Les principes de l'apprentissage

August 21, 2019

Mots clefs. Après avoir lu le cours, ces mots doivent vous être familiés.

- Variables d'entrées=variables explicatives=descripteur=X, variable de sortie=à expliquer=Y.
- Fonction objectif=$f^?$, estimateur=$\widehat{f}$.
- Variable quantitative / qualitative, classification / regression
- Echantillon=Observations, Train / Test / Validation.
- Loss (=Perte=Coût). Distance.
- Modèle paramétrique / Non paramétrique.
- Prévoir / interpréter

1 variable aléatoire et réalisation

- Les variables aléatoires sont notées avec des majuscules ex : X représente une donnée générale, pas encore observée. On ne connait pas sa valeur, mais on peut imaginer sa loi (= la proba qu'elle fasse ceci ou cela).
- Les constantes sont notées avec des minuscules ex : x représente une donnée numérique, la réalisation d'un X.
- On peut par exemple écrire $\mathbf{P}[X = x]$.

Il est parfois difficile de savoir quand il faut mettre une majuscule ou pas. Voici un exemple qui vous éclairera peut-être.

Exemple : Considérons un quartier dans lequel toutes les maisons sont identiques. Notons $X_1, X_2, X_3...$ le prix des différentes maisons. Ces prix sont aléatoires car ils dépendent des négociations futures entre acheteurs et vendeurs. On suppose de plus que la vente d'une maison n'influe pas sur la vente d'une autre. Ainsi $X_1, X_2, ...$ sont des "copies indépendantes" d'une même variable aléatoire X (qui représente le prix "générique" d'une maison). Supposons maintenant que les 10 premières maisons se sont vendues. Puisque la vente a eu lieu, ces prix ne sont plus aléatoires, on les notes avec des petites lettres $x_1, x_2, ..., x_{10}$. L'espérance du prix générique X peut-être estimée par la moyenne $\frac{1}{10}(x_1 + ... + x_{10})$.

Le héros de ce cours sera un couple aléatoire (X, Y), avec le plus souvent $X \in \mathbb{R}^p$ et $Y \in \mathbb{R}$. Nous disposerons de copies indépendantes (X_i, Y_i), que

nous appellerons échantillon. Et de temps en temps, nous considérerons des réalisations (x_i, y_i) de (X_i, Y_i) qui nous permettrons d'estimer des espérances : $\mathbf{E}[\varphi(X,Y)] \simeq \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i, y_i)$.

2 Les principes de l'apprentissage

2.1 Construire un estimateur $\widehat{f}$ d'une fonction inconnue $f^?$.

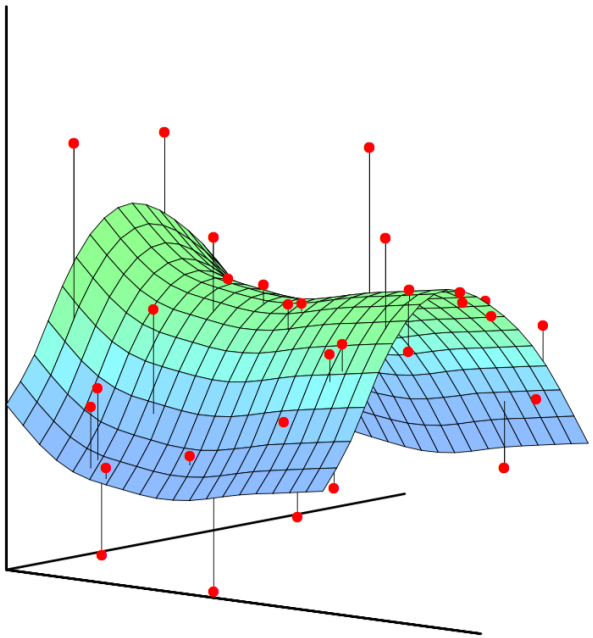
On considère des variables explicatives (=variables d'entrée=input) $X = (X^1, ..., X^p) \in \mathbb{R}^p$ (ex : la surface, le quartier, le nombre de pieces d'un appartement).

On considère une variable à expliquer (=variable de sortie=output) Y (ex: le prix d'un appartement).

On suppose l'existence d'une relation $Y = f^?(X) + Bruit$.

On ne connait pas $f^?$ mais l'on dispose d'observations indépendantes X_i et $Y_i = f^?(X_i) + Bruit_i$.

On va construire un estimateur $\widehat{f}$ à partir des observations, en espérant in fine que $\widehat{f}$ et $f^?$ soient proches.



2.2 Deux objectifs

- Prévoir : on a une entrée $x = (x^1, ..., x^p)$ dont on ne connait pas la sortie correspondante. On la prédira avec $\widehat{f}(x)$.
- Interpréter = comprendre l'influence des entrées sur la sortie. Par exemple X^1=quantité de tabac consommé, X^2=quantité de lait consommé, $Y \in \{0; 1\}$= avoir ou ne pas avoir un cancer du poumon. Dans quelle mesure les entrées X^1 et X^2 influencent la sortie Y ?

