# Apprentissage en grande dimension (Partie #1)

Thomas Verdebout

Université de Lille

# Contenu du cours

1. Introduction.
2. Apprentissage supervisé: régression.
3. Apprentissage supervisé: classification.
4. Apprentissage non-supervisé

# 1: Introduction.

Ce cours utilise des livres de référence:

- Theory of Multivariate Statistics, de Bilodeau et Brenner
- An introduction to Statistical learning with Applications in R, de James, Witten, Hastie et Tibshirani
- Statistics for high-dimensional data, Bulhmann et van de Geer

Bien que le terme d'apprentissage statistique soit récent, bon nombre des concepts qui sous-tendent ce domaine ont été introduits il y a relativement longtemps.

La méthode d'apprentissage la plus classique fondée sur la méthode des moindres carrés date en effet de travaux de Legendre et Gauss.

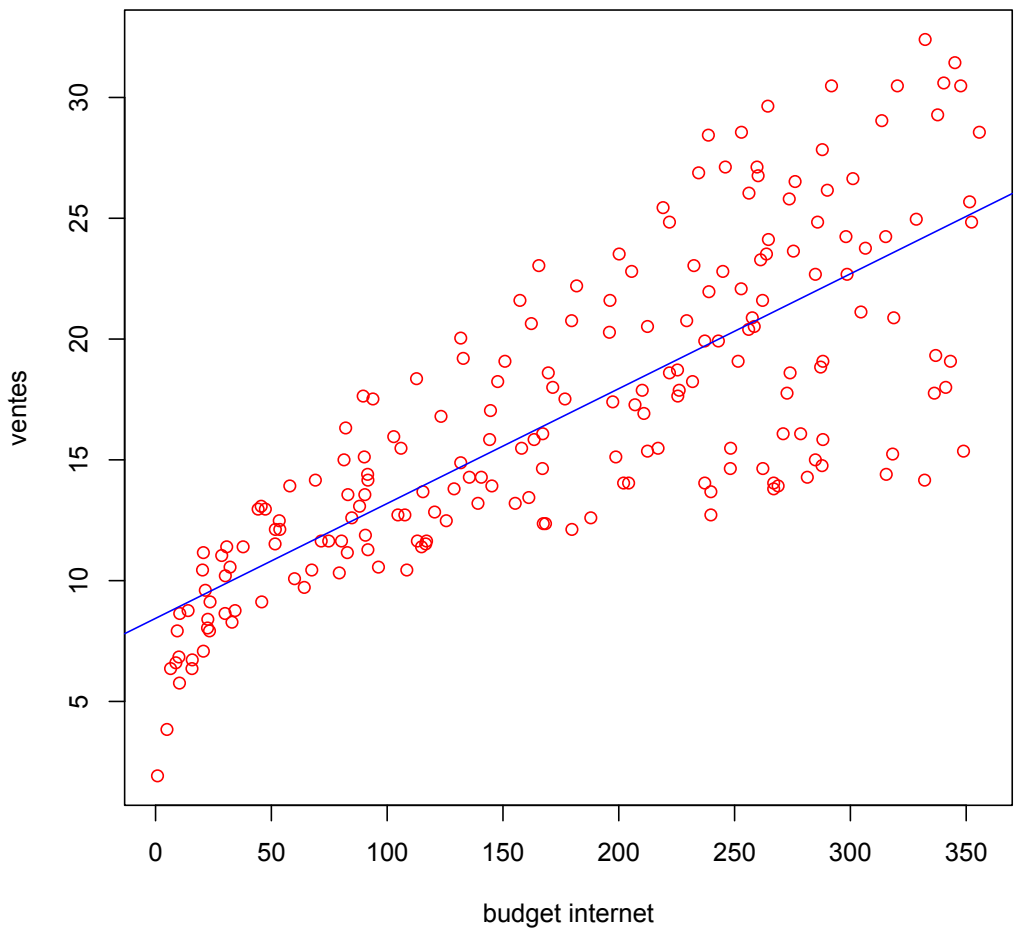
Fisher a proposé l'analyse discriminante linéaire en 1936. Un peu plus tard fut développée la régression logistique. Au début des années 1970, on a commencé à parler de modèles linéaires généralisés pour désigner une classe entière de méthodes d'apprentissage statistique qui incluent à la fois la régression linéaire et la régression logistique comme cas particuliers.

