étape de prétraitement des données avant l'application d'autres algorithmes d'apprentissage automatique. En réduisant la dimensionnalité et en décorrélation des variables, PCA peut améliorer les performances des modèles prédictifs en réduisant le risque de surajustement et en accélérant la convergence des algorithmes d'optimisation.

7. INTERPRÉTABILITÉ DES RÉSULTATS :

Les composantes principales produites par PCA sont ordonnées en fonction de leur contribution à la variance totale des données, ce qui permet de les interpréter facilement en termes de leur importance relative dans la représentation des données. Cela facilite l'interprétation des résultats et la prise de décision basée sur les analyses PCA.

6. Avantages et Inconvénients de PCA :

1. LES AVANTAGES :

- **Réduction de la dimensionnalité :**

PCA permet de réduire le nombre de variables dans un ensemble de données tout en préservant au mieux leur structure et leurs relations sous-jacentes.

- **Simplicité et interprétabilité :**

Les composantes principales obtenues par PCA sont des combinaisons linéaires des variables originales, ce qui facilite leur interprétation et leur compréhension.

- **Élimination de la corrélation :**

PCA permet de décorréler les variables en transformant l'ensemble de données en un nouvel espace où les variables sont orthogonales les unes par rapport aux autres.

- **Visualisation des données :**

PCA permet de projeter les données dans un espace de dimension réduite, ce qui facilite leur visualisation et leur compréhension à travers des graphiques et des représentations visuelles.

- **Amélioration des performances des modèles :**

En réduisant la dimensionnalité des données et en éliminant la corrélation entre les variables, PCA peut améliorer les performances des modèles d'apprentissage automatique en réduisant le risque de surajustement et en accélérant la convergence des algorithmes d'optimisation.

2. LES INCONVÉNIENTS :

- **Perte d'information :**

Lors de la réduction de la dimensionnalité, PCA peut entraîner une perte d'information, car certaines informations peuvent être sacrifiées pour simplifier la représentation des données.

- **Sensibilité aux valeurs aberrantes :**

PCA est sensible aux valeurs aberrantes dans les données, ce qui peut affecter les résultats de manière significative, en particulier dans les ensembles de données avec des valeurs aberrantes importantes.

- **Interprétation limitée des composantes :**

Bien que les composantes principales obtenues par PCA soient facilement interprétables, leur signification peut être limitée dans certains cas, en particulier lorsque les variables originales sont fortement corrélées ou lorsque les données sont très complexes.

- **Dépendance à la linéarité :**

PCA repose sur l'hypothèse de linéarité des relations entre les variables, ce qui peut limiter son efficacité dans la capture de structures non linéaires dans les données.

7. Domaines d'application:

L'Analyse en Composantes Principales (PCA) trouve des applications dans divers domaines en raison de sa capacité à réduire la dimensionnalité des données tout en préservant leur structure sous-jacente. Voici quelques-uns des domaines d'application où PCA est couramment utilisée :

1. EXPLORATION DE DONNÉES :

PCA est souvent utilisée pour explorer de grands ensembles de données en identifiant les tendances, les patterns et les relations entre les variables. Elle permet de visualiser les données de manière concise et de détecter les structures sous-jacentes.

2. BIOLOGIE ET GÉNOMIQUE :

Dans le domaine de la biologie, PCA est utilisée pour analyser les données génomiques, telles que les données d'expression génique et les données de séquençage d'ADN. Elle permet d'identifier les gènes ou les variations génétiques qui sont associés à des phénotypes spécifiques.

3. FINANCE :

En finance, PCA est utilisée pour l'analyse de portefeuille et la gestion des risques. Elle permet de réduire la dimensionnalité des données financières tout en conservant les principales sources de variation, ce qui facilite la prise de décision et la modélisation des risques.

4. TRAITEMENT DU SIGNAL :

En traitement du signal, PCA est utilisée pour extraire les composantes principales des signaux et réduire leur dimensionnalité. Cela permet de filtrer le bruit et d'identifier les caractéristiques importantes des signaux.