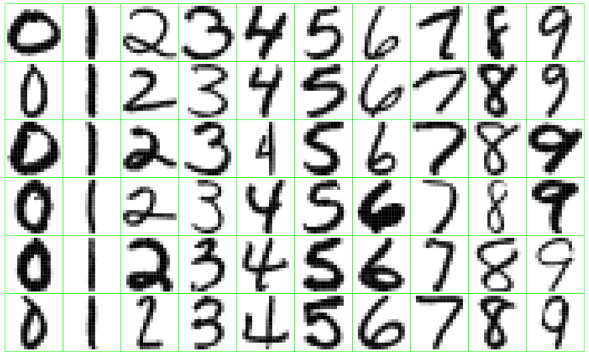
2.3 Types de variables

2 types de variables :

- variable quantitative : age, prix, surface, niveau de gris
- variable qualitative[1] : ex : sexe, région, catégorie d'age.
- la sortie Y est quantitative. On parle de régression
- la sortie Y est qualitative. On parle de classification

Exemple MNIST (Mixed National Institute of Standards and Technology)

- Entrée quantitative : $X \in \mathbb{R}^{28\times 28}$: une image en niveau de gris.
- Sortie qualitative $Y \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$: les 10 chiffres possibles.
- Problème : associé à chaque image le chiffre correspondant. C'est de la regression/classification ?



[1]= categorielle, discrète, factors (anglais)

Exemple : Prix d'une assurance.

- Entrées quantitatives : âge du conducteur, coefficient bonus-malus, prix du véhicule.
- Et aussi des entrées qualitatives : sexe du conducteur, département, marque de véhicule.
- Sorties quantitatives : prix de la prime d'assurance.
- Problème : associé à chaque conducteur le tarif qu'il mérite. C'est de la regression/classification ?

3 Construire et juger $\widehat{f}$.

3.1 Juger $\widehat{f}$.

On sépare nos observations (X_i, Y_i) en deux parties :

- Les données d'entrainement: $Train$, que nous utilisons pour construire $\widehat{f}$.
- Les données de test : $Test$, qui nous servent à juger le $\widehat{f}$ qu'on vient de construire. Ce jugement se fait à l'aide d'une fonction de coût (= de perte=Loss)

$$Loss_{Test} = \sum_{Test} \text{dist}(Y_i, \hat{f}(X_i)) \qquad \text{avec par exemple dist}(a,b) = (a-b)^2$$

Si $Loss$ est petit, on a gagné !

On verra d'autres exemples de $Loss$. En particulier, quand Y est qualitative (ex: un numéro de département), on ne plus utiliser $\text{dist}(a, b) = (a - b)^2$.

3.2 Construire $\widehat{f}$ en minimisant un $Loss_{Train}$

Puisque l'estimateur est jugé avec une fonction coût, pourquoi utiliser cette même fonction pour construire l'estimateur !

On se donne une famille de fonction $\mathcal{F}$ possible et on pose :

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\, Loss_{Train} = \underset{f \in \mathcal{F}}{\text{argmin}} \sum_{Train} \text{dist}(Y_i, f(X_i)) \tag{1}$$

Mais attention, pour définir ce $Loss_{train}$, nous n'avons pas utilisé les même données (X_i, Y_i) que pour $Loss_{test}$. Car il ne faut pas être juge et parti !

On peut aussi vouloir expressément que $\widehat{f}$ soit régulière en mettant une pénalité sur les f irréguliers. Par exemple :

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}} \sum_{Train} \text{dist}(Y_i, f(X_i)) + \lambda Penalisation(f)$$

où la $Penalisation$ associe aux fonctions "compliquée" une valeur importante. Ainsi, quand on minimise le terme ci-dessus, on doit faire un compromis entre "coller aux donnée train" et "garder un modèle simple".

Souvent l'ensemble de fonctions possibles $\mathcal{F}$ est décrit à l'aide de paramètres. On parle alors de modèle paramétrique. Par exemple

$$\mathcal{F} = \{f(x) = \theta_0 + \theta_1 x : \theta \in \Theta = \mathbb{R}^2\}$$

Lorsque l'on a paramétré $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, la formule (1) se décompose ainsi

