ENSIAS

Machine Learning Project

Présentée par :

Manal NHILI

Assistée par :

Pr. Abdelhak MAHMOUDI

Plan

1. Random Forest
2. Gaussian Mixture

Part 1

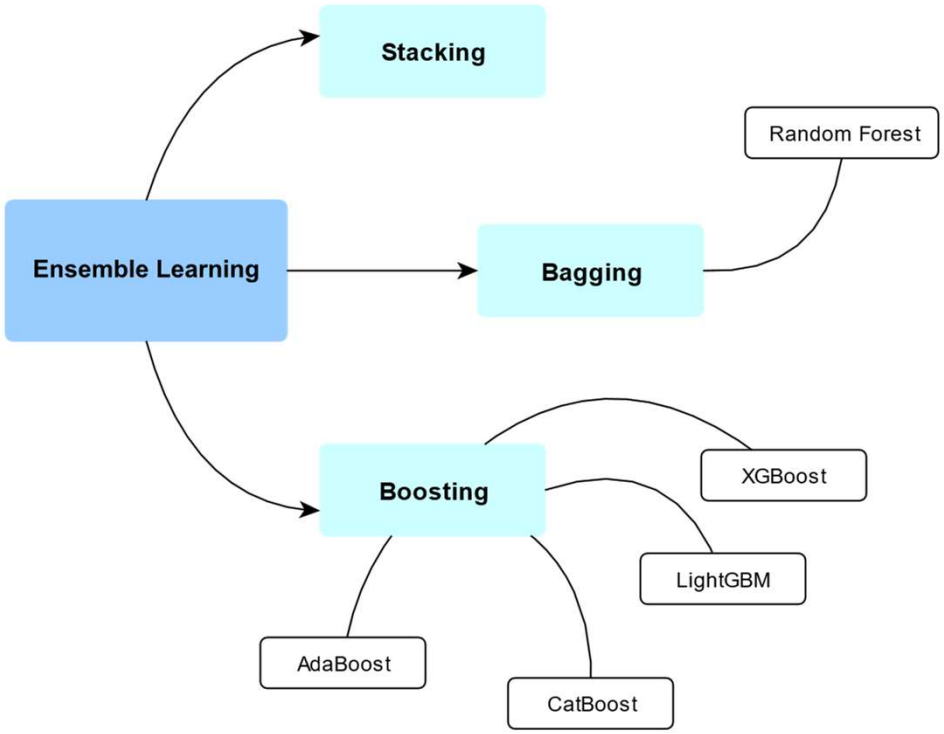
Random Forest

Re-look at Decision tree problem

Comment peut-on atteindre un équilibre entre le biais et la variance ?

Ensemble Learning

What is Random Forest ?



Bootstrapping

Le **bootstrapping** est une méthode d'estimation de la distribution d'échantillonnage d'un estimateur par rééchantillonnage avec remplacement à partir de l'échantillon d'origine.

