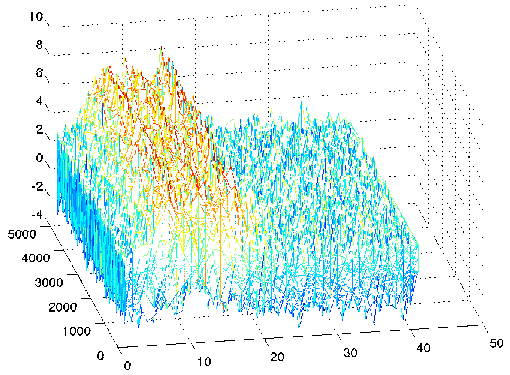
Département	**Informatique**	**Année**	**5A**
Matière	**ML_Big-Data**		
Enseignant	**Khalid Benabdeslem**		
Intitulé TD/TP :	**TP Apprentissage semi-supervisé**		
Durée	**3h**		

I. Découpage de la base en apprentissage/test

Créer un programme qui permet de découper votre base de données **X** avec un échantillonnage stratifié par rapport aux labels en deux sous-ensembles d'apprentissage **A** et de test **T** de tailles respectivement **1/2** et **1/2**.

Fichier de données à utiliser : « Wave.txt ». Cette base appelée « les vagues de Brieman » contient 5000 individus, 40 variables et 3 classes.



II. Simulation de l'aspect semi-supervisé

Ecrire un programme permettant de rendre la base **A** partiellement étiquetée avec un argument permettant de renseigner le % des données labélisées par rapport à la taille totale de **A**.

III. Sélection de variables semi-supervisée

Développer une procédure permettant de calculer la pertinence (sous forme de score) de chaque variable *v* de la base **A** comme suit :

$$\text{Score}\ (v) = S_1(v) \,/\, S_2(v) \text{ où}$$

$S_1(v)$ ne doit se calculer que sur la partie labélisée de **A**, en se basant sur le **score de Fisher** dont la formule est la suivante :

$$S_1(v) = \frac{\sum_{i=1}^{c} n_i(\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i \sigma_i^2}$$

c : le nombre de classes

n_i : l'effectif de la classe i

μ_i : la moyenne de la classe i sur la variable v

μ : la moyenne de toute la base sur la variable *v*

σ_i : l'écart-type de la classe *i* sur la variable *v*

$S_2(v)$ ne doit se calculer que sur la partie non-labélisée de **A**, en se basant sur le **score Laplacien** dont la formule est la suivante :

$$S_2(v) = \frac{\sum_{i,j}(v_i - v_j)^2 s_{ij}}{var(v)}$$

Tel que : *var*(*v*) représente la variance de la variable *v* et $S_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{t}\right)$, on prendra $t = 10$

IV. Evaluation de la sélection

Plus le score d'une variable est élevé, plus elle est pertinente. Sur ce principe, trier les variables selon leurs pertinences (décroissant).

- Tracer l'histogramme des pertinences de toutes les variables selon leurs scores
- Tracer une courbe d'efficacité (selon les performances d'un perceptron multi couches (MLP) appris sur **A**) sur la base **T**, en fonction du nombre de variables pertinentes sélectionnées (par tranche de 5 variables)
- Tracer cette même courbe avec les données non-normalisées
- Tracer la courbe sur les variables non-pertinentes
- Tracer la courbe sans sélection de variables
- En fixant le nombre de variables sélectionnées à 20, tracer la courbe de performance en fonction du % de données labélisées