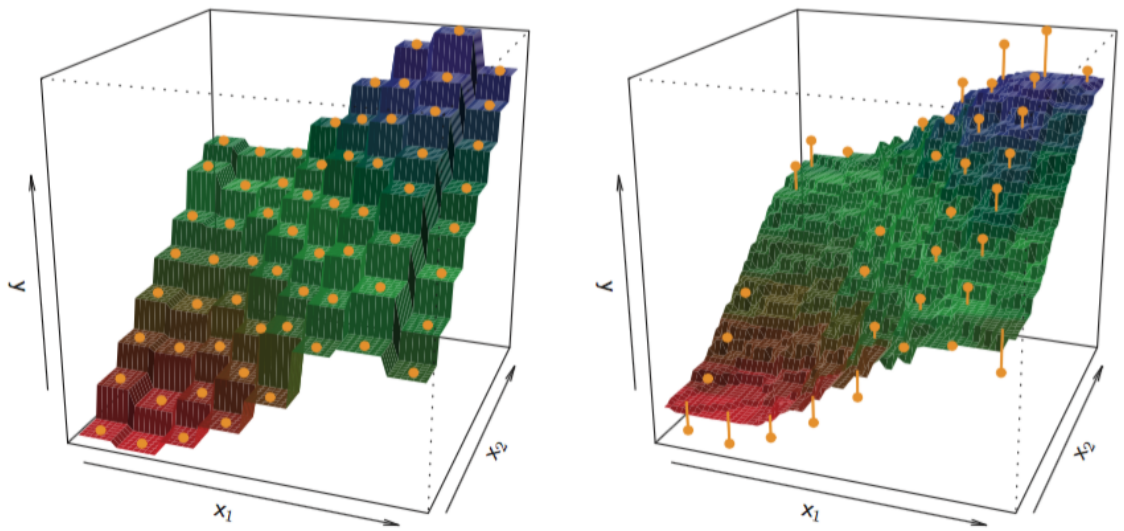
$$\hat{f} = f_{\widehat{\theta}} \qquad \text{avec } \widehat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_{Train} \operatorname{dist}(Y_i, f_\theta(X_i)) \tag{2}$$

3.3 Construire $\widehat{f}$ sans minimiser de $Loss_{Train}$

La technique des k plus proches voisins :

$$\widehat{f}(x) = \frac{1}{k} \sum_{i:X_i\in V_k(x)} Y_i$$

où V_k est le voisinage de x constitué des k plus proches X_i dans $Train$. C'est une technique non-paramètrique (on n'a même pas décrit l'ensemble des fonctions possibles $\mathcal{F}$).



La technique des noyaux. On se donne un noyau $N(x, y)$, par exemple $N(x, y) = e^{-\frac{1}{2}(\frac{x-y}{\sigma})^2}$, puis

$$\widehat{f}(x) = \frac{\sum_{Train} N(x, X_i) Y_i}{\sum_{Train} N(x, X_i)}$$

On reconnait là les techniques d'interpolation, et de lissage.

3.4 Qu'est-ce que la modélisation

Modéliser signifie faire un pari, (= faire des hypothèses) sur $f^?$, sur le bruit, ou plus généralement sur la loi jointe de (X, Y). Ce pari peut-être fondé sur notre connaissance du problème (avis d'expert), ou bien sur l'observation des

données (statistiques descriptives). Cette modélisation conduit naturellement à choisir une technique pour construire $\widehat{f}$.

Exemple : Il est naturel de supposer une relation affine entre le prix d'un appartement Y et sa surface X. Donc $Y = w_0^? + w_1^? X + Bruit$. Ce qui nous conduit à l'estimateur suivant :

$$\widehat{f}(x) = \widehat{w}_0 + \widehat{w}_1 x \qquad \text{avec } (\widehat{w}_0, \widehat{w}_1) = \underset{(w_0,w_1)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{Train} (w_0 + w_1 X_i - Y_i)^2$$

La modélisation peut aller plus loin : on peut parier que $Bruit$ est une v.a. Gaussienne centrée de variance σ^2. Cette hypothèse supplémentaire permet de calculer des intervalles de confiance, de faire des tests etc.

Ainsi, la regression linéaire (qu'on verra en détail plus loin) revient à parier que les données sont disposée selon une droite, un plan, un hyperplan...

Par contre, quand on utilise la technique des plus proches voisins, avec $k = 1$ petit,

$$\widehat{f}(x) = Y_i \text{ avec } X_i \text{ le plus proche de } x$$

on parie plutôt que la fonction $f^?$ est localement constante. Ainsi $f^?(x)$ est proches des $f^?(x_i)$ pour x_i voisins de x.

3.5 Sélection de modèle

Une technique plus avancée consiste à choisir plusieurs modèles, ce qui conduit à plusieurs estimateurs $(\widehat{f}_1, \widehat{f}_2, ...)$, puis il faut choisir le meilleur. Pour faire

marcher cette technique, nous devons séparer nos observations (X_i, Y_i) en trois parties: $Train$, $Validation$, $Test$.

- $Train$ sert à construire les estimateurs $(\widehat{f}_1, \widehat{f}_2, ...)$. Par exemple on se donne plusieurs ensembles de fonctions $\mathcal{F}_1, \mathcal{F}_2, ...$ et plusieurs constantes $\lambda_1, \lambda_2, ...$

$$\widehat{f}_k = \underset{f \in \mathcal{F}_k}{\operatorname{argmin}} \sum_{Train} \operatorname{dist}(f(X_i), Y_i) + \lambda_k \, penalisation(f)$$

- $Validation$ sert à sélectionner le bon estimateur $\widehat{f}_k$.

$$\widehat{f} = \widehat{f}_{\widehat{k}} \qquad \text{avec } \widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{Validation} \operatorname{dist}(\widehat{f}_k(X_i), Y_i)$$

- $Test$ sert à tester si notre modèle final est bon à l'aide du critère :

$$\sum_{Test} \operatorname{dist}(\widehat{f}(X_i), Y_i)$$