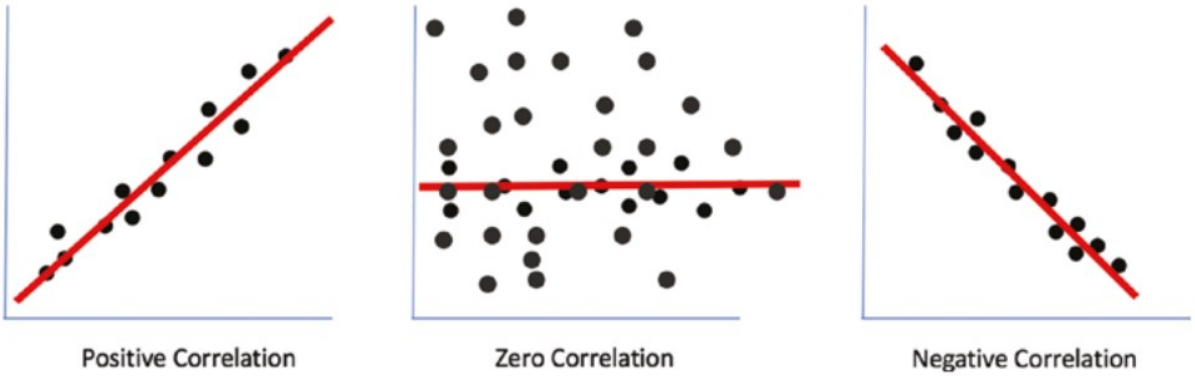
POINTS IMPORTANT DU CHAPITRE 5 : MLLIB : MACHINE

- la bibliothèque d'apprentissage automatique de Spark (Mllib) a une capacité à former des modèles à grande échelle et à fournir une formation distribuée.
- la permet aux utilisateurs de créer rapidement des modèles sur un énorme ensemble de données, en plus de prétraiter et de préparer des flux de travail avec le framework Spark lui-même.
- La corrélation est une mesure importante permettant de déterminer s'il existe une relation entre deux variables continues. Elle peut être positive ou négative ou tout simplement qu'il n'y ait pas de corrélation entre deux variables.



- La corrélation concerne la relation entre les caractéristiques numériques, alors que d'autres types de variables peuvent également être catégorielles. L'un des moyens de valider la relation entre deux variables catégorielles consiste à utiliser un test du chi carré.
- Nous pouvons convertir la variable numérique/continue en caractéristiques catégorielles (0/1) en utilisant Binarizer dans Mllib.
- Nous devons déclarer la valeur seuil, afin de convertir la caractéristique numérique en une caractéristique binaire.
- L'analyse en composantes principales (ACP) est l'une des techniques de transformation qui permet de réduire les dimensions des données tout en gardant intacte au maximum la variation des données.
- La normalisation fait référence à la transformation des données de manière à ce que les nouvelles données normalisées aient une moyenne de 0 et un écart type de 1. La normalisation se fait à l'aide de la formule suivante :

$$\frac{\left(x - mean(x)\right)}{standard\ dev(x)}$$

- La normalisation permet de standardiser les données d'entrée et parfois d'améliorer les performances des modèles d'apprentissage automatique.
- La mise à l'échelle est une autre technique pour normaliser les données, de sorte que les valeurs se situent dans une plage spécifique. Sa formule est :

$$\frac{x - min(x)}{max(x) - min(x)}$$

- La mise à l'échelle min-max est une autre version de la mise à l'échelle standard, car elle vous permet de redimensionner les valeurs des caractéristiques entre des limites spécifiques (généralement entre 0 et 1).
- MaxAbsScaler est un peu différent des outils de mise à l'échelle standard, car il redimensionne chaque valeur de caractéristique entre -1 et 1. Cependant, il ne déplace pas le centre des données et, par conséquent, n'a aucun impact sur la parcimonie.
- On faire du binning à l'aide de Bucketizer dans Spark
- utiliser la bibliothèque d'apprentissage automatique de Spark (MLlib) pour créer des modèles de classification.