4. MARKETING ET ANALYSE DE DONNÉES CLIENT :

PCA est utilisée dans le domaine du marketing pour segmenter les clients en fonction de leurs comportements d'achat et de leurs préférences. Elle permet de réduire la dimensionnalité des données client tout en identifiant les segments de marché les plus importants.

8. Conclusion:

Dans ce chapitre, nous avons exploré en détail l'Analyse en Composantes Principales (PCA). Nous avons commencé par présenter les étapes fondamentales de l'algorithme PCA, allant de la standardisation des données à la projection dans le nouvel espace de caractéristiques. Ensuite, nous avons examiné les relations mathématiques sous-jacentes à chaque étape, mettant en évidence l'importance de la matrice de covariance et de la décomposition en vecteurs propres et valeurs propres.

En poursuivant, nous avons discuté des caractéristiques de PCA, soulignant sa capacité à réduire la dimensionnalité des données tout en préservant leur structure essentielle. Nous avons également examiné les modes de fonctionnement de PCA, illustrant comment il capture les principales sources de variation dans les données.

En outre, nous avons analysé les avantages et les inconvénients de PCA, mettant en lumière sa simplicité d'utilisation et sa capacité à gérer des données complexes, tout en notant sa sensibilité aux valeurs aberrantes et son exigence de données standardisées. Enfin, nous avons discuté des domaines d'application de PCA, montrant comment il est utilisé dans divers domaines tels que l'analyse des données, la vision par ordinateur et la biologie.

CHAPITRE 3 : K-Nearest Neighbors (KNN)

1. INTRODUCTION:

Dans ce chapitre, nous explorons l'algorithme des k plus proches voisins (KNN), une méthode simple mais puissante d'apprentissage supervisé utilisée pour la classification et la régression. Nous présentons en détail la description de l'algorithme KNN, ses caractéristiques, ses modes de fonctionnement, ainsi que ses avantages et inconvénients. Ensuite, nous examinons quelques domaines d'application où KNN est largement utilisé, en mettant l'accent sur ses avantages par rapport à d'autres techniques d'apprentissage automatique.

2. DÉFINITION:

L'algorithme des k plus proches voisins (KNN) est une méthode d'apprentissage supervisé utilisée pour la classification et la régression:

- En classification, l'algorithme attribue une étiquette de classe à une instance en fonction de la majorité des étiquettes de classe de ses voisins les plus proches.
- En régression, il prédit la valeur cible d'une instance en prenant la moyenne des valeurs cibles de ses voisins les plus proches.

3. Étapes de k plus proches voisins (KNN):

Les étapes clés de l'algorithme K-Nearest Neighbors (KNN) sont les suivantes :

1. SÉLECTION DU NOMBRE DE VOISINS (K):

Déterminez le nombre de voisins à considérer pour la prédiction K .

2. CALCUL DE LA DISTANCE:

Utilisez une mesure de distance comme la distance euclidienne pour calculer la distance entre le nouvel exemple et tous les exemples du jeu de données.

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

Où :

- x et y sont deux vecteurs représentant des points dans l'espace.
- xi et yi sont les composantes respectives de ces vecteurs.
- n est la dimension de l'espace .

3. SÉLECTION DES K VOISINS LES PLUS PROCHES :

Identifiez les K voisins les plus proches en fonction des distances calculées.

4. ATTRIBUTION DE L'ÉTIQUETTE DE CLASSE MAJORITAIRE :

Sélectionnez l'étiquette de classe majoritaire parmi les K voisins.

$$\hat{y} = \textbf{mode}(y_{\textbf{voisins}})$$

Où :

- $y_{\textbf{voisins}}$ est l'ensemble des étiquettes de classe des voisins les plus proches de x.
- $\hat{y}$ est la prédiction de classe pour le point x.

4. Caractéristiques de KNN :

Les caractéristiques de l'algorithme K-Nearest Neighbors (KNN) sont les suivantes :

1. MÉTHODE NON-PARAMÉTRIQUE:

KNN est une méthode d'apprentissage non paramétrique, ce qui signifie qu'elle ne fait aucune hypothèse explicite sur la distribution des données.

