Fouille de données et medias sociaux
TP2 : sélection de modèles

Olivier Schwander <olivier.schwander@lip6.fr>

25 septembre 2017

On s'intéresse dans ce TP à au problème du choix du degré lors d'une régression polynomiale. L'objectif est de comparer différentes méthodes de sélection de modèles.

Remarque : ne pas hésiter à utiliser les fonctions `numpy.polyval` et `numpy.polyfit`.

Question 1

Chargez les données disponible à l'adresse https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/bluegills.txt.

Question 2

Utilisez tous les outils qui vous semblent utiles pour faire une première analyse de ces données

Question 3

En utilisant un découpage fixe (80% / 20%), choisissez le degré du polynôme permettant la meilleure généralisation (en utilisant la MSE sur l'ensemble de test pour mesurer la qualité).

Essayez plusieurs découpages, le résultat est-il stable ?

Question 4

Même question avec une validation croisée de type k-fold.

Question 5

Même question avec une validation croisée leave-one-out.

Question 6

Même question avec le critère AIC.

La log-vraisemblance négative du modèle sera évaluée à partir de la MSE (voir par exemple http://users.monash.edu/~dschmidt/ModelSectionTutorial1_SchmidtMakalic_2008.pdf).

Question 7

Même question avec le critère BIC.

Question 8

Comparez les “meilleurs” degrés obtenus avec ces différentes méthodes. Les réponses sont-elles cohérentes ? Que faire si ce n’est pas cohérent ?

Augmentez puis diminuez le nombre de points. Que se passe-t-il ?

Question 9

Utilisez les résultats de vos analyses pour conclure sur le vrai modèle ayant servi à générer les données.

Question 10

Quelle est l’importance du “vrai” modèle dans un projet d’apprentissage ?