

Introduction à l'apprentissage statistique

Cristina Butucea - Ensa-e-Crest

Cours : 12 heures - TP : 6 heures

Objectif

Ce cours constitue une introduction d'ensemble aux méthodes de machine learning. On introduira les problématiques typiques de description et modélisation des données dans le but de prédire au mieux la réponse d'un nouvel individu. Nous décrivons les algorithmes et quantifierons leur bon comportement et, parallèlement, à travers des séances de TP sous R, nous verrons comment utiliser ces méthodes en pratique.

Plan

1 Introduction.

Différence entre estimation (statistique) et prédiction (ML); définition des fonctions de perte, risque, risque empirique.

2 Algorithmes de classification.

Méthodes provenant de la statistique, discrimination linéaire. Méthode des plus proches voisins et autres méthodes universellement consistantes. Arbres de décisions et Random forests.

3 Algorithmes de régression.

Méthode des moindres carrés. Méthodes par pénalisation: estimateur RIDGE, estimateur LASSO et Elastic Net.

4 Sélection d'estimateurs.

Méthodes de minimisation du risque empirique. Données d'apprentissage et de test. Validation croisée.

Références

- Devroye, Györfi, Lugosi - A Probabilistic Theory of Pattern Recognition - (1996) Springer-Verlag
- Hastie, Tibshirani, Friedman - The Elements of Statistical Learning - (2008) Springer Series in Statistics