2. CLASSIFICATION ET RÉGRESSION:

KNN peut être utilisé à la fois pour la classification et la régression, ce qui en fait un algorithme polyvalent.

3. SIMPLICITÉ:

KNN est simple à comprendre et à mettre en œuvre. Il ne nécessite pas de phase d'apprentissage coûteuse car il mémorise simplement les données d'entraînement.

4. INTERPRÉTABILITÉ:

Les prédictions de **KNN** sont souvent faciles à interpréter, en particulier dans le cas de la classification où la classe majoritaire parmi les voisins les plus proches est utilisée comme prédiction.

5. ROBUSTESSE AUX DONNÉES BRUITÉES:

KNN peut être robuste aux données bruitées, en particulier si K est choisi de manière appropriée pour lisser les fluctuations.

6. SENSIBILITÉ AUX CARACTÉRISTIQUES:

KNN peut être sensible aux valeurs aberrantes et au bruit dans les données, car il se base sur la proximité des points. Un mauvais choix de k peut conduire à une classification incorrecte ou à une mauvaise prédiction.

5. Avantages et Inconvénients de KNN :

1. LES AVANTAGES :

- **Facilité d'implémentation :**

KNN est facile à comprendre et à mettre en œuvre, ce qui en fait un bon choix pour les problèmes simples où la complexité de l'algorithme n'est pas une préoccupation majeure.

- **Adaptabilité aux données complexes :**

KNN peut s'adapter à des distributions de données complexes et à des frontières de décision non linéaires sans nécessiter de modifications de l'algorithme lui-même.

- **Utilisation dans les systèmes de recommandation :**

En raison de sa simplicité et de sa capacité à capturer les similitudes entre les instances, KNN est souvent utilisé dans les systèmes de recommandation pour recommander des éléments similaires à ceux que l'utilisateur a précédemment appréciés.

- **Pas de formation de modèle :**

Comme KNN n'a pas de phase d'apprentissage, il peut être utilisé dans des situations où les données changent fréquemment ou lorsque le coût de la formation d'un modèle est prohibitif.

2. LES INCONVÉNIENTS :

- **Sensibilité à la dimensionnalité :**

KNN devient moins efficace à mesure que le nombre de dimensions des données augmente, en raison du phénomène de la malédiction de la dimensionnalité.

- **Calcul intensif :**

La prédiction de la classe d'une nouvelle instance dans KNN implique le calcul de la distance entre cette instance et toutes les instances d'apprentissage, ce qui peut être intensif en calcul pour de grands ensembles de données.

- **Besoin de données étiquetées :**

Comme KNN est un algorithme d'apprentissage supervisé, il nécessite des données étiquetées pour fonctionner. Si les données ne sont pas étiquetées ou si les étiquettes sont incorrectes, cela peut affecter la qualité des prédictions.

- **Sensibilité à l'échelle des fonctionnalités :**

KNN est sensible à l'échelle des fonctionnalités, ce qui signifie que les caractéristiques avec des échelles différentes peuvent biaiser les calculs de distance.

6. Domaines d'application:

L'algorithme K-Nearest Neighbors (KNN) trouve des applications dans de nombreux domaines, notamment :

1. CLASSIFICATION D'IMAGES :

KNN est largement utilisé dans la classification d'images pour identifier des objets, des caractéristiques ou des motifs similaires dans des ensembles de données d'images volumineux.

2. SYSTÈMES DE RECOMMANDATION :

KNN est utilisé dans les systèmes de recommandation pour recommander des produits, des films, des chansons ou d'autres éléments similaires à ceux que l'utilisateur a déjà appréciés.

3. BIOINFORMATIQUE :

Il est utilisé dans l'analyse de séquences génétiques pour classer les gènes, prédire les fonctions des protéines ou identifier les similitudes entre les séquences.

4. ANALYSE MÉDICALE :

KNN est utilisé dans la classification des patients en fonction de leurs caractéristiques médicales pour diagnostiquer des maladies, prédire des résultats ou recommander des traitements.

