Séance 1: Introduction à l'Apprentissage Statistique

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Première partie I

Introduction à l'Apprentissage Statistique

Objectifs

- Phénomène physique, biologique, financier, etc trop complexe pour être décrit de manière déterministe $\Longrightarrow$ Utilisation de techniques statistiques d'apprentissage.
- Exemple : Reconnaissance de la parôles, d'images, prédiction de données climatiques, du comportement d'un client, etc.
- Techniques statistiques basées sur des modèles faisant intervenir :
  - des variables explicatives
  - des variables à expliquer
  - une composante de bruit statistique
- But du statisticien : estimer au mieux des paramètres du modèle pour obtenir la meilleure fiabilité de prédiction
- Mots clefs : Machine Learning, Reconnaissance de formes, Intelligence Artificielle

Problématique

- Apprentissage Supervisé :
  - Variable Y à expliquer, décrite par n individus dont on connaît p variables explicatives synthétisées dans X.
  - Ensemble d'apprentissage $\mathcal{D}_{Train} = \{(X_1, Y_1), \dots (X_n, Y_n)\}$
  - Connaissant $\mathcal{D}_{Train}$, on cherche ϕ fonction des p prédicteurs telle que la variable Y s'explique au mieux en fonction des p prédicteurs :
$$Y = \phi(X) + \epsilon$$
  - L'apprentissage est SUPERVISE puisque conditionné par la donnée d'étiquettes (labels, valeurs, etc.) pour chacun des n individus : les Y_i.
  - Exemples classiques : Modèle de régression simple, multiple, arbre binaire de classifications, Réseaux de Neurones, Support Vector Machine, k-N.N., etc.
- Apprentissage Non-Supervisé :
  - Pas de variable Y à expliquer, mais toujours n individus décrits par p variables chacun
  - Objectif : recherche d'une taxinomie (caractéristiques communes) des observations (Exemple : Clustering, CAH, k-means)

Taille des données - Choix de la méthode

- n petit ou modèle statistique des échantillons connu : utilisation des techniques classiques comme modèle linéaire généralisé, vraisemblance, etc. (optimale). Mais quel est le problème dans ce cas ?
- n grand, ou relations non linéaires entre variables : autres méthodes... Quels sont les problèmes et avantages de travailler dans ce cadre là ?
- Exemple : $Y = \phi(X^1, \ldots, X^p) + \epsilon$
  - si ϕ linéaire et p petit : classique
  - si ϕ non linéaire : on ne peut modéliser Y comme c.l. des X^j
  - interaction entre variables $\Longrightarrow$ augmente dimension du modèle
  - Problème de calculabilité informatique pour grandes dimensions
- Choix de méthode :
  - Il existe rarement une "meilleure méthode" !
  - Pertinence et comparaisons des méthodes à étudier au cas par cas.
  - Calcul de l'erreur d'une méthode statistique : ce n'est pas si simple...

Dimension du modèle - Équilibre Biais/Variance

- Importance capitale : construire un modèle parcimonieux. Par exemple :
  - Nombre de variables explicatives réduit
  - Nombre de feuilles dans un arbre de classification
  - Nombre de couches cachées dans un réseau de Neurones
- Modèle complexe $\Longrightarrow$ bon ajustement aux données (Ajouter des variables en régression réduit le biais). Erreur d'ajustement faible.
- Modèle réduit $\Longrightarrow$ Faible variance du modèle.
- MAIS : Ajouter des variables augmente la variance
- MAIS : Modèle réduit : mauvaise qualité d'ajustement
- Objectif : optimiser un dosage entre biais et variance en contrôlant l'ajustement aux données, et la complexité du modèle.

Espérance conditionnelle et Compromis Biais/Variance

- On minimise en f, à la vue des observations x, ce qui justifie :
- On recherche f, fonction de x qui minimise : $\mathbb{E}\left[(y-f(x))^2 |x\right]$
- Proposition : Parmi toutes les fonctions de x, la solution qui minimise (1) est donnée par $f(x) = \mathbb{E}[y|x]$
- Proposition : Connaissant $\mathcal{D} = \{(X_1, Y_1), \dots (X_n, Y_n)\}$, on a le compromis biais/variance :

$$\mathbb{E}\left[(y-f(x,\mathcal{D})^2|\mathcal{D}\right] = \mathbb{E}\left[(y-\mathbb{E}[y|x])^2\right] + \underbrace{\mathbb{E}_{\mathcal{D}}\left[(f(x,\mathcal{D})-\mathbb{E}[y|x])^2\right]}_{\text{Biais}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[(f(x,\mathcal{D})-\mathbb{E}_{\mathcal{D}}[f(x,\mathcal{D})])^2\right]}_{\text{Variance}}$$

- Problème : Loi de y en général inconnue, $\mathbb{E}[y|x]$ souvent incalculable. Approche "optimale" pas toujours réalisable.