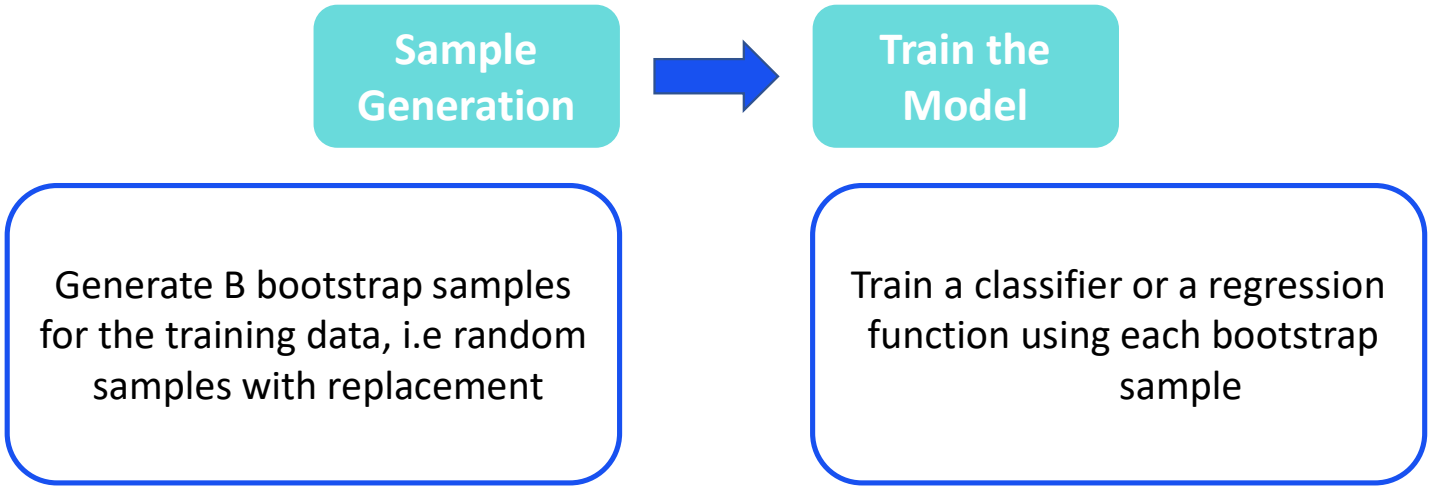
- La méthode est particulièrement utile dans les situations où la distribution d'échantillonnage de l'estimateur n'est pas une distribution standard.
- La méthode peut être utilisée dans n'importe quel problème d'inférence statistique.
- L'utilisation du terme "bootstrap" vient de l'expression “To pull oneself up by one's bootstraps”- généralement interprétée comme réussir malgré des ressources limitées.

Bagging

- Le terme "Bagging" a été introduit par Breiman (1996).
- "Bagging" signifie "Bootstrap Aggregating".
- Il s'agit d'une méthode d'ensemble : une méthode de combinaison des résultats de plusieurs rééchantillonnages.
- La méthode d'ensemble peut également être appliquée en utilisant différents classificateurs pour un échantillon donné.

Bagging Method Framework

Bagging has two basic steps:



- Model for **classification** → Majority vote on the classification
- Model for **regression** → Average of the predicted value

Bagging improves performance for unstable classifiers which vary significantly with small changes in the data set