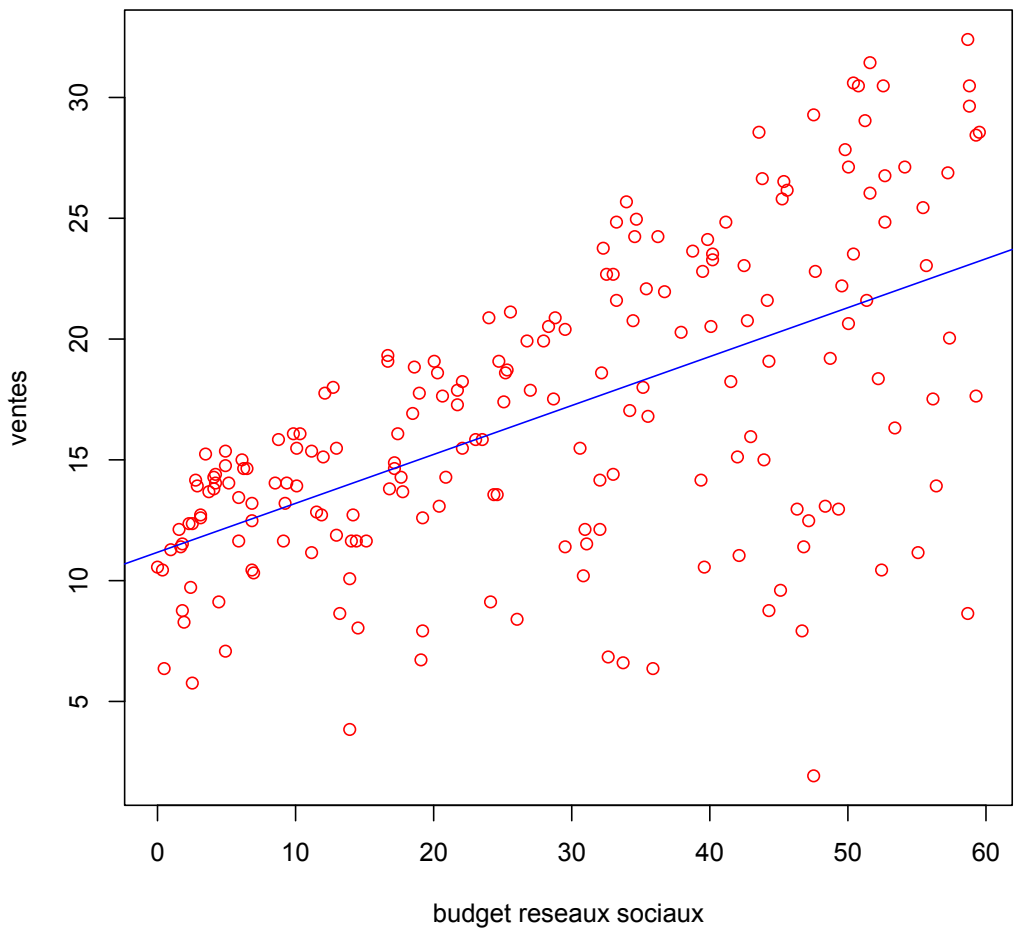
On peut également citer Breiman, Friedman, Olshen et Stone qui au milieu des années 1980 ont introduit les *arbres* de classification et de régression. Hastie et Tibshirani ont inventé le terme de modèles additifs généralisés (GAM) en 1986 pour une classe d'extensions non linéaires des modèles linéaires généralisés.

L'apprentissage statistique est aujourd'hui un nouveau domaine de la statistique, axé sur la modélisation et la prédiction *supervisées et non supervisées*.

Supposons que nous soyons des Statisticiens engagés pour fournir des conseils sur la manière d'améliorer les ventes d'un produit.

Nous avons des données qui concernent la vente du produit dans 200 marchés différents ainsi que les budgets publicitaires pour le produit sur chacun de ces marchés pour trois médias différents : internet, réseaux sociaux et les journaux.


ventes
5
10
15
20
25
30
0
50
100
150
200
250
300
350
budget internet


ventes
5
10
15
20
25
30
0
10
20
30
40
50
60
budget reseaux sociaux

Il n’est clairement pas possible ici d’augmenter directement les ventes du produit (nous n’avons pas le contôle sur ceci).

En revanche, nous pouvons contrôler les dépenses publicitaires dans chacun des trois médias. Par conséquent, si nous déterminons qu’il existe une association entre la publicité et les ventes, nous pouvons alors ajuster les budgets publicitaires, augmentant ainsi indirectement les ventes.

En d’autres termes, notre objectif est de développer un modèle qui puisse être utilisé pour prédire les ventes sur la base des budgets des trois médias.

