Validation croisée

I. De l'utilité de la validation croisée

La validation croisée est un algorithme utilisé par le statisticien afin d'optimiser un ou plusieurs paramètres lors de l'utilisation d'une méthode. Par exemple, il permet d'essayer de choisir le meilleur nombre de voisins dans l'algorithme des k plus proches voisins, ou de trouver le nombre d'arbres optimal dans l'algorithme de forêt aléatoire.

II. Algorithme de la validation croisée

- Déterminer les n valeurs possibles du paramètre à tester

Pour les n valeurs :

- Commencer par découper l'échantillon étudié en K parties d'effectifs sensiblement égaux (en général, K choisi entre 5 et 15)

  Pour j allant de 1 jusqu'à K :

  - Déterminer un modèle à partir des K-1 autres échantillons restant, puis tester sur l'échantillon j, qu'on appelle échantillon test
  - Calculer l'erreur
- Moyenner les erreurs obtenues

- Sélectionner la valeur du paramètre ayant obtenu l'erreur moyenne la plus faible

III. Validation croisée sur Python

Lorsque le volume de données est raisonnable, il est convenable d'utiliser la librairie sk-learn de python (le stockage des données se fait sur la RAM). Pour de plus gros volumes, il est préférable sur python de s'intéresser aux modules de Spark.

`train_test_split` : découpage en échantillons d'apprentissage

http://scikit-learn.org/stable/modules/cross_validation.html

Sources

Validation croisée et modèles statistiques appliquées, Matthieu Cornec, 04/06/2009

Nelo Magalhães. Validation croisée et pénalisation pour l'estimation de densité. Probabilités [math.PR]. Université Paris Sud - Paris XI, 2015. Français.

Validation croisée pour le choix de paramètres de méthodes de régularisation, Julien Chiquet, 23/12/2009 (donne une explication très simplifiée de la validation croisée et explique la fonction cross.validation sur R)

http://wikistat.fr/pdf/st-m-app-risque-estim.pdf