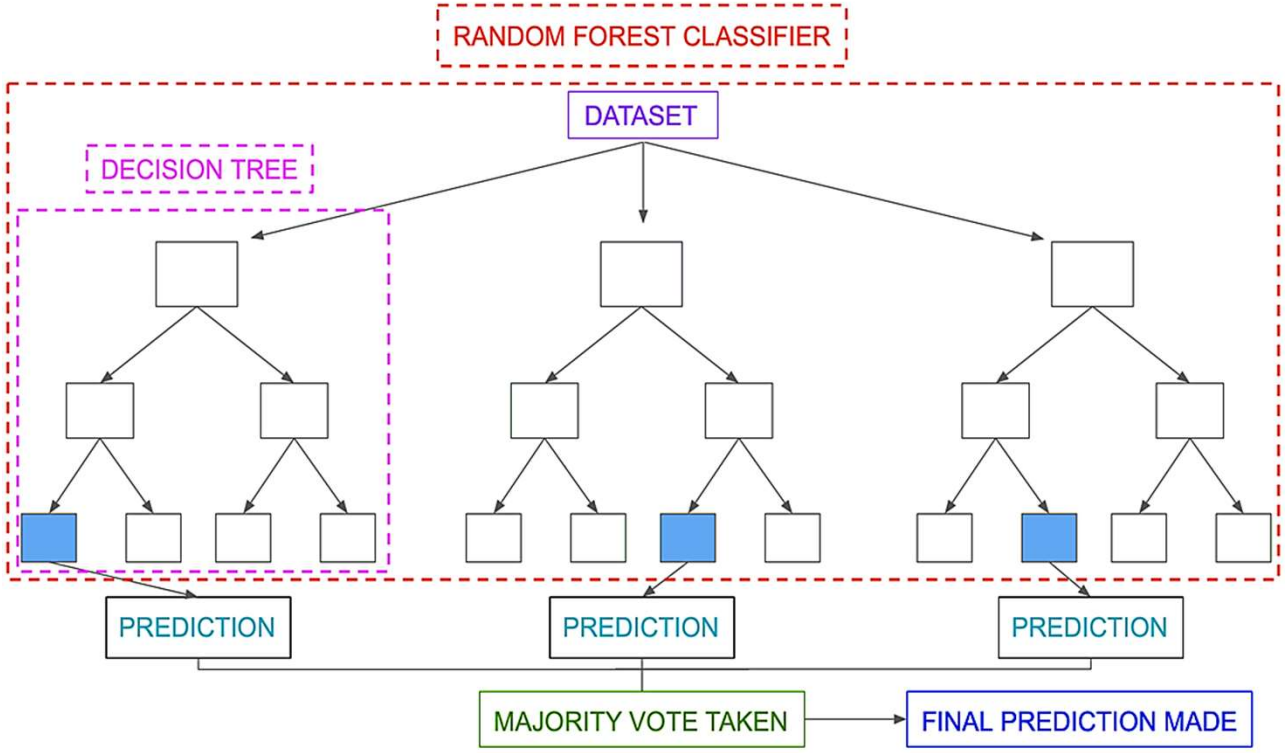
Random forest Classifier

Random Forest (RF) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.

Random Forest is Bagging of Decision Trees

- The method combines Breiman's "Bagging" idea and the random selection of features.

Random Forest Algorithme



In what tasks it is used ?

- Classification
- Regression
- Feature Selection
- Outlier / Anomaly Search
- Clustering

La formule du classificateur final est :

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x)$$

Où

N : the number of trees;

i : the counter for trees;

b : the decision tree;

x : a sample generated by us based on the data.

Part 2

Gaussian Mixture Model aka (GMM)

What is a gaussian ?

- Une distribution gaussienne multivariée d'un vecteur $x = (x_1, x_2, \dots, x_n)$ à n dimensions peut s'écrire :

$$N(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)\right)$$

Où μ est le vecteur moyenne à n dimension et Σ est la matrice de covariance $n \times n$

What is a Mixture Model ?

- Mixture Model : un modèle de mélange est la somme pondérée d'un nombre de densités de probabilité où les poids sont déterminés par une distribution π.

$$p(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \cdots + \pi_K f_K(x) \quad \text{avec} \quad \sum_{k=1}^{K} \pi_k = 1$$

$$p(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

What is a Gaussian Mixture Model ?

- GMM : somme pondérée d'un nombre de gaussiens où les poids sont déterminés par une distribution π.

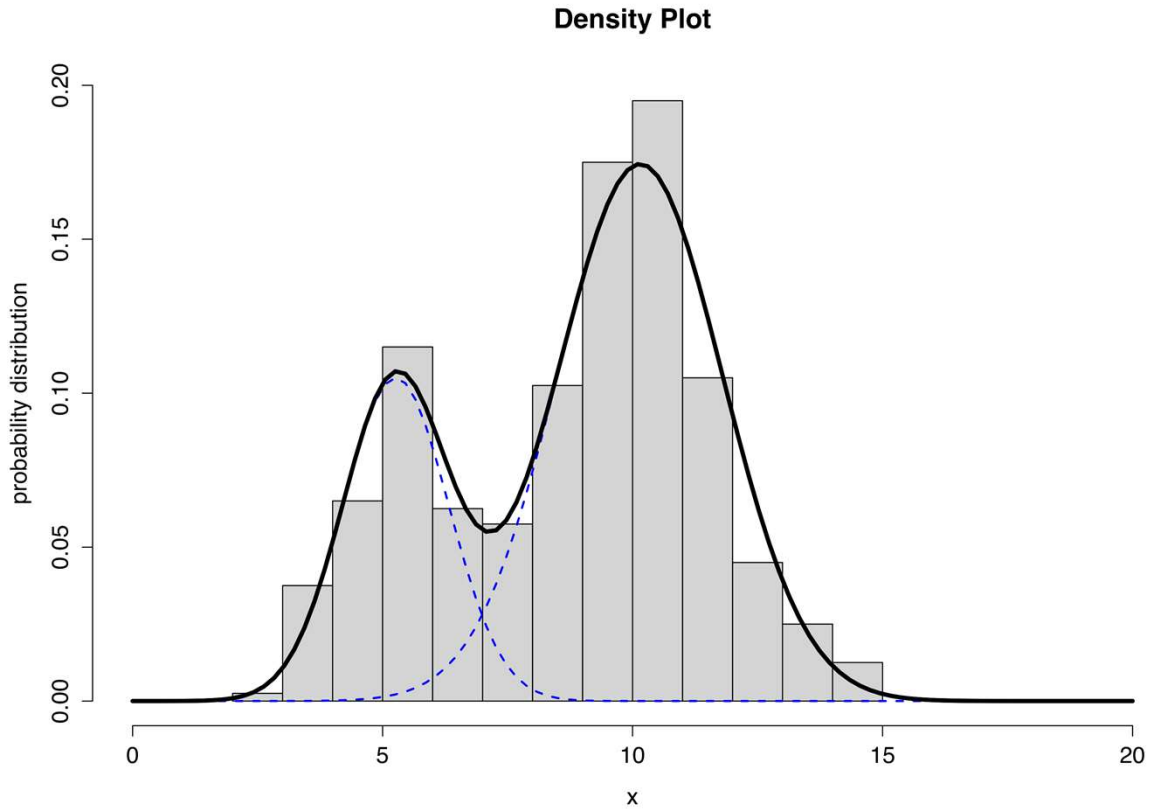
$$p(x) = \pi_1 N(x|\mu_1, \Sigma_1) + \pi_2 N(x|\mu_2, \Sigma_2) + \cdots + \pi_n N(x|\mu_K, \Sigma_K)$$

