ML 01 Introduction

1 Contexte

A l'occasion des trois premières révolutions industrielles, des tâches, auparavant réservées au travail manuel de l'Homme, ont été automatisées. Il semble envisageable d'associer au tournant du 21ème siècle une quatrième révolution portée par l'automatisation de la capacité à prédire, essentielle pour le processus de prise de décision dans les secteurs de l'industrie, du commerce et des services.

Cette transformation se fonde sur des évolutions scientifiques et techniques majeures. Elle est ainsi associée à une discipline, le machine learning ou apprentissage automatique de modèles prédictifs par extrapolation à partir de données générées par des processus physiques, numériques ou biologiques. Ces développements algorithmiques, en particulier la redécouverte des réseaux de neurones profonds, ont révélé sous un nouveau jour leur potentiel autour des années 2010 grâce, d'une part, à la création de jeux de données volumineux dans des domaines variés comme la reconnaissance de la parole, la vision par ordinateur, les données multimedia, le traitement de la langue naturelle, la robotique, les véhicules autonomes... et, grâce d'autre part, à une croissance rapide des capacités de calcul et de stockage aux coûts toujours plus abordables.

Par ailleurs, cette automatisation des prédictions s'accompagne d'un renouveau des formes de jugement dans les processus de prise de décision avec un couplage de plus en plus fin entre d'un côté, des experts humains et de l'autre, des chaînes de traitement automatique des données qui aboutissent sur la mise en production d'algorithmes prédictifs. Pour la réussite de cette intégration, les compétences de l'ingénieur informaticien sont essentielles. Ce dernier, puisqu'il comprend en profondeur le fonctionnement et les limites des algorithmes qu'il déploie, est capable d'en mesurer les risques et les biais pour éclairer le jugement de ceux qui utiliseront ses réalisations logicielles pour prendre des décisions.

Ainsi, ce cours introductif à l'apprentissage automatique a pour objectif d'offrir des connaissances fondamentales et des compétences pratiques qui aideront l'ingénieur à tenir ce rôle essentiel.

2 Objectifs

La discipline de l'apprentissage automatique, ou machine learning, élabore des algorithmes et des méthodes pour découvrir des régularités dans des données multidimensionnelles afin, entre autres, d'automatiser la prédiction. Elle peut se subdiviser en trois catégories. D'abord, l'apprentissage non (ou semi) supervisé qui s'attache à découvrir des structures dans les données non étiquetées à travers des approches comme le clustering, la réduction dimensionnelle, les modèles génératifs... Ensuite, l'apprentissage par renforcement, dans le cadre duquel un agent interagit avec son environnement en adaptant son comportement pour maximiser une fonction de récompense. Enfin, l'apprentissage supervisé, qui fait l'objet de ce module, a quant à lui pour objectif d'apprendre à prédire l'association entre un objet décrit selon plusieurs dimensions et une étiquette.

Par exemple, il peut s'agir d'associer aux quartiers d'une ville le prix médian d'un logement. Dans ce cas, un quartier peut être décrit par la proportion de zones résidentielles, le taux de criminalité, le nombre moyen de pièces par habitat, etc. Ici, nous faisons face à un problème dit de « régression » où la valeur à prédire, autrement l'étiquette associée à chaque observation, est continue. Lorsque la variable à prédire est discrète, il s'agit d'un problème dit de « classification », comme détecter un objet dans une image ou décider si une transaction bancaire risque d'être une fraude. Nous considérerons ces deux catégories de problèmes.

Ainsi, ce cours a pour objectif d'introduire quelques concepts fondamentaux de l'apprentissage supervisé et de montrer leurs interconnexions variées dans le cadre de développements algorithmiques qui permettent d'analyser des jeux de données dans une visée avant tout prédictive. Ainsi, les propositions théoriques mèneront

à l'écritures de programmes qui implémentent ou utilisent quelques modèles essentiels de l'apprentissage supervisé.

Pour faciliter l'acquisition des connaissances, le cours est accompagné de notebooks manipulables, rédigés dans le langage de programmation R. Ces mises en pratique systématiques doivent permettre de faire le lien entre des concepts fondamentaux et leur application dans des projets d'analyse de données.

3 Ajustement de courbe

$\mathbf{x}^{(1)} \dots \mathbf{x}^{(P)}$ sont des vecteurs de $\mathbb{R}^N$ associés aux valeurs, aussi appelées étiquettes, $y^{(1)} \dots y^{(P)}$ de $\mathbb{R}$. Nous cherchons une fonction $f(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}$ qui modélise la relation entre les observations $\mathbf{x}$ et les étiquettes y.

