Fondements statistiques de l'apprentissage automatique

Laurent Risser
CNRS - Institut de Mathématiques de Toulouse (IMT UMR5219)
3IA Artificial and Natural Intelligence Toulouse Institute (ANITI)

ISAE-SUPAERO - 2021/22

Table des matières

Chapitre 1

Introduction

1.1 Modèle linéaire en Sciences de la Décision ?

Motivation

Afin d'étudier quantitativement un phénomène à $p \geq 1$ variables d'entrée $X = \left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right)$ et une variable de sortie Y, il est bien pratique de construire un modèle g qui explique par une relation mathématique les valeurs observées de Y en fonction des variables d'entrée :

$$Y = g\left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right) .$$

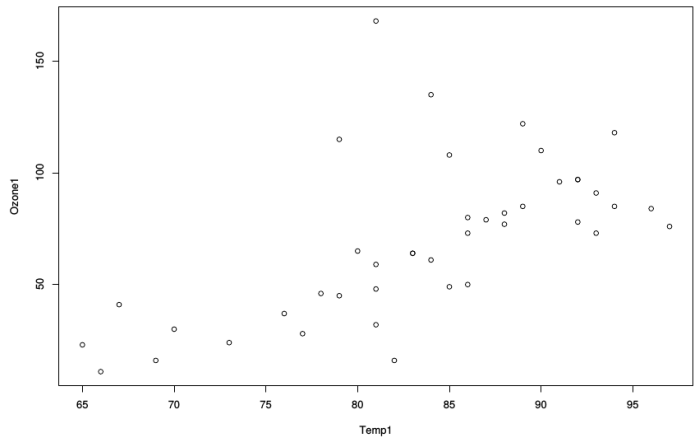
Ce modèle essaie de refléter le plus fidèlement possible la réalité à partir de n observations du phénonène. Il permet ainsi de mieux comprendre le phénomène étudié, mais potentiellement aussi de prédire les sorties inconnues Y en lien avec de nouvelles données d'entrée X.

On distingue deux types de modèles :

1. *Modèles déterministes* : C'est une équation ou un ensemble d'équations qui émanent souvent de lois physiques, chimiques, économiques, ..., et représentent le comportement attendu du phénomène.
2. *Modèles statistiques* : Souvent, il est difficile de développer un modèle théorique car le phénomène étudié est trop complexe. On a alors recours à un modèle statistique basé non pas sur une théorie, mais sur des données observées.

Exemple

On étudie la pollution de l'air à New-York. On a mesuré pendant 111 jours la concentration en ozone, noté 0_i (en ppm), et la température de l'air, notée T_i (en degrés Farenheit). Le tableau ci-dessous représente une partie des observations (celles pour lesquelles la vitesse du vent et le rayonnement solaire sont dans une certaine plage).



On constate que la concentration en ozone croit avec la température. La relation est approximativement linéaire dans la zone représentée ici. En considérant les T_i variables (ou observations) d'entrées et les O_i comme variables (ou observations) de sorties, on introduit alors le modèle :

$$O_i = a + bT_i + \varepsilon_i \,, \tag{1.1}$$

pour chaque observation $i = 1, \cdots, n$, où ε_i représente un bruit entre les observations de sorties réelles et celles prédites par le modèles. Ce modèle est appelé modèle de régression linéaire simple et sera étudié dans ce cours.

Questions posées dans ce cours

La résolution et l'étude du problème introduit ci-dessus sont discutés au début de ce cours (Chapitre 2). Beaucoup d'autres questions permettent de bien comprendre les bases de l'apprentissage statistique, qui est une composante importante de l'Intelligence Artificielle :

- Peut-on s'assurer qu'il y a une relation entre les entrées et les sorties ?
- Quel est le niveau d'incertitude sur cette relation ?
- Peut-on détecter des valeurs abérantes ?
- Que faire si la dimension des entrées (p) est plus grande que le nombre d'observations (n) ?
- Que faire si le niveau de bruit n'est pas le même pour differents groupes de variables ou si différents groupes de variables ont un *bruit* de moyenne non nulle.
- ...