5. DÉTECTION DE FRAUDE :

Il est utilisé dans les systèmes de détection de fraude pour identifier les transactions frauduleuses en fonction de schémas de comportement similaires à ceux observés dans des transactions frauduleuses précédentes.

6. DÉTECTION D'ANOMALIES :

Il est utilisé dans la détection d'anomalies pour identifier les comportements anormaux dans les systèmes informatiques, les réseaux de communication ou les processus industriels.

7. Conclusion:

L'algorithme des k plus proches voisins (KNN) est une méthode simple mais puissante en apprentissage automatique, offrant une approche intuitive pour la classification et la régression. En examinant ses caractéristiques, ses avantages et ses inconvénients, ainsi que ses domaines d'application, nous avons pu comprendre comment il peut être utilisé efficacement dans divers scénarios. Bien que KNN présente quelques limitations, telles que la sensibilité aux données bruitées et la nécessité de choisir le bon nombre de voisins, il reste un outil précieux dans de nombreux domaines, de la reconnaissance de formes à la recommandation de produits. En fin de compte, en comprenant ses fonctionnalités et ses compromis, nous sommes mieux équipés pour utiliser KNN de manière judicieuse et efficace dans nos projets d'apprentissage automatique.

CHAPITRE 4 : Gaussian Mixture (GMM)

1. INTRODUCTION:

Dans ce chapitre, nous explorons l'algorithme Gaussian Mixture Model (GMM), une technique d'apprentissage non supervisée largement utilisée pour la modélisation de données. Nous examinons en détail les étapes de l'algorithme, ses caractéristiques, ses avantages et ses inconvénients, ainsi que ses domaines d'application.

2. DÉFINITION:

Le Gaussian Mixture Model est une méthode probabiliste qui suppose que les données sont générées à partir d'un mélange de plusieurs distributions gaussiennes. Il cherche à estimer les paramètres de ces distributions pour mieux comprendre la structure sous-jacente des données.

3. Étapes de Gaussian Mixture (GMM) :

L'algorithme Gaussian Mixture Model (GMM) est une méthode d'apprentissage non supervisé largement utilisée pour la modélisation de données. Voici les étapes clés de l'algorithme GMM :

1. INITIALISATION DES PARAMÈTRES :

Initialiser les paramètres du modèle de manière aléatoire, y compris les moyennes, les matrices de covariance et les pondérations des composantes.

$$\mu_{k0}, \Sigma_{k0}, \pi_{k0} \quad \textbf{pour} \quad k = 1, \ldots, K$$

- μ_0, Σ_0, π_0 : Les paramètres initiaux du modèle, comprenant les moyennes (μ_k^0) , les matrices de covariance (Σ_k^0) et les poids des composantes (π_k^0) pour chaque composante gaussienne *K* dans le mélange.

2. ESTIMATION DE L'ESPÉRANCE DE LA LOG-VRAISEMBLANCE :

Calcul de la probabilité conditionnelle (probabilité postérieure) de chaque point d'appartenance à chaque cluster, basée sur les paramètres actuels du modèle. Utilisation de la formule de la loi normale multivariée pour calculer les probabilités. La probabilité d'un point xi appartenant à un cluster *K* est calculée comme suit

$$P(z_{ik} = 1 \mid \mathbf{x}_i, \boldsymbol{\theta}) = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

- $P(z_{ik} = 1 \mid x_i)$ **:** La probabilité que l'échantillon x_i appartienne à la composante gaussienne *K* , donnée par la formule de Bayes en utilisant les paramètres actuels du modèle.
- π_k **:** Le poids de la composante *K*
- $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$: La densité de probabilité gaussienne multivariée pour l'échantillon xi avec les paramètres μ_k (moyenne) et Σ_k (matrice de covariance) de la composante *K*.