Avec $\sum_{k=1}^{K} \pi_k = 1$ et $0 \leqslant \pi_k \leqslant 1$

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

Exemple de GMM (K = 2)

Exemple de mélange Gaussien dans le cas unidimensionnel montrant deux Gaussiennes en bleu et leur somme en noir.



Parameter Estimation

- Pour un GMM avec k composantes, sur des données à n-dimension, les paramètres $\Theta = \{\pi_k, \mu_k, \Sigma_k\}$ à estimer:
  - k coefficients de mélange (*mixing coefficients*)
  - k vecteurs moyenne de n - dimension
  - k matrices de covariance $n \times n$
- Likelihood :

$$p(x|\Theta) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \right)$$

Parameter Estimation 2

- Log Likelihood:

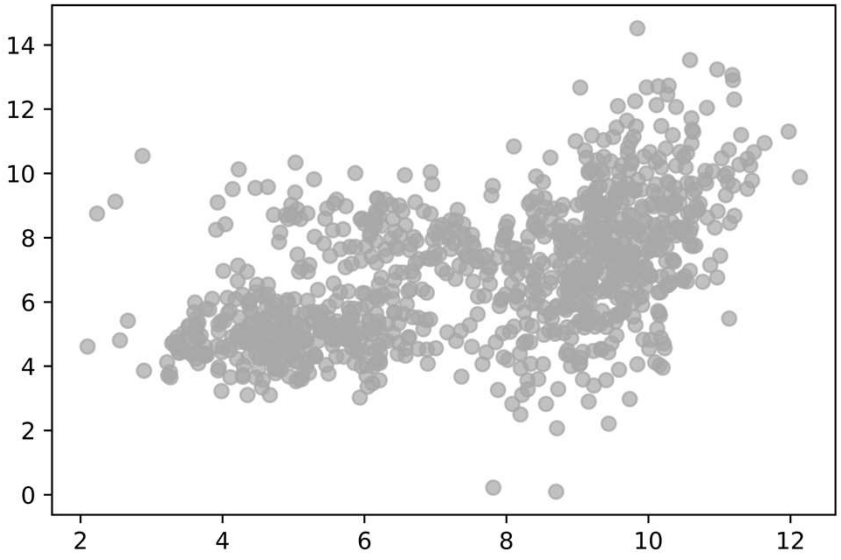
$$\log p(x|\Theta) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \right)$$