Ces questions seront abordées dand le cadre de ce cour.

1.2 Rappels en Probabilités/Statistique

1.2.1 Notions de variable aléatoire et de densité de probabilité

Variable aléatoire Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre

de ce cours ses résultats possibles seront toujours dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On distinguera en particulier le *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple $X \in \{0, 1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face.

Loi de probabilité La *loi de probabilité* d'une v.a. décrit la probabilité d'obtenir les différents résultats de cette variable.

Loi de probabilité discrète Par exemple si l'on joue à pile ou face avec une pièce parfaitement équilibrée, on a $\mathbb{P}(X = 0) = 1 - p = 0.5$ et $\mathbb{P}(X = 1) = p = 0.5$. On remarquera que la somme des probabilités de tous les résultats possibles dans le cas discret est toujours 1.

Loi de probabilité continue Dans le cas continu, écrire $\mathbb{P}(X = x)$ n'a aucun sens puisque la probabilité d'une valeur exacte est infinitésimale. On pourra par contre utiliser la *fonction de répartition* $F_X(x) = \mathbb{P}(X \leq x)$ pour représenter comment se répartissent les probabilités des différents résultats de X. Il sera alors possible de quantifier les chances que X soit sur une certaine gamme de valeurs $\mathbb{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. Naturellement, on aura toujours $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$. De manière purement équivalente à la fonction de répartition $p_X(x)$, la *densité de probabilité* pourra de même représenter la loi de probabilité d'une v.a. X suivant :

$$p_X(x) = \frac{\partial F_X}{\partial x}(x)$$

En utilisant les densités de probabilités, les chances que X tombe sur une gamme de valeurs $[x_1, x_2]$ sera alors

$$\mathbb{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p_X(x)dx\,.$$

.

1.2.2 Théorème central limite

Afin de montrer l'importance de la loi Normale en probabilités/statistique, ainsi que de manipuler les concepts énoncés ci-dessus, il est intéressant de présenter maintenant le Théorème Central Limite (TCL).

Supposons que n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes mais suivant une même loi de probabilité soient tirés. L'espérance (ou moyenne) m et l'écart type s de leur loi est connue. Le nombre d'observations n est aussi supposé grand (typiquement $n > 30$). Alors, la somme des X_i peut être approchée par une loi normal de moyenne nm et d'écart type $s\sqrt{n}$, *i.e.* :

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(nm, s^2 n)\,,$$

où la *densité de probabilité* de la loi normale $\mathcal{N}(\mu, \sigma^2)$ est (voir aussi appendice A) :

$$f_{\theta=\{\mu,\sigma\}}(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

On peut de même montrer que la loi de $\sum_{i=1}^{n} X_i$ tend de même vers $\mathcal{N}(nm, s^2 n)$ lorsque n tend vers l'infini. Nous ne le montrerons pas ici, mais il est aisé de trouver la preuve de ce théorème.

Afin de nous familiariser avec les notions énoncées ci-dessus, nous proposons de vérifier empiriquement le TCL dans le cas d'une pièce tirée à pile ou face. Le protocole expérimental sera le suivant :

- Chaque étudiant de la classe tire $n = 10$ fois une pièce à pile ou face avec et compte le nombre de fois que la pièce est tombée sur pile. Pile correspond alors à $X_i = 1$ et face à $X_i = 0$.
- On suppose que $\mathbb{P}(X = 1) = 0.5$ et $\mathbb{P}(X = 0) = 0.5$, ce qui est sans doute très proche de la réalité. Ainsi l'espérance (moyenne) de X est $m = 0.5$ et son écart type est $s = 0.5$.
- On va dessiner un graphique dans lequel l'abscisse représente le nombre de 'piles' potentiellement obtenus par un étudiant (entre 0 et 10) et l'ordonnée représente le nombre d'étudiant qui ont obtenus ce nombre de 'piles' divisé par le nombre total d'étudiants.
- On constatera que cette courbe approche la densité de la loi normale de moyenne $10m$ et d'écart type $s\sqrt{10}$ (voir appendice A).

