Mélanges Gaussiens

Ou le Gaussian mixture model (GMM)

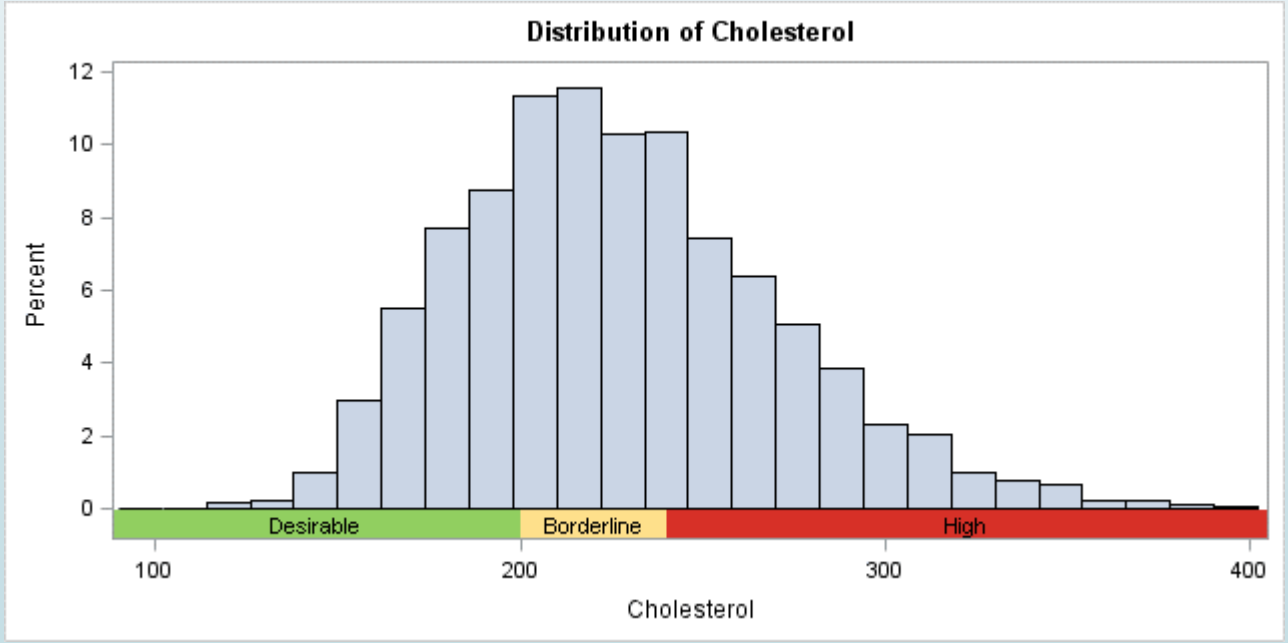
Rafik LACHAAL

Généralité

La distribution gaussienne (ou normale):

On dira d'une variable continue qu'elle a une distribution normale lorsque cette distribution prend la forme d'une cloche symétrique centré sur sa moyenne.

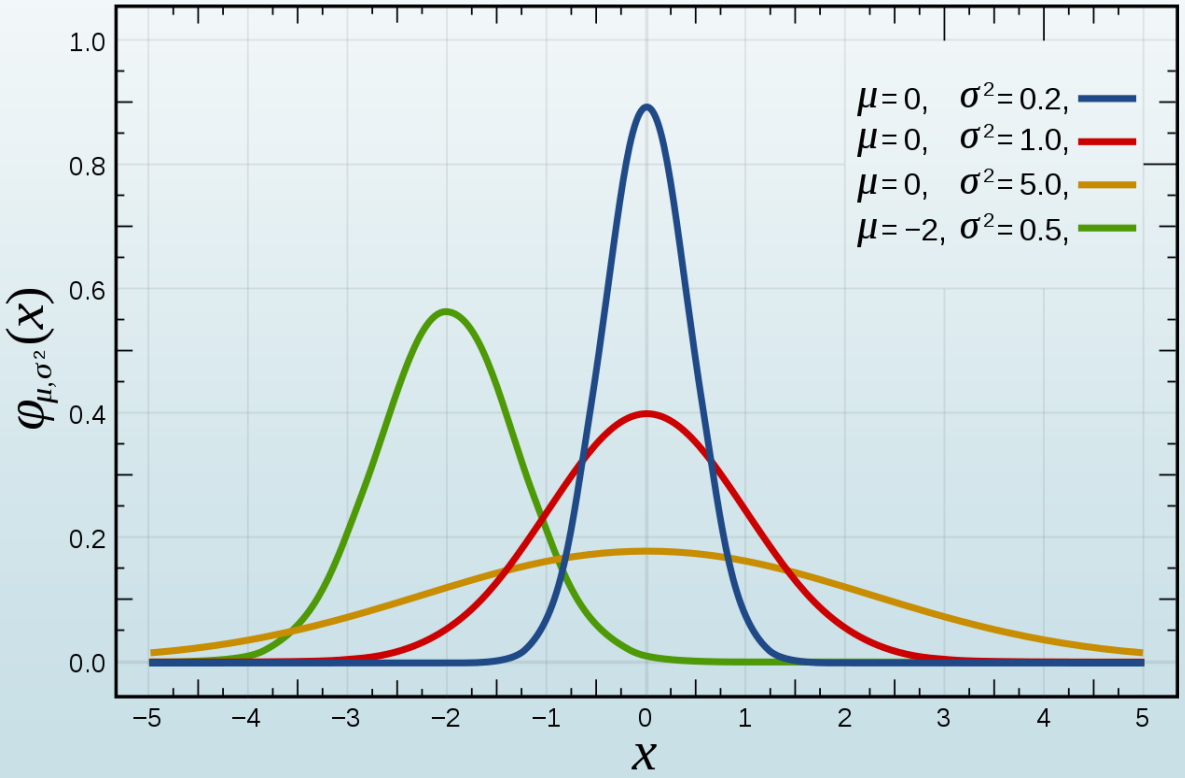
Plus on s'écarte de la moyenne et moins on à d'observations prenant ces valeurs.



