Datamining 5: les grande dimension Introduction

M2 STD

November 26, 2014

Le probleme

Supposons qu'on ait des variables X de grande dimension (grande dimension peut signifier qu'il y a beaucoup de variables ou qu'il y a plus de variables que d'individus) On se pose toujours les 3 memes questions:

1. Comment représenter X (projection) ?
2. Comment "prédire" une variable Y quantititative (regression) ?
3. Comment "prédire" une variable Y qualitative (classification) ?

Quelques cas concrets de grande dimension

- Les données fonctionelles (cas extreme de la dimension infinie)
- **Les "grandes bases" (le bigdata)**

Les méthodes “classiques” face a de tels cas

On va voir les problèmes inérhents a l'ACP, l'AFD et la regression linéaire dans ces cas

Supposons qu'on ait n individus et $p \gg n$ variables. Alors on va diagonaliser $X'X$ qui est de dimension n, bref on est intrinsèquement limités par le nombre d'observation qu'on fait. Le seul moyen de "s'en sortir" est de considérer qu'en réalité les données "reposent" dans un espace de dimension $k \ll n \ll p$. Sous cette hypothèse (uniquement) l'ACP donnera des résultats "corrects". et les méthodes vues au chapitre 1 fonctionneront bien sous condition d'avoir réalisé au préliminaire une ACP

La regression

Rappelons que pour effectuer une regression il faut inverser $X'X$. Ici $X'X$ n'est pas inversible (de dimension p et de rang $n < p$). On ne peut donc, même pas effectuer une regression linéaire...

1. $n < p$ mais $k < p$ regression PLS
2. $n < p$ et $k < p$ mais hypothèse supplémentaire **la sparsité** Y ne dépend que d'un petit nombre r de variables avec $r \ll n \ll p$

L'analyse discriminante

Meme probleme que la regression