Au delà de la connaissance du TCL lui même et de l'illustration des notions de la section 1.2.1, cet exemple nous amène un enseignement qui est (à mes yeux) l'essence de la modélisation statistique. En assemblant plusieurs variables aléatoires, nous avons créé un modèle aléatoire dont on peut étudier les propriétés statistiques telles que la moyenne mais aussi d'une certaine manière la précision/étendue/sensibilité. Ce type de modélisation se distingue alors de la modélisation déterministe qui ne s'intéresse qu'à l'équivalent de la moyenne ici.

1.2.3 Estimation empirique des paramètres d'un modèle

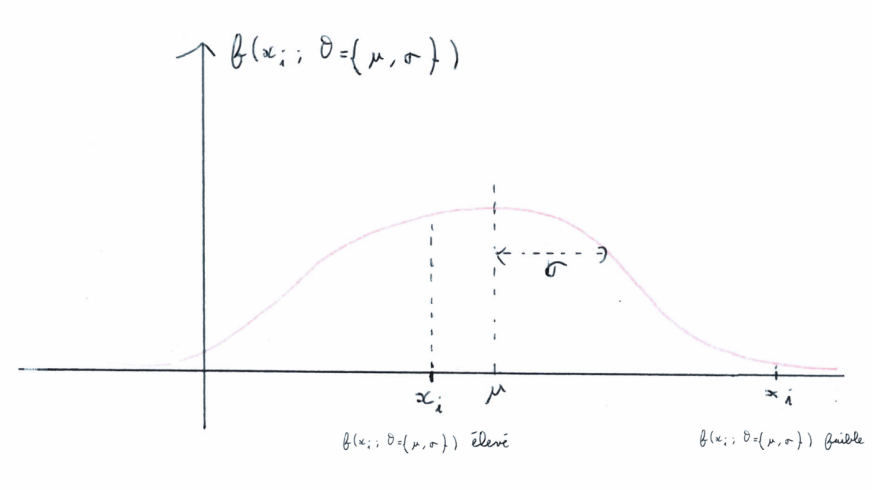
Un des composantes importantes de ce cour est de donner des méthodes pour l'estimation des paramètres de lois à partir d'observations, ou plus spécifiquement de paramètres de modèles contenant des variables aléatoires (c'est à dire avec des sources d'aléa). Cette estimation est classiquement effectuée en suivant le principe du maximum de vraisemblance ou plus simplement une estimation au sens des moindres carrés.

Maximum de vraisemblance

On dénote X une variable aléatoire (v.a.) supposée suivre une loi discrète (e.g. Bernoulli) ou continue (e.g. Normale) de paramètres θ. On note aussi $x_1, \ldots, x_i, \ldots, x_n$ les observations de X.

Pour une observation x_i donnée, on modélise alors la loi de X avec la fonction $f(x_i; \theta)$. Cette fonction vaut $f(x_i; \theta) = \mathbb{P}_\theta(X = x_i)$ si X est une v.a. discrète et $f(x_i; \theta) = f_\theta(x_i)$ si X est continue, où $f_\theta(x_i)$ est la densité de la loi en fonction de ses paramètres θ.

Pour des paramètres θ donnés (ex : moyenne et écart type d'une loi normale), $f(x_i; \theta)$ sera alors d'autant plus élevée que x_i a des chances d'être tirée en fonction des θ.