Dans le probème qui nous intéresse ici, les budgets publicitaires sont des *variables d'entrée* (*input variables*) tandis que les ventes sont une *variable de sortie* (*output variable*).

Les variables d'entrée sont généralement notées avec le symbole $X$. Nous avons ici trois variables d'entrées: $X_1$ le budget internet, $X_2$ le budget réseaux sociaux et $X_3$ le budget journaux. Ces trois variables d'entrée peuvent nous permettre de prédire la variables de sorties que sont les ventes notée $Y$.

Dans la suite, les termes *prédicteurs*, *covariables* seront aussi utilisés pour désigner les variables d'entrée tandis que le terme *réponse* sera aussi utilisé pour désigner la variable de sortie.

Plus généralement, supposons que nous observons $p$ prédicteurs $\mathbf{X} = (X_1, \ldots, X_p)$ et une réponse $Y$. Supposons qu'il existe un lien entre la réponse et les prédicteurs qui peut s'écrire de la forme

$$Y = f(\mathbf{X}) + \epsilon, \tag{1}$$

où $f$ est ici une fonction inconnue de $X_1, \ldots, X_p$ et où $\epsilon$ est un *terme d'erreur* aléatoire indépendant de $\mathbf{X}$ et d'espérance nulle.

Comme indiquer ci-dessus, la fonction $f$ qui relie la variable d'entrée à la variable de sortie est en général inconnue: il faut l'estimer!

## Pourquoi estimer $f$?

- Dans beaucoup de situations, des covariables $\mathbf{X}$ sont disponibles, mais la réponse $Y$ ne peut pas être facilement obtenue. Dans ce cas, puisque le terme d'erreur est d'espérance nulle, une prédiction raisonnable de $Y$ est

$$\hat{Y} = \hat{f}(\mathbf{X}),$$

où $\hat{f}$ est notre estimateur de $f$. Lorsque c'est un prédiction qui nous intéresse, la forme ou la *qualité* de $\hat{f}$ nous importe peu. Ce qui est important, c'est la qualité de la prédiction $\hat{Y}$.

La qualité de la prédiction $\hat{Y}$ de $Y$ dépend clairement de deux quantités. Supposons que nous mesurions la qualité de notre prédiction avec un *risque* (de type $L_2$) de la forme $\mathrm{E}[(\hat{Y} - Y)^2]$. Supposons un instant que $\hat{f}$ et $\mathbf{X}$ sont fixés (ne sont pas aléatoires). Nous avons donc dans le modèle ci-dessus que

$$\begin{aligned} \mathrm{E}[(\hat{Y} - Y)^2] &= \mathrm{E}[(\hat{f}(\mathbf{X}) - f(\mathbf{X}) - \epsilon)^2] \\ &= (\hat{f}(\mathbf{X}) - f(\mathbf{X}))^2 + \mathrm{Var}(\epsilon) \end{aligned}$$

Le risque provient donc de deux facteurs: l'estimation de $f$ et la variance de l'erreur $\epsilon$. On peut donc clairement améliorer la prédiction en obtenant des estimateurs de $f$ performants.

## Pourquoi estimer $f$?

- Pour faire de l'inférence. Nous sommes souvent intéressés par comprendre la nature du lien entre $Y$ et $\mathbf{X}$. Dans cette situation, nous allons aussi estimer f, mais le but n'est de faire des prédictions sur $Y$ . Nous voulons ici comprendre la relation entre $\mathbf{X}$ et $Y$, ou plus précisément, comprendre comment $Y$ évolue en fonction de $X_1, \ldots, X_p$. En particulier, nous pourrions avoir envie de répondre aux questions suivantes :
  1. Quels prédicteurs jouent un rôle sur la réponse ?
  2. Quelle est la relation entre la réponse et chaque prédicteur ?
  3. La relation entre Y et chaque prédicteur peut-elle être résumée de manière adéquate à l'aide d'une équation linéaire, ou la relation est-elle plus compliquée ?

Pour revenir à notre exemple:

1. Quels sont les médias qui contribuent aux ventes ?
2. Quels sont les médias qui génèrent la plus forte augmentation des ventes ?
3. Dans quelle mesure l'augmentation des ventes est-elle associée à une augmentation donnée de la publicité via internet ?

Dans les deux applications liées à l'estimation de $f$ (la prédiction et l'inférence), la qualité de l'estimation de $f$ joue un rôle très important. Pour estimer $f$, nous supposerons toujours que nous avons observé un ensemble de $n$ points $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ de données différents. Ces observations sont appelées données d'apprentissage car nous les utiliserons pour *entraîner*, ou *apprendre*, à notre méthode comment estimer $f$.