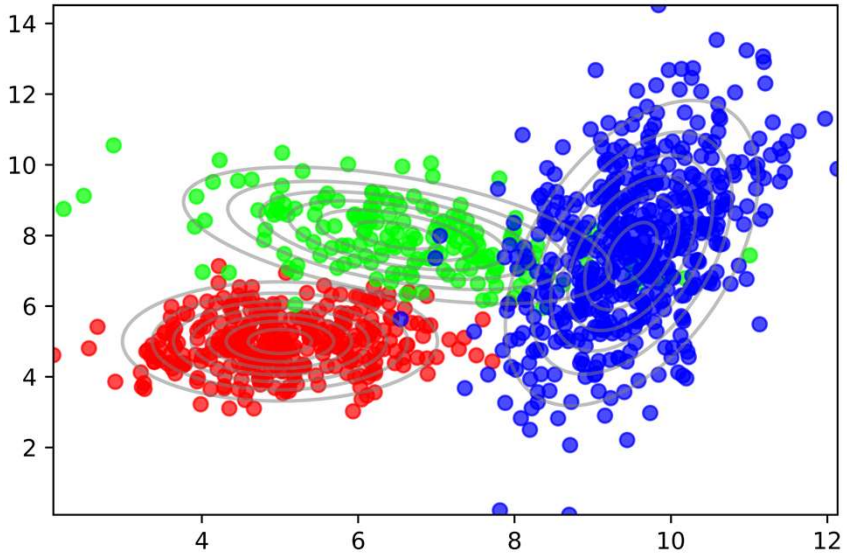
- Parameter estimation:

$$\Theta_{MLE} = \underset{\Theta}{\operatorname{argmax}}\, log(p(x \mid \Theta))$$

Latent variable

- Une *variable latente* est une variable qui ne peut pas être observée directement mais qui peut être déduite des variables et paramètres observés.



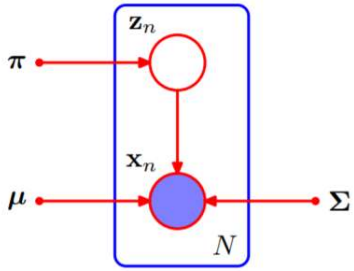


Latent Variable Representation

- On peut représenter un GMM en introduisant une variable latente discrète z.

$$p(x|\Theta) = \sum_{k=1}^{K} p(z_k = 1 \mid \Theta) p(x \mid z_k = 1, \Theta)$$

$$p(x|\Theta) = \sum_{k=1}^{K} \pi_k \, N(x \mid \mu_k, \Sigma_k)$$



- La *responsabilité* que prend un composant de mélange pour expliquer une observation x.