La vraisemblance des paramètres θ en fonction des observations $x_1, \ldots, x_n$ est alors :

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Dans l'exemple de pile ou face, supposons que l'on souhaite vérifier empiriquement si une pièce est équilibrée ou non. On modélisera $\mathbb{P}(X = 1) = f(X_i = 1; \theta = \{p\}) = p$ et $\mathbb{P}(X = 0) = f(X_i = 0; \theta = \{p\}) = 1 - p$, puis on réalisera n observations de X en tirant à pile ou face. La vraisemblance sera alors $L(\theta = \{p\}) = \prod_{i=1}^{n}(1_{X_i=1}p + 1_{X_i=0}(1-p))$. Supposons que sur $n = 10$ tirages, on observe 4 'piles' et 6 'faces'. En simplifiant légèrement les notations, la vraisemblance du paramètre p par rapport à notre modèle et nos observations empiriques sera alors $L(p) = p^4(1-p)^6$. Calculons alors la vraisemblance pour plusieurs valeurs de p : $L(0.2) = 0.00042$, $L(0.5) = 0.00098$, $L(0.8) = 0.00002$. De ces trois valeurs, $p = 0.5$ semble le plus vraisemblable.

De manière générale, on calculera le maximum de vraisemblance :

$$\hat{\theta} = \arg\max_{\theta} L(\theta)\,,$$

qui renverra les paramètres les plus vraisemblables en fonction des observations et de la loi choisie.

Dans l'exemple de pile ou face, la meilleur vraisemblance sera obtenue pour $p = 0.4$ avec $L(0.4) = 0.00119$. Si la pièce est bien équilibré, le nombre de 'piles' et de 'faces' obtenus sera de plus en plus proche quand $n \to +\infty$ et $p = 0.5$ aura ainsi la meilleur vraisemblance.

Pour des raisons numériques, il est aussi bien pratique de maximiser la log-

vraisemblance au lieu de la vraisemblance brute :

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \log\left(L(\theta)\right) \\ &= \arg\max_{\theta} \log\left(\prod_{i=1}^{n} f(x_i;\theta)\right) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i;\theta)\end{aligned}$$

Vu que la fonction *log* est strictement croissante les paramètres optimum $\hat{\theta}$ seront les mêmes avec la log-vraisemblance ou la vraisemblance.

Estimation au sens des moindres carrés

On suppose disposer d'observations $\{y_i\}_{i=\{1,\dots,n\}}$ que l'on souhaite prédire/deviner à partir de observations correspondantes $\{x_i\}_{i=\{1,\dots,n\}}$, où chaque y_i correspond à x_i (voir l'exemple introductif par exemple). Dans ce cours, et très souvent en apprentissage automatique, on va alors optimiser les paramètres θ d'un modèle f_θ pour prédire au mieux les y_i avec $\hat{y_i} = f_\theta(x_i)$.

Faisons l'hypothèse que les erreurs d'approximation du modèle $e_i = y_i - f_\theta(x_i)$ suivent une loi normale centrée, *i.e.* $e_i \sim \mathcal{N}(0,\sigma)$. Ce choix par défaut est commun et semble raisonnable quand f_θ est bien calibré. Nous pouvons alors utiliser le principe de maximum de vraisemblance pour estimer les paramètres θ du modèle f_θ.

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) \\ &= \arg\max_{\theta} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} e_i^2\right) \\ &= \arg\min_{\theta} \sum_{i=1}^{n} e_i^2 \\ &= \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2\end{aligned}$$

Cette technique d'estimation est celle dite au sens des moindres carrés. Nous la retrouvons très couramment en apprentissage automatique et son interprétation est particulièrement intuitive. Elle doit notamment sa popularité au fait qu'il est aisé de calculer son gradient par rapport aux paramètres θ si on sais calculer le gradient de f_θ par rapport à θ :

$$\nabla_\theta e_i^2 = 2(y_i - f_\theta(x_i))\nabla_\theta f_\theta(x_i)$$

Cela ouvre la porte aux techniques d'optimisation par descente de gradient qui sont quasi systématiques en apprentissage automatique.

Pour un public avisé, il faudra se souvenir du fait que la pertinence de l'estimation de paramètres d'un modèle au sens des moindres carrés repose sur une hypothèse de normalité de l'erreur.