3. MISE À JOUR DES PARAMÈTRES :

estimation des moyennes, covariances et poids des composantes en utilisant les données et les probabilités postérieures calculées. Utilisation de formules d'estimation de maximum de vraisemblance pour ajuster les paramètres. Les nouvelles estimations des paramètres sont calculées comme suit :

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik}) \cdot x_i$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik}) \cdot (x_i - \mu_k)(x_i - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

- μ_k **:** La nouvelle moyenne de la composante K, calculée comme la moyenne pondérée des échantillons en utilisant les poids $P(z_{ik} = 1 \mid x_i)$.
- Σ_k **:** La nouvelle matrice de covariance de la composante K, calculée comme la covariance pondérée des échantillons en utilisant les poids $P(z_{ik} = 1 \mid x_i)$.
- pi_k : Les nouveaux poids de la composante K, calculés comme la proportion des échantillons attribués à la composante K .

4. CALCUL DE LA LOG-VRAISEMBLANCE :

Calcul de la log-vraisemblance totale pour évaluer la convergence de l'algorithme.Utilisation de la formule de la log-vraisemblance pour évaluer la qualité du modèle. La log-vraisemblance totale est calculée comme suit :

$$\log p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

- $p(X \mid \mu, \Sigma, \pi)$ **:** La log-vraisemblance du modèle GMM, qui est la probabilité de générer l'ensemble des données X à partir des paramètres μ,Σ et π du modèle.

5. CONVERGENCE :

- Vérification de la convergence de l'algorithme en comparant les log-vraisemblances successives.
- Si Non ,Ces étapes sont itérées jusqu'à la convergence du modèle, et elles permettent d'estimer les distributions de probabilité sous-jacentes des données et d'assigner chaque échantillon à une ou plusieurs composantes gaussiennes en fonction de leur vraisemblance.

4. Caractéristiques de GMM :

Les caractéristiques de l'algorithme Gaussian Mixture (GMM) sont les suivantes :

1. CLUSTERING SOUPLE:

Contrairement à K-Means, où chaque point de données est assigné à un seul cluster, GMM permet à un point de données d'appartenir à plusieurs clusters avec des probabilités différentes, ce qui permet une segmentation plus flexible des données.

2. ESTIMATION DES PARAMÈTRES:

GMM est capable d'estimer les paramètres du modèle, y compris les moyennes, les covariances et les poids de mélange, à partir des données d'entraînement, ce qui en fait une méthode d'apprentissage non supervisé puissante.

3. GESTION DES DONNÉES NON-LINÉAIRES:

Bien que GMM soit souvent utilisé pour modéliser des données gaussiennes, il peut également être utilisé pour des données non linéaires en combinant plusieurs composantes gaussiennes pour représenter la distribution des données.

4. GESTION DES DONNÉES NON-LINÉAIRES:

Comme GMM prend en compte la covariance des données, il est plus robuste aux valeurs aberrantes par rapport à K-Means, ce qui le rend approprié pour les ensembles de données contenant des valeurs aberrantes ou des structures de cluster complexes.

5. Avantages et Inconvénients de GMM :

1. LES AVANTAGES :

- **Flexibilité dans la Modélisation :**

GMM peut modéliser des distributions de données complexes en utilisant plusieurs composantes gaussiennes, ce qui le rend approprié pour un large éventail de données.

- **Robustesse aux Formes de Cluster Variées :**

En permettant une modélisation probabiliste des clusters, GMM peut gérer des formes de cluster variées et des structures de données complexes.

- **Génération de Données :**

Une fois le modèle entraîné, GMM peut être utilisé pour générer de nouvelles données qui suivent la même distribution que l'ensemble de données d'entraînement.

2. LES INCONVÉNIENTS :

- **Sensibilité au Nombre de Composantes :**

GMM nécessite de spécifier le nombre de composantes gaussiennes, ce qui peut être difficile à déterminer, surtout pour des ensembles de données de grande dimension ou lorsque les clusters sont fortement chevauchants.

- **Temps de Calcul :**

L'entraînement de GMM peut être intensif en termes de calcul, en particulier pour des ensembles de données de grande taille ou avec un grand nombre de dimensions, car il implique l'estimation de plusieurs paramètres.

