

Théorie bayésienne de la décision - L3 Info SD

TP - feuille n°3 - Recherche de la meilleure frontière de décision

1 Recherche de la frontière sur un premier jeu de données

Le but de cette partie est de déterminer, à partir des données d'apprentissage, un intervalle dans lequel rechercher la meilleure frontière de décision, c'est-à-dire la frontière qui donne le taux d'erreur le plus petit sur les données de validation.

On travaille dans cette partie sur les données du TP n°1 : `tp1_data_train.txt` et `tp1_data_valid.txt`.

- 1) On note `X_train` l'ensemble des caractéristiques des données d'apprentissage. Déterminez, à partir de `X_train`, les bornes `D_min` et `D_max` d'un intervalle dans lequel se situe forcément la meilleure frontière de décision.
- 2) On note `X_valid` et `y_valid` l'ensemble des caractéristiques et des classes des données de validation.
 - a) En balayant l'intervalle obtenu à la partie précédente, déterminer une valeur de la frontière de décision pour laquelle le taux d'erreur sur les données de validation est le plus petit possible. La valeur de la frontière trouvée devra avoir au moins 4 chiffres significatifs.
 - b) Calculez le taux d'erreur et la matrice de confusion obtenus avec cette valeur de frontière.
 - c) Comparez avec les valeurs obtenues au TP n°1.

2 Recherche de la frontière sur d'autres jeux de données

Vous avez à votre disposition 7 autres jeux de données (réparties en données d'apprentissage et de validation) :

- `tp3_data1_train.txt` et `tp3_data1_valid.txt`
- `tp3_data2_train.txt` et `tp3_data2_valid.txt`
- ...

Pour chacun de ces jeux, déterminez une frontière de décision en utilisant la méthode précédente. On pourra ajouter de nouvelles fonctions au fichier `tools.py`.