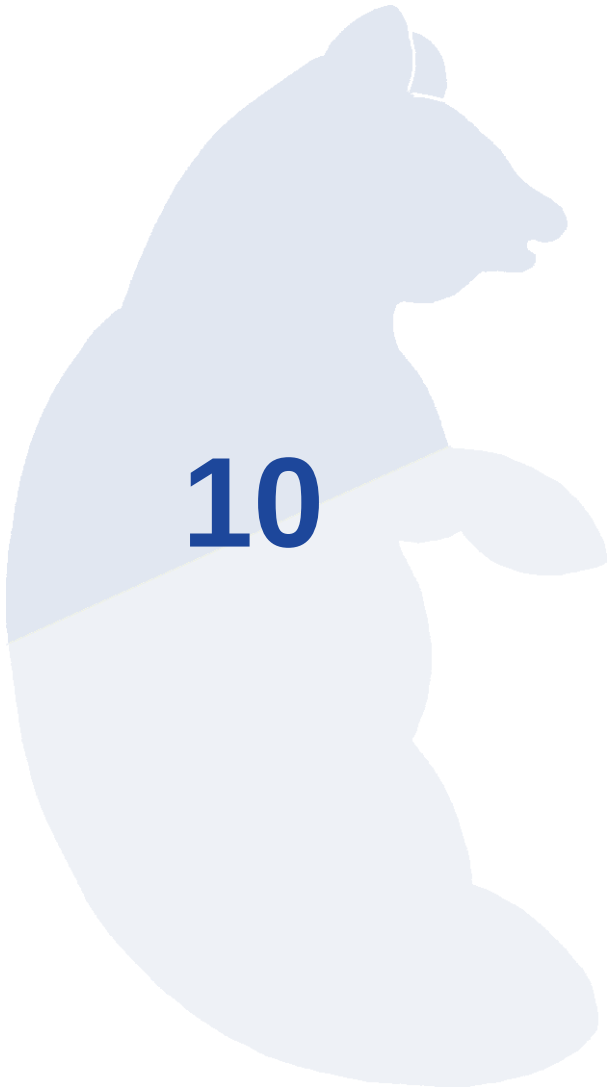


10

GMM

10. GMM

GaussianMixture : un modèle de mélange gaussien est un modèle probabiliste qui suppose que tous les points de données sont générés à partir d'un mélange d'un nombre fini de distributions gaussiennes avec des paramètres inconnus.
On peut considérer que les modèles de mélange généralisent le regroupement par k-means pour intégrer des informations sur la structure de covariance des données ainsi que sur les centres des gaussiennes latentes.

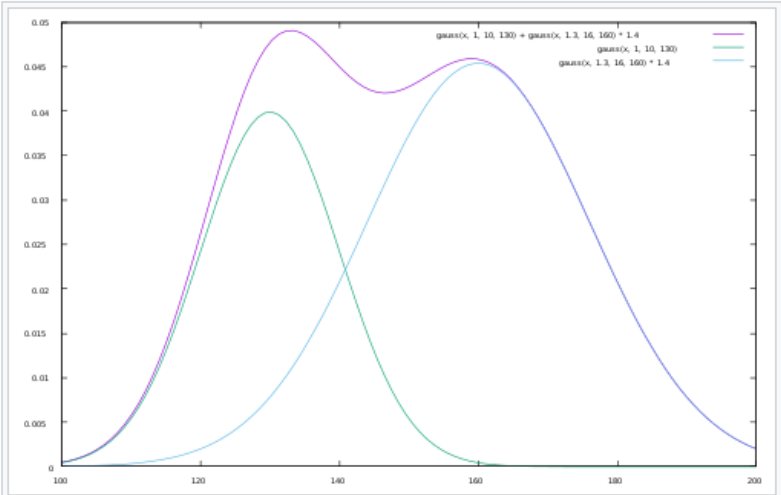
L'objet GaussianMixture met en œuvre l'algorithme d'espérance-maximisation (EM) pour l'ajustement des modèles de mélange de gaussiens. Il peut également dessiner des ellipsoïdes de confiance pour les modèles multivariés et calculer le critère d'information bayésien pour évaluer le nombre de groupes dans les données.
Une méthode GaussianMixture.fit est fournie pour apprendre un modèle de mélange gaussien à partir des données d'entraînement. Étant donné les données de test, il peut attribuer à chaque échantillon le modèle de mélange gaussien auquel il appartient le plus probablement en utilisant la méthode GaussianMixture.predict.

Le GaussianMixture est fourni avec différentes options pour contraindre la covariance des classes de différence estimées : sphérique, diagonale, liée ou covariance complète.

Source

10. GMM

Un **modèle de mélange gaussien** (usuellement abrégé par l'acronyme anglais **GMM** pour *Gaussian Mixture Model*) est un modèle statistique exprimé selon une densité mélange. Il sert usuellement à estimer paramétriquement la distribution de variables aléatoires en les modélisant comme une somme de plusieurs gaussiennes (appelées *noyaux*). Il s'agit alors de déterminer la variance, la moyenne et l'amplitude de chaque gaussienne. Ces paramètres sont optimisés selon un critère de maximum de vraisemblance pour approcher le plus possible la distribution recherchée. Cette procédure se fait le plus souvent itérativement via l'algorithme espérance-maximisation (*EM*).
Les modèles de mélange gaussien sont réputés pour reconstruire de manière particulièrement efficace les données manquantes dans un jeu de données expérimentales.