Généralité

Toute les distributions normales sont caractérisés par deux valeurs :

- L'Esperance μ (que l'on peut assimiler à la moyenne)
- La variance σ^2 qui représente la dispersion autour de la moyenne



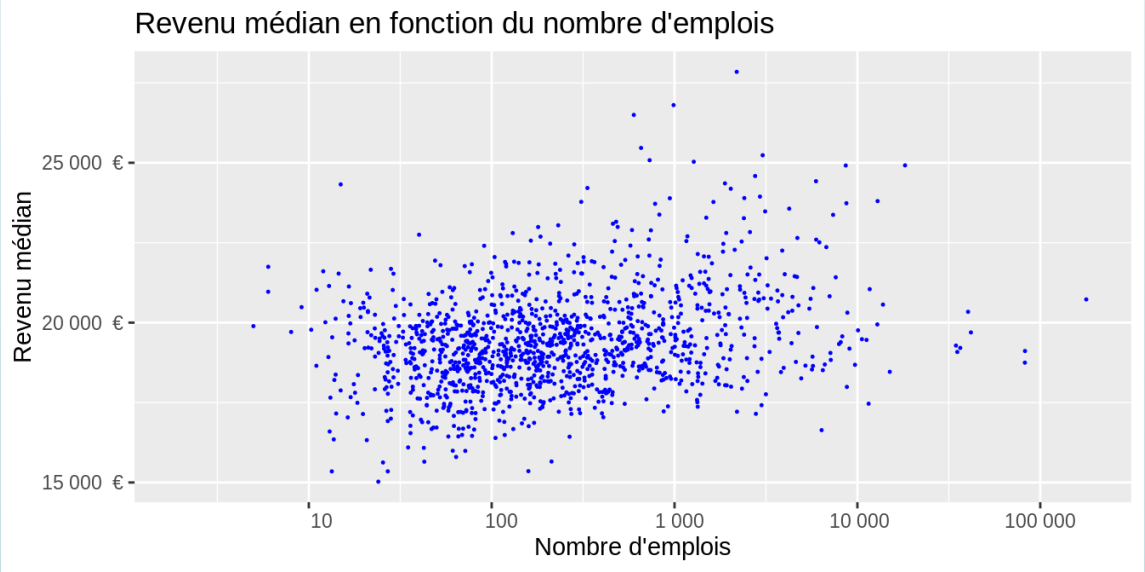
Généralité

On peut généraliser cette distribution normale avec des observation non plus décrite par seulement une variable (cholestérol) mais par plusieurs variables (cholestérol, ferritine, taux de globule rouge, …).

C'est ce qu'on appelle une distribution **Normale multivariée** ou **multinormale** ou **de Gauss à plusieurs variables.**

Toutes les distribution multinormale sont caractérisé par :

- Un vecteur μ représentant sont centre
- Une matrice de variance – covariance représentant ses dispersions Σ.



Le mélange gaussien

Le modèle de mélange gaussien est un modèle probabiliste suppose que :

- Les observations sont issue d'un mélange de plusieurs distributions gaussiennes dont les paramètres (espérance et variance) sont inconnue.
- Toute les observations ayant une même distribution gaussienne forme un cluster ayant typiquement une forme ellipsoïdale.
- L'ellipsoïde de chacun des clusters peut être de forme, de taille, de densité ou d'orientation différentes.

