DM

November 21, 2024

Université Paul Sabatier, M2 SID

1 Apprentissage Statistique Avancé

Un jeu de données est chargé dans les arrays `X,y` ci-dessous. Il s'agit d'un problème de classification binaire.

[ ]:
```
%matplotlib inline
from matplotlib import pyplot as plt
import math
import numpy as np
import scipy as sp
import sklearn as sk
```

[ ]:
```
X = np.loadtxt("X.txt")
y = np.loadtxt("y.txt")
```

[ ]:
```
n = X.shape[0]
p = X.shape[1]
print("Sample size: " + str(n))
print("Number of variables: " + str(p))
```

```
Sample size: 100
Number of variables: 10
```

1.1 Q. 1

Analyser le jeu de données `X,y`. Quelles sont ses spécificités?

[ ]:
```

```

1.2 Q. 2

Effectuez une ACP des données `X` et représentez les deux premières dimensions par un scatter plot. Utilisez des couleurs pour représenter les classes `y`. Qu'observez vous?

[ ]:
```

```

1.3 Q. 3

Reprenez la question 2 mais avec une ACP à noyau Gaussien (rbf). Voyez-vous une différence?

[ ]:

1.4 Q. 4

Au vu des questions précédentes, mettez en oeuvre une méthode de classification pour ce jeu de données. Vous règlerez ses paramètres (au moins 1) de la manière de votre choix et donnerez une estimation de l'erreur de votre modèle.

[ ]:

1.5 Q. 5

Le jeu de validation `X_val` contient 100 individus dont 90 distribuées comme `X` et 10 outliers. Donner un fichier csv `y_pred.txt` contenant vos prédictions faites à partir de `X_val` : la ligne i doit contenir -1 si `X_val[i]` est prédit comme un outlier et 0 ou 1 sinon suivant la classe prédite. *(Vous pouvez utiliser la fonction* `np.savetxt("y_pred.txt",y)`*)*

Pour vous aider, je vous donne un jeu de 40 individus `X_try`, distribués comme `X_val`, avec les vraies valeurs `y_try`.

Cette question est notée suivant votre pourcentage de bonne prédiction sur mon jeu privé `y_val`*.*

[ ]:
```
X_val = np.loadtxt("X_val.txt")
```

[ ]:
```
X_try = np.loadtxt("X_try.txt")
y_try = np.loadtxt("y_try.txt")
```