Deux gaussiennes forment une distribution bimodale

10. GMM

Dans sa forme la plus simple, le GMM est également un type d'algorithme de regroupement.
Comme son nom l'indique, chaque cluster est modélisé selon une distribution gaussienne différente.
Cette approche flexible et probabiliste de la modélisation des données signifie qu'au lieu d'avoir des affectations dures dans les clusters comme les k-means, nous avons des affectations souples.
Cela signifie que chaque point de données pourrait avoir été généré par n'importe laquelle des distributions avec une probabilité correspondante.
En fait, chaque distribution a une certaine "responsabilité" dans la génération d'un point de données particulier.

Comment pouvons-nous estimer ce type de modèle ?
Eh bien, une chose que nous pouvons faire est d'introduire une variable latente γ (gamma) pour chaque point de données.
Cela suppose que chaque point de données a été généré en utilisant certaines informations sur la variable latente γ.
En d'autres termes, cela nous indique quelle gaussienne a généré un point de données particulier.
En pratique, cependant, nous n'observons pas ces variables latentes et nous devons donc les estimer.

Comment s'y prendre ?

Heureusement pour nous, il existe déjà un algorithme à utiliser dans ce genre de cas, l'algorithme de maximisation de l'espérance (EM), que nous allons aborder.

10. GMM

L'algorithme EM se compose de deux étapes, une étape E ou étape d'espérance et une étape M ou étape de maximisation. Disons que nous avons des variables latentes γ (qui ne sont pas observées et sont désignées par le vecteur Z ci-dessous) et nos points de données X. Notre objectif est de maximiser la vraisemblance marginale de X étant donné nos paramètres (désignés par le vecteur θ). Essentiellement, nous pouvons trouver la distribution marginale comme étant le joint de X et Z et faire la somme de tous les Z (règle de la somme des probabilités).

$$ln\, p(X|\Theta) = ln\left\{\Sigma_z p(X, Z|\Theta)\right\}$$

Équation 1 : Vraisemblance marginale avec variables latentes

L'équation ci-dessus donne souvent lieu à une fonction compliquée qu'il est difficile de maximiser. Ce que nous pouvons faire dans ce cas est d'utiliser l'inégalité de Jensen pour construire une fonction de limite inférieure qui est beaucoup plus facile à optimiser. Si nous optimisons cette fonction en minimisant la divergence KL (écart) entre les deux distributions, nous pouvons nous rapprocher de la fonction originale. Ce processus est illustré dans la figure 1 ci-dessous. J'ai également fourni un lien vidéo ci-dessus qui montre une dérivation de la divergence KL pour ceux d'entre vous qui souhaitent une explication mathématique plus rigoureuse.

Pour estimer notre modèle, il suffit essentiellement d'effectuer deux étapes. Dans la première étape (étape E), nous voulons estimer la distribution postérieure de nos variables latentes γ conditionnellement à nos poids (π), nos moyennes (μ) et notre covariance (Σ) de nos gaussiennes. Le vecteur des paramètres est désigné par θ dans la figure 1. L'estimation de l'étape E nécessite d'abord d'initialiser ces valeurs et nous pouvons le faire avec les k-means qui sont généralement un bon point de départ (plus d'informations à ce sujet dans le code ci-dessous). Nous pouvons ensuite passer à la deuxième étape (étape M) et utiliser γ pour maximiser la vraisemblance par rapport à nos paramètres θ. Ce processus est répété jusqu'à ce que l'algorithme converge (la fonction de perte ne change pas).

10. GMM

Hyperparamètres :

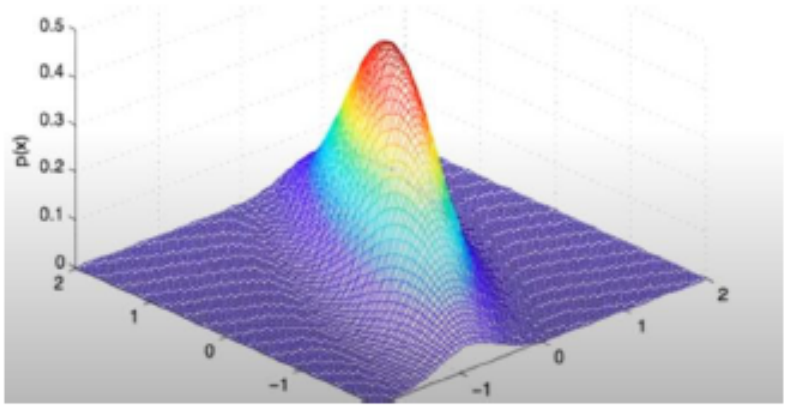
•**n_composants** : par défaut=1, le nombre de composants du mélange.
•**covariance_type** : {'full', 'tied', 'diag', 'spherical'}, default='full'. Chaîne décrivant le type de paramètres de covariance à utiliser. Doit être l'un de :

- full : (complet) chaque composant a sa propre matrice de covariance générale.
- tied : tous les composants partagent la même matrice de covariance générale.
- diag : chaque composant a sa propre matrice de covariance diagonale.
- spherical : chaque composante a sa propre variance unique.

10. GMM

Gaussian Mixture Model – Principe:

Méthode assez similaire au k-means, à la différence qu'ici on va considérer que les clusters sont distribués en suivant une loi normale (gaussiennes)
=> On ne va donc plus construire nos clusters via la distance euclidienne des points au centroïde mais en se basant sur le principe de maximum de vraisemblance



Grâce à cette méthode, on ne se limite donc plus à des clusters sphériques !

10. GMM

10. GMM