Eppur si muove !
Institut Galilée

UNIVERSITÉ PARIS 13

Master Informatique Specialités EID², PLS

Traitement numérique des données

TP2
Prétraitement et visualisation de données

Les TP sont notés. Les étudiants doivent sauvegarder le code de chaque exercice et doivent présenter le résultat au chargé du TP. Répondez à toutes les questions et notez la réponse dans un fichier.

Le package **sklearn.preprocessing** offre plusieurs fonctions pour la transformations de données, i.e. changer les caractéristiques initiales de données en une représentation qui est plus approprié pour le traitement de ces données. Ce traitement est souvent nécessaire pour les données en grande dimension.

A. Normalisation de données

La normalisation de données est une étape importante dans le processus de traitement de données. Par exemple, de nombreux éléments utilisés dans la fonction objective d'un algorithme d'apprentissage (tels que le noyau RBF de Support Vector Machines ou la L1 et L2 régularisé des modèles linéaires) supposent que toutes les variables sont centrées autour de zéro et ont la variance dans le même ordre. Si une caractéristique a une variance qui est des ordres de grandeur plus grand que les autres, il pourrait dominer la fonction objectif et de faire l'estimateur incapable d'apprendre correctement comme prévu.
En pratique, nous ignorons souvent la forme de la distribution de données et on simplement transforme les données en les centrent en retirant la valeur moyenne de chaque variable, puis en divisant les variables par leur écart-type.

Importez les librairies **numpy** (calcul scientifique) et preprocessing (prétraitement de données)

1- Créez la matrice **X** suivante :

```
1, -1,  2,
2,  0,  0,
0,  1, -1
```

2- Visualisez X et calculez la moyenne et la variance de X.

3- Utilisez la fonction **scale** pour normaliser la matrice X. Que constatez vous ?

4- Calculer la moyenne et la variance de la matrice X normalisé. Expliquez le résultat obtenu.

B. Normalisation MinMax

Un autre type de normalisation est de normaliser les caractéristiques (variables) de données entre un minimum et une valeur maximale donnée, souvent entre zéro et un. Ceci peut être réalisé en utilisant la fonction **MinMaxScaler**.

1- Créez la matrice de données **X2** suivante :

```
1, -1,  2,
2,  0,  0,
0,  1, -1
```

2- Visualisez la matrice et calculez la moyenne sur les variables.
3- Normalisez les données dans l'intervalle [0 1]. Visualisez les données normalisées et calculez la moyenne sur les variables. Que constatez-vous ?

C. visualisation de données

1- Chargez les données Iris

2- Visualisez le nuage de points en 2D avec des couleurs correspondant aux classes en utilisant toutes les combinaisons de variables. Quelle est la meilleure visualisation ? Justifiez votre réponse.

D. Réduction de dimensions et visualisation de données

L' Analyse en Composantes Principales (ACP) a comme objectif d'identifier la combinaison d'attributs (composants principaux, ou les directions dans l'espace de caractéristique), qui représentent le plus la variance dans les données.
L'Analyse discriminante linéaire (ADL) tente d'identifier les attributs qui représentent le plus la variance entre les classes. En particulier, l'ADL, contrairement à l'APC, est un procédé supervisé en utilisant les étiquettes de classe connus.

1- Les méthodes **PCA** et **LDA** peuvent etre importé à partir des package suivants :

```
import from sklearn.decomposition
import PCA from sklearn.lda import LDA
```

2- Analysez le manuel d'aide pour ces deux fonctions (**pca** et **lda**) et appliquez les sur la base Iris. Il faudra utiliser **pca.fit(Iris).transform(Iris)** et sauvegardez les résultats dans IrisPCA pour la PCA et IrisLDA pour la LDA.

3- Visualisez les nuages de points avec les nouvelles axes obtenus : une image pour l'ACP et une autre pour l'ADL et utiliser la classe de Iris comme couleurs de points. Quelle différence constatez-vous entre les deux visualisations? Expliquer votre raisonnement.