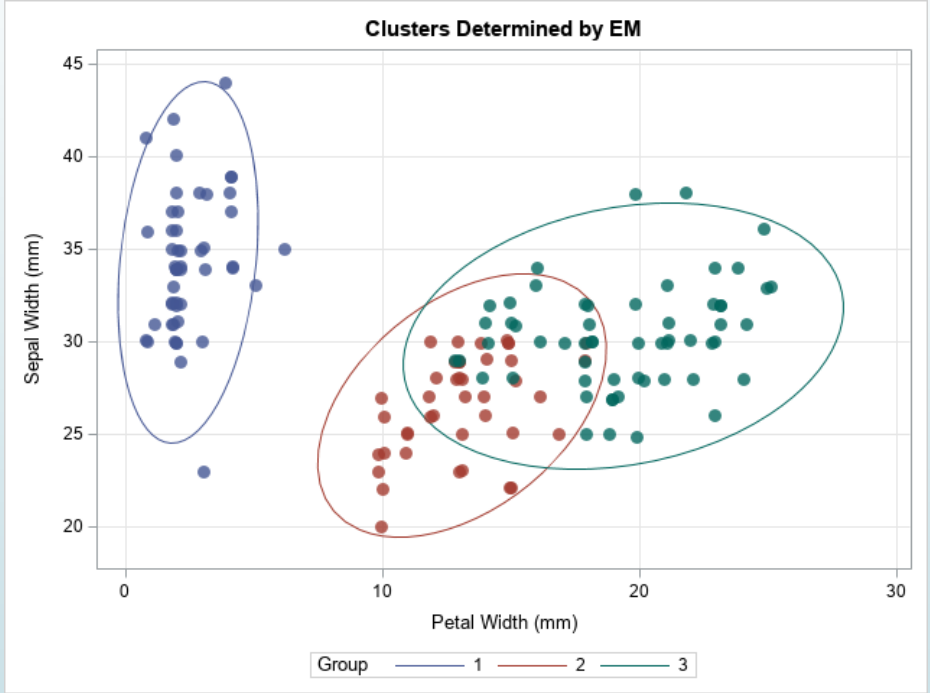
Pour chacune des observations de notre jeux de données on va supposer qu'elle est issue de l'une des distributions gaussiennes (dont on ignore les paramètres).

Déterminer à quel cluster appartient notre observation revient à déterminer à quel distribution elle appartient.

Le mélange gaussien

Les questions que l'on se pose :

- à quel distribution appartient notre observation ?
- Combien y a-t-il de clusters (de distribution gaussienne différentes) dans mon jeu de données ?



Le mélange gaussien

Sont fonctionnement :

- On lui fourni en entré un nombre **k de cluster**.
- A l'aide de l'algorithme **espérance-maximisation, EM,** (fonctionnement proche de celui des kmeans) le GaussianMixture de Sklearn va calculer les vecteurs des espérances et les matrice de variances-covariances de chacun des clusters supposés.
- Puis en se basant sur le **théorème de bayes** on va calculer pour chacune des observation (x) la probabilité d'appartenir à chacune des distributions (G_k) sachant les vecteurs des espérances et les matrice de variances-covariances .

$$\mathbb{P}(\boldsymbol{x} \in G_k | \boldsymbol{x}) = \frac{\mathbb{P}(\boldsymbol{x} | \boldsymbol{x} \in G_k) \,.\, \mathbb{P}(\boldsymbol{x} \in G_k)}{\mathbb{P}(x)}$$

- Enfin on affectera l'observation à la classe dont la distribution normale est la plus probable.

Le mélange gaussien

- Se modèle sera efficace si nos variables on réellement une distribution gaussienne (ce qui arrive assez souvent)
- Ne fonctionne ni pour des variables catégoriel même avec un tableau de confusion, ni pour des variables binaire.
- Tout comme les kmeans l'initialisation de l'algorithme **espérance-maximisation** se fait de manière aléatoire. Il convient donc de relancer plusieurs fois l'apprentissage pour voire si les résultats sont stable.
- Lorsqu'il y a trop de cluster ou peu d'observation l'EM peut avoir du mal à converger.
- Dans l'hyperparamètre « covariance_type » on peut lui spécifier la forme que peuvent prendre les clusters (va mettre des contrainte sur la matrice de variance-covariance).
  - **« sphérical »** → tout les clusters sont sphérique mais avec différentes tailles.
  - **« diag »** → forme ellipsoïdale de n'importe quel tailles, mais les axes de l'éllipsoïde sont parrallèles aux axes des coordonnées.
  - **« tied »** (lié) : tous les clusters doivent avoir la même forme ellipsoïdale, être de même taille et la même orientation.