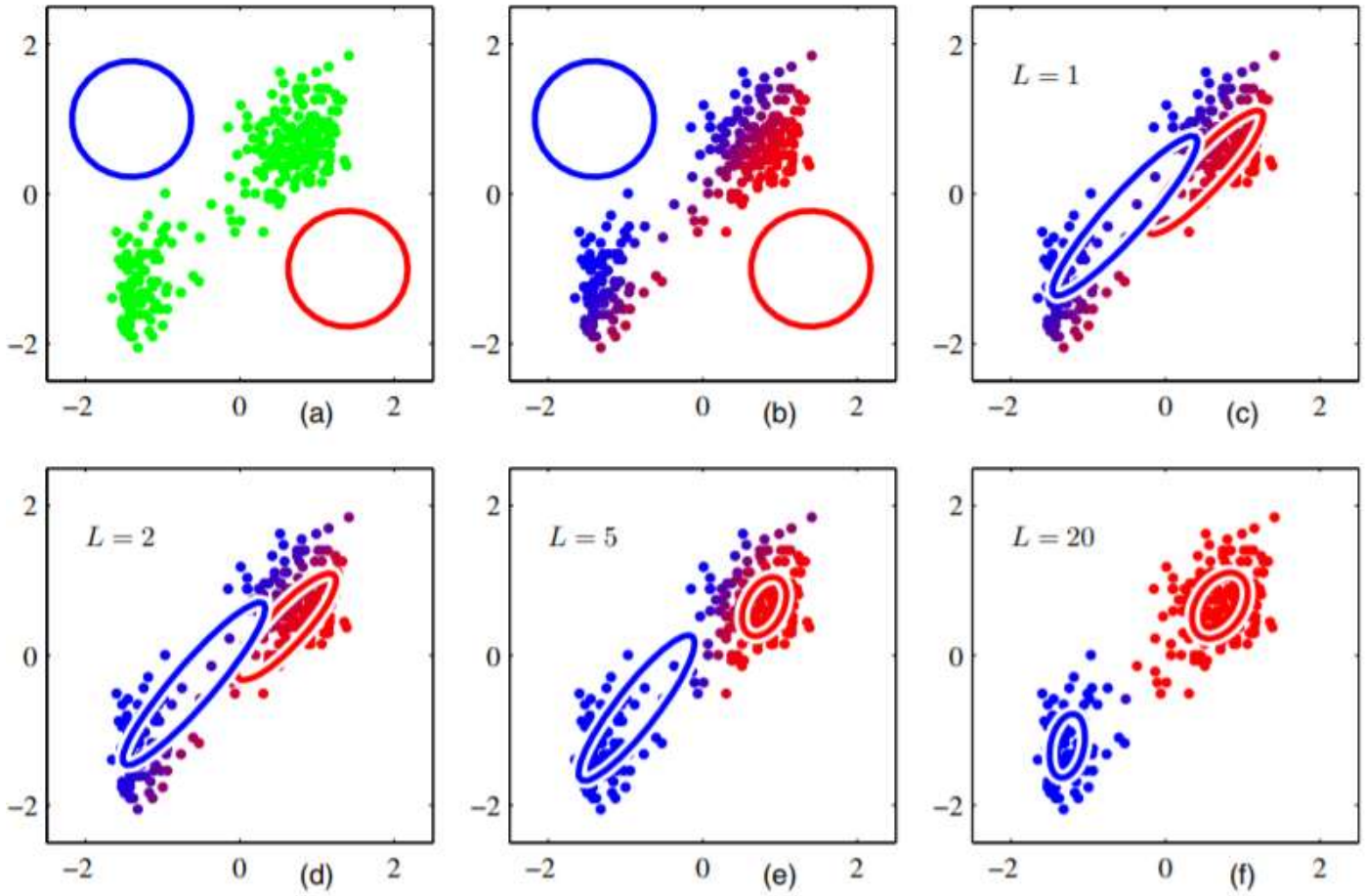
$$\tau(z_k) = p(z_k = 1|x, \Theta)$$

Expectation Maximization

- L'algorithme d'entraînement des GMMs avec des variables latentes peut être réalisé en utilisant Expectation-Maximization.

1. Initialiser les paramètres et évaluer la valeur initiale du log likelihood.
2. E-step : Évaluer les « responsabilités » pour chaque cluster avec les paramètres courant.
3. M-step : Réestimer les paramètres en utilisant les « responsabilités » existantes.
4. Évaluer le log likelihood et vérifier la convergence.

Visual example of EM



In what tasks it is used ?

- Clustering (Soft Clustering)
- Density Estimation
- Outliers detection

Thank You !

Références

1. 1. Education, I. C. (2021, January 26). Random Forest. IBM. Retrieved from https://www.ibm.com/cloud/learn/random-forest
2. 2. RandomForest Machine Learning – Oracle Machine Learning (OML). (2020, June 24). Oralytics. Retrieved from https://oralytics.com/2020/06/24/randomforest-machine-learning-oracle-machine-learning-oml/
3. 3. Ampadu, H. (2021, May 10). Random Forests Understanding. Ai-Pool. https://ai-pool.com/a/s/random-forests-understanding
4. 4. Gaussian Mixture Model | Brilliant Math & Science Wiki. (n.d.). Brilliant.Org. https://brilliant.org/wiki/gaussian-mixture-model/
5. 5. Ghassany, M. (2021, October 18). 8 Gaussian Mixture Models & EM | Machine Learning. Mghassny.Com. https://www.mghassany.com/MLcourse/gaussian-mixture-models-em.html#the-gaussian-distribution
6. 6. Team, D. (2021, March 8). Gaussian Mixture Model with Case Study – A Survival Guide for Beginners. DataFlair. https://data-flair.training/blogs/gaussian-mixture-model/
7. 7. Gupta, P., & Sehgal, N. K. (2021). Introduction to Machine Learning in the Cloud with Python: Concepts and Practices (1st ed. 2021 ed.). Springer. https://doi.org/10.1007/978-3-030-71270-9
8. 8. Vannieuwenh, A. (2019). Intelligence artificielle vulgarisée - Le Machine Learning et le Deep Learning par la pratique (French Edition) [E-book]. ENI.
9. Christopher M. Bishop. Pattern Recognition and Machine Learning (PDF), Chapter 2,9.
10. Kevin P. Murphy. Machine Learning, A Probabilistic Perspective, Chapter 11.