On peut distinguer essentiellement deux types de méthodes d'apprentissage: les méthodes paramétriques et les méthodes non-paramétriques.

# Les méthodes paramétriques

Les méthodes paramétriques fonctionnent en deux étapes:

1. Faire une hypothèse sur la forme de $f$ de sorte à n'avoir qu'à estimer un paramètre fini-dimensionnel. Par exemple, on suppose que $f$ est une fonction linéaire des covariables $X_1, \ldots, X_p$:

   $$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

   Dans ce modèle linéaire, on a beaucoup simplifier l'estimation de $f$. On est passé de l'estimation d'un paramètre infini-dimensionnel $f$ à l'estimation des $p + 1$ paramètres $\beta_0, \beta_1, \ldots, \beta_p$.
2. Après avoir sélectionné un modèle, nous avons besoin d'une procédure qui utilise les données pour *ajuster* ou *entraîner* le modèle. Dans le cas d'un modèle linéaire, la méthode la plus classique est la méthode des moindres carrés.

L'avantage des méthodes paramétriques est donc qu'elle simplifie grandement l'estimation du modèle ou l'entrainement.
L'inconvénient potentiel d'une approche paramétrique est que le modèle que nous choisissons ne correspondra généralement pas à la vraie forme inconnue de $f$. Si le modèle choisi est trop éloigné de $f$, l'estimation sera mauvaise. Nous pouvons essayer de résoudre ce problème en choisissant des modèles flexibles qui peuvent s'adapter à de nombreuses formes fonctionnelles différentes pour $f$. Mais en général, l'adaptation d'un modèle plus flexible nécessite l'estimation d'un plus grand nombre de paramètres. Ces modèles plus complexes peuvent conduire à un phénomène connu sous le nom d'ajustement excessif des données (*overfitting*).

## Les méthodes non-paramétriques

Les méthodes non paramétriques ne font pas d'hypothèses explicites sur la forme fonctionnelle de f . Au lieu de cela, elles cherchent à obtenir une estimation de *f* qui se rapproche "le plus possible des points de données". Ces approches peuvent présenter un avantage majeur par rapport aux approches paramétriques : en évitant l'hypothèse d'une forme fonctionnelle particulière pour *f*, elles ont la possibilité d'ajuster avec précision un plus grand nombre de formes possibles pour *f*. Toute approche paramétrique comporte la possibilité que la forme fonctionnelle utilisée pour estimer *f* soit très différente de la vraie *f*, auquel cas le modèle résultant ne s'ajustera pas bien aux données. En revanche, les approches non-paramétriques évitent complètement ce danger, puisqu'aucune hypothèse sur la forme de *f* n'est formulée.

Cependant les approches non paramétriques souffrent d'un inconvénient majeur : elles ne réduisent pas le problème de l'estimation de $f$ à un petit nombre de paramètres, un très grand nombre d'observations (bien plus que ce qui est généralement nécessaire pour une approche paramétrique) est donc généralement nécessaire pour obtenir une estimation précise de $f$.

On peut raisonnablement se poser la question: pourquoi choisiri une méthode plus restrictive plutôt qu'une approche plus souple ? Il y a plusieurs raisons pour lesquelles nous pourrions préférer un modèle plus restrictif. Par exemple, si nous sommes principalement intéressés par l'inférence, les modèles restrictifs sont beaucoup plus faciles à interpréter; l'interprétation d'un modèle linéaire est très simple. Certaines méthodes d'estimation de $f$ sont tellement compliquées qu'elles ne permettent pas de comprendre comment les covariables agissent sur la prédiction de $Y$.

## Apprentissage supervisé ou non-supervisé

La plupart des problèmes d'apprentissage statistique tombent dans l'une des deux catégories suivantes : apprentissage supervisé ou non supervisé. Jusqu'ici nous avons décrit des problèmes du domaine de l'apprentissage supervisé. Pour chaque observation du prédicteur, il existe une mesure de réponse associée. Nous souhaitons ajuster un modèle qui relie la réponse aux prédicteurs, dans le but de prédire avec précision la réponse pour les observations futures (prédiction) ou de mieux comprendre la relation entre la réponse et les prédicteurs (inférence).

En revanche, l'apprentissage non supervisé décrit une situation dans laquelle chaque observation est un vecteur de mesures mais aucune réponse n'y est associée. On parle d'apprentissage non supervisé parce que nous n'avons pas de variable de réponse qui puisse superviser notre analyse. Dans ces situations, l'objectif est souvent comprendre les relations entre les variables observées ou de former des groupes (clusters) d'observations.