- **Initialisation Sensible :**

Les performances de GMM dépendent de l'initialisation des paramètres du modèle, et une mauvaise initialisation peut entraîner des résultats sous-optimaux ou des problèmes de convergence.

6. Domaines d'application:

1. CLUSTERING DE DONNÉES :

GMM est largement utilisé pour regrouper des données non étiquetées en identifiant des structures de cluster sous-jacentes basées sur la similarité des données.

2. SEGMENTATION D'IMAGE :

En traitement d'image, GMM peut être utilisé pour segmenter des régions d'intérêt en identifiant des groupes de pixels similaires dans une image.

3. MODÉLISATION DE DONNÉES BIOMÉDICALES :

En biologie et en médecine, GMM peut être utilisé pour modéliser des ensembles de données complexes, comme des profils d'expression génique ou des données de séquençage d'ADN.

4. DÉTECTION D'ANOMALIES :

GMM peut être utilisé pour détecter des anomalies dans des ensembles de données en identifiant des observations qui ne suivent pas le modèle de distribution estimé.

5. ANALYSE DE SÉRIES TEMPORELLES :

GMM peut être appliqué à des données séquentielles pour identifier des motifs et des tendances cachés dans des séries temporelles, telles que les modèles de comportement des utilisateurs ou les tendances du marché financier.

7. Conclusion:

En conclusion, Gaussian Mixture Models (GMM) offrent une approche flexible et puissante pour la modélisation de distributions de données complexes. À travers une combinaison de distributions gaussiennes, GMM peut capturer la structure latente des données et est utilisé dans divers domaines tels que le clustering, la segmentation d'image, la compression d'image, la modélisation biomédicale, la détection d'anomalies et l'analyse de séries temporelles. Cependant, GMM présente des inconvénients tels que la sensibilité aux paramètres initiaux et la complexité de calcul. Néanmoins, avec une bonne compréhension de ses caractéristiques et de ses étapes, GMM reste un outil précieux pour l'analyse et l'exploration de données.

Conclusion Generale

En conclusion, ce projet a permis d'explorer trois algorithmes d'apprentissage automatique : l'Analyse en Composantes Principales (PCA), le K-Nearest Neighbors (KNN) et le Gaussian Mixture Model (GMM). Chaque algorithme a été présenté en détail, mettant en lumière ses caractéristiques, ses modes de fonctionnement, ses avantages et ses inconvénients, ainsi que ses domaines d'application. Nous avons également implémenté ces algorithmes dans des cas d'utilisation concrets, démontrant leur efficacité dans la résolution de problèmes réels. En fin de compte, ce projet nous a permis de mieux comprendre les concepts fondamentaux de l'apprentissage automatique et de leur application pratique dans divers domaines.

Références

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
https://hastie.su.domains/Papers/ESLII.pdf
[2] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
https://github.com/peteflorence/MachineLearning6.867/blob/master/Bishop/Bishop%20-%20Pattern%20Recognition%20and%20Machine%20Learning.pdf
[3] Raschka, S., & Mirjalili, V. (2017). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow. Packt Publishing Ltd.
https://radio.eng.niigata-u.ac.jp/wp/wp-content/uploads/2020/06/python-machine-learning-2nd.pdf
[4] Comprendre le Machine Learning: L'algorithme du KNN .
www.youtube.com/watch?v=9pvbEP1eyNY
[5] Principal Component Analysis | PCA | Dimensionality Reduction in Machine Learning by Mahesh Huddar.
www.youtube.com/watch?v=ZtS6sQUAh0c
[6] Gaussian Mixture Models (GMM) Explained
www.youtube.com/watch?v=wT2yLNUfyoM
[7] EM algorithm: how it works.
www.youtube.com/watch?v=REypj2sy_5U
[8] KNN Algorithm Explained with Simple Example.
www.youtube.com/watch?v=XgYETToAHDU
[9] Data Analysis 6: Principal Component Analysis (PCA) - Computerphile.
www.youtube.com/watch?v=TJdH6rPA-TI