Enseignantes : A. NAJJAR- I. BEN OTHMEN - F. JENHANI **TP : Z. ZOUAGHIA - I. BEN AYCHA**	**TP2** **Machine Learning**	**Classe : 3ème GLSI**

Partie 1 : Préparation des données

Plusieurs techniques de traitement des données peuvent être utilisées. Le choix de la technique appropriée dépend du contexte.

Dans cette partie, nous utilisons la base "**pima-indians-diabetes.data.csv**" manipulée lors du premier TP.

a. Gérer les données manquantes

Certaines instances peuvent avoir des attributs ayant des valeurs nulles. Ce qui peut dégrader les performances d'un algorithme d'apprentissage. La solution la plus simple est de supprimer les individus dont les valeurs de certains attributs sont manquantes. Mais, ceci peut causer la perte de données importantes. Un seconde alternative et de déterminer, la valeur médiane de chaque attribut, puis de remplacer les valeurs nulle par la médiane.

Questions

1. Lire le contenu du fichier "**pima-indians-diabetes.data.csv**" et en extraire les valeurs des attributs.
2. Filtrer les valeurs de l'attribut '**SkinThick'** pour déterminer les valeurs non nulles.
3. Calculer la médiane de ces valeurs
4. Remplacer les valeurs nulles de l'attribut "**SkinThick"** par la médiane**.**
5. Pourquoi on ne peut pas faire la même chose avec l'attribut **" NumTimesPrg"** ?

b. Uniformisation d'échelle

Le changement d'échelle est une sorte de normalisation. L'intervalle dans lequel vari les variables numériques peut être différent selon l'attribut. Ceci peut influencer les performances de certains algorithmes d'apprentissage automatique, surtout ceux qui se basent sur le calcul de distances. Pour cela, le but de cette étape est de ramener toutes les valeurs dans l'intervalle [0,1]. Une des techniques utilisées pour la normalisation est la suivante :

$$Val_{norm} = \frac{Val - Val_{min}}{Val_{max} - Val_{min}}$$

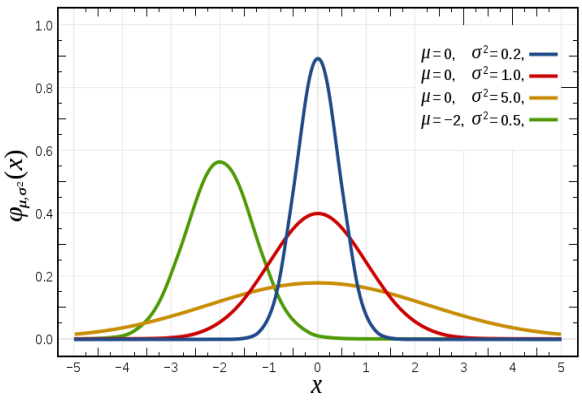
Questions

1. Ramener les valeurs de tous les attributs dans un intervalle [0-1].
2. Réafficher les données.

c. Normalisation

La standardisation des données est aussi une sorte de normalisation. Ramener les valeurs des attributs à l'intervalle [0-1] est parfois insuffisant surtouts dans le cas des bases qui contiennent beaucoup de zéros ou des algorithmes qui multiplient ces valeurs par une certaine pondération. Les valeurs d'un attribut sont transformées pour suivre une loi gaussienne ayant une moyenne **μ=0** et un écart type **σ=1**.

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$



Questions

1. Normaliser les valeurs des attributs.
2. Réafficher les données.

Partie 2 : Régression Linéaire Simple

On souhaite effectuer de la **régression linéaire simple**. Dans cette partie du TP, on se basera sur la base **"Weather.csv"** contenant des données météorologiques.

Questions

1- Importer les données à partir du fichier.
2- Afficher la dimension de l'ensemble de données et en déduire le nombre des variables utilisées.
3- Afficher les noms de ces variables.

Pour faire de la régression linéaire simple, on se restreindra à deux variables : la valeur minimale de température (**MinTemp**) et sa valeur maximale (**MaxTemp**). On considère que "**MinTemp**" est la variable explicatif et "**MaxTemp"** et la variable expliquée. Par conséquent, on souhaite prédire la valeur de "**MaxTemp"** en fonction de la valeur enregistrée de "**MinTemp**".

Questions

4- Afficher toutes les observations de la base en considérant la valeur minimale de la température sur l'axe des abscisses et sa valeur maximale sur l'axe des ordonnées.
5- Construire le nouvel ensemble de données, composé seulement des deux variables "**MinTemp"** et **"MaxTemp".**
6- Diviser l'ensemble de données en un ensemble d'apprentissage qui contient 80% des observations et un ensemble de test (20%).
7- Déterminer les paramètres **"a"** et **"b"** de la droite de régression (**y=ax+b**) en utilisant l'ensemble d'apprentissage.
8- Afficher les paramètres du modèle : la pente **"a"** et l'ordonné à l'origine **"b"**.
9- Déterminer et afficher le **coefficient de détermination R^2**. Conclure.
10- En utilisant les températures minimales de l'ensemble de test et le modèle de régression calculé dans la question précédente, prédire les températures maximales.
11- Afficher, sur la même figure, les valeurs réelles observées et celles prédites des températures maximales en fonction des températures