Introduction

— Slides : link

1 Qu'est-ce que l'analyse de données ?

L'analyse de données est un ensemble de méthodes permettant de retirer de l'information d'un jeu de données. On parle aussi d'apprentissage statistique (*statistical learning*). L'idée est d'utiliser des modèles statistiques pour comprendre comment les données sont structurées et comment elles intéragissent l'une avec l'autre.

> Exemple
>
> Imaginons que vous êtes employé par l'Organisation des Nations Unies (ONU). Votre mission est d'analyser l'espérance de vie à travers le monde. Pour celà, vous disposez d'une mesure de l'espérance de vie dans chaque pays membre de l'ONU, bien sûr, mais aussi le PIB par habitant, les montants des dépenses liés à la santé, le taux de fertilité, le taux d'urbanisation, le niveau d'éducation du pays, etc. Le but de l'analyse de données est de trouver des liens entre ses différentes variables et la variable d'intérêt, l'espérance de vie, de visualiser ces données, et éventuellement de prédire l'espérance de vie à partir des autres variables.

2 Objectifs du cours

Dans ce cours, on cherche à introduire des méthodes qui permettent une étude d'un jeu de données de "haute dimension" (dans le sens où l'on ne peut pas faire un simple graphique de l'ensemble des variables pour chaque observation) sans avoir recours à un modèle probabiliste. Les différentes techniques que l'on va voir peuvent servir à :

— visualiser les données ;
— réduire la dimension des données ;
— identifier certains liens entre les variables ;

— diviser le jeu de données en groupes/classes.

Ce cours n'a pas vocation à être exhaustif, dans le sens de présenter toutes les méthodes possibles. Ce cours n'a pas non plus vocation à être à l'état de l'art, dans le sens où on ne s'intéressera pas aux derniers développements en apprentissage machine. Ce cours n'est pas non plus un cours de programmation.