Introduction au Machine Learning
TD 4. Noté. 15 mars 2018
Détection de Pulsars

`Consignes :`
Chaque étudiant envoie à *vincent.runge@univ-evry.fr* le fichier *MON_NOM.ipynb* à la fin du TP4. La deuxième partie de la note consiste à terminer ce TP en autonomie et à envoyer une version achevée à la même adresse avant le lundi 2 avril 12h.
Télécharger les fichiers `TP4_test.ipynb` et `HTRU_2.csv` à l'adresse suivante:
`https://github.com/vrunge/TP_Python`

Statistique descriptive

1) Afficher la taille du dataframe, son résumé numérique, les boxplots de chaque variable, les histogrammes et la matrice de corrélation entre les variables. Commenter brièvement les résultats obtenus.

L'objectif principal de ce TP est de déterminer le meilleur modèle SVM pour notre problème de détection de pulsars. Ce modèle sera comparé aux autres méthodes d'apprentissage vues en cours. Enfin, nous déciderons, selon un objectif préétabli, si une méthode de machine learning peut être utilisée sur ces données.

SVM

2) Séparer les données en un "*train*" et un "*test*". Utiliser les modèles **svm.SVC(kernel = 'type'), C=1)** avec type = 'linear', 'rbf' puis 'sigmoid'. La taille du *train* pourra être choisie de façon à ne pas trop attendre à l'exécution de svm.SVC. Comparer les scores obtenus.

Remarque: pour créer le *train* et le *test*, on pourra utiliser:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, ...
                    test_size=0.6, random_state=0)
```

3) Rechercher par cross-validation sur les données de *train* le paramètre C donnant le meilleur score dans le cas d'un kernel linéaire. Tracer la courbe du score en fonction de C. Donner le C optimum et le score obtenu sur les données de *test*.

4) Notre but étant d'automatiser la détection de pulsars et les pulsars étant assez rares dans nos données, on va pénaliser le fait de "rater un pulsar" (faux négatif) plutôt que d'en détecter un à tort (faux positif). Le score sera simplement le nombre de pulsars correctement détectés sur le *test* divisé par le nombre total de pulsars dans le *test* (cas extrême où on ne considère pas les faux positifs).

Répéter les questions 2 et 3 avec ce nouveau score. Comment expliquer les scores avec rbf et sigmoid?

Naive Bayes / LDA / QDA

5) Calculer le score de chacune des méthodes Naive Bayes, LDA et QDA sur les données de *test* avec

les 2 scores.

k-NN

6) Avec la méthode des k plus proches voisins, trouver le k qui donne le meilleur score sur les données de *test*. On considérera une fois encore les 2 scores possibles.
7) Répéter cette méthode après avoir centré et réduit les variables.

Régression logistique

8) Calculer les 2 scores sur les données de *test* après avoir déterminé les coefficients d'une régression logisitique sur les données de *train*.

9) De toutes les méthodes utilisées, laquelle retiendriez-vous?

10) Cette étude a été confiée à un ingénieur afin qu'il fournisse un avis d'expert. Si l'ingénieur peut exhiber une méthode de machine learning donnant moins de 10% de faux négatifs alors la méthodologie pour construire le dataset sera validée. Dans le cas contraire, il faudrait transformer la construction des covariables en collaboration avec les physiciens et spécialistes du traitement du signal. Pouvez-vous valider cette approche?