La fonction f peut avoir une forme paramétrique, comme par exemple :

$$f(\mathbf{x}) = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_N x_N$$

Si $P = N + 1$, les paramètres $a_0, a_1, \dots a_N$ sont solution d'un système linéaire :

$$\begin{cases} y^{(1)} & = a_0 + a_1 x_1^{(1)} + a_2 x_2^{(1)} + \dots + a_N x_N^{(1)} \\ y^{(2)} & = a_0 + a_1 x_1^{(2)} + a_2 x_2^{(2)} + \dots + a_N x_N^{(2)} \\ \dots & = \dots \\ y^{(P)} & = a_0 + a_1 x_1^{(P)} + a_2 x_2^{(P)} + \dots + a_N x_N^{(P)} \end{cases}$$

Ce système s'écrit également sous forme matricielle :

$$\begin{pmatrix} 1 & x_1^{(1)} & \dots & x_N^{(1)} \\ 1 & x_1^{(2)} & \dots & x_N^{(2)} \\ \dots & \dots & \dots & \dots \\ 1 & x_1^{(P)} & \dots & x_N^{(P)} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_N \end{pmatrix} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(P)} \end{pmatrix}$$

Chaque ligne i de la matrice du terme de gauche de l'égalité ci-dessus est le vecteur ligne $\mathbf{x}^{(i)T}$ avec l'addition d'un premier terme constant qui correspond au paramètre a_0. En nommant cette matrice $\mathbf{X}^T$, le système linéaire ci-dessus s'écrit :

$$\mathbf{X}^T \mathbf{a} = \mathbf{y}$$

Soit le cas particulier où x est un scalaire et f est un polynome de degré N :

$$f(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_N x^N$$

Avec $P = N + 1$ observations et les étiquettes associées $\left(x^{(k)}, y^{(k)}\right)$, les coefficients de ce polynôme sont solution d'un système linéaire :

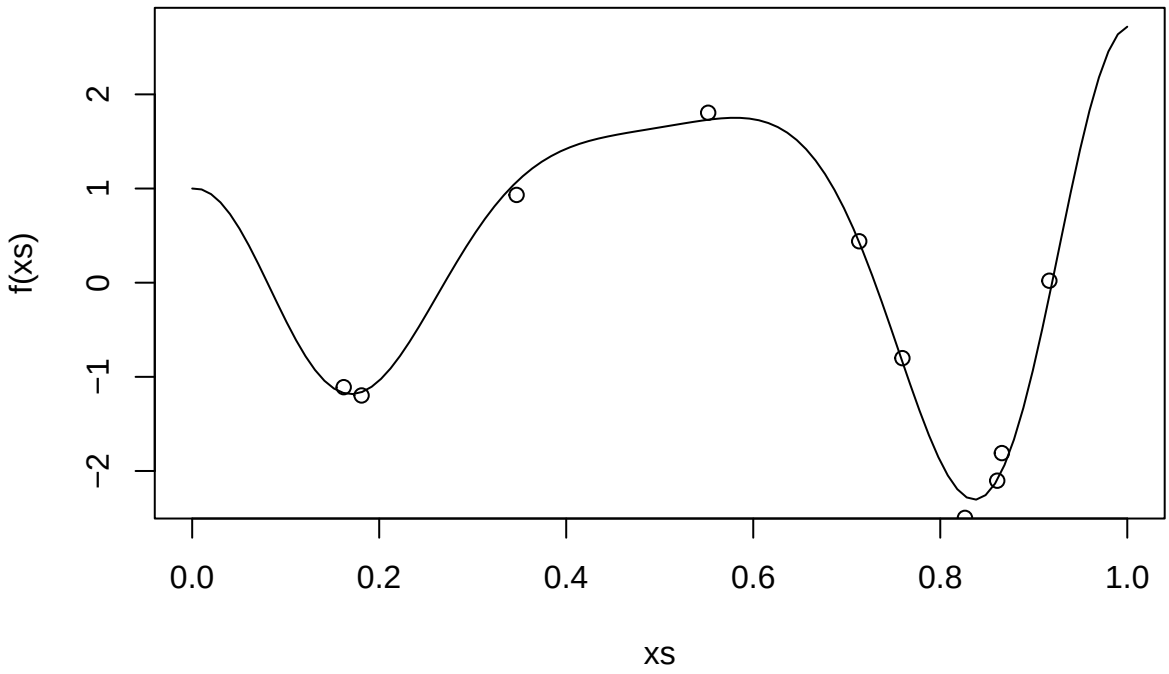
$$\begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & \dots & (x^{(1)})^N \\ 1 & x^{(2)} & (x^{(2)})^2 & \dots & (x^{(2)})^N \\ \dots & \dots & \dots & \dots & \\ 1 & x^{(P)} & (x^{(P)})^2 & \dots & (x^{(P)})^N \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_N \end{pmatrix} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(P)} \end{pmatrix}$$

La matrice du terme de gauche de l'égalité ci-dessus est traditionnellement appelée "matrice de Vandermonde".

4 Interpolation polynomiale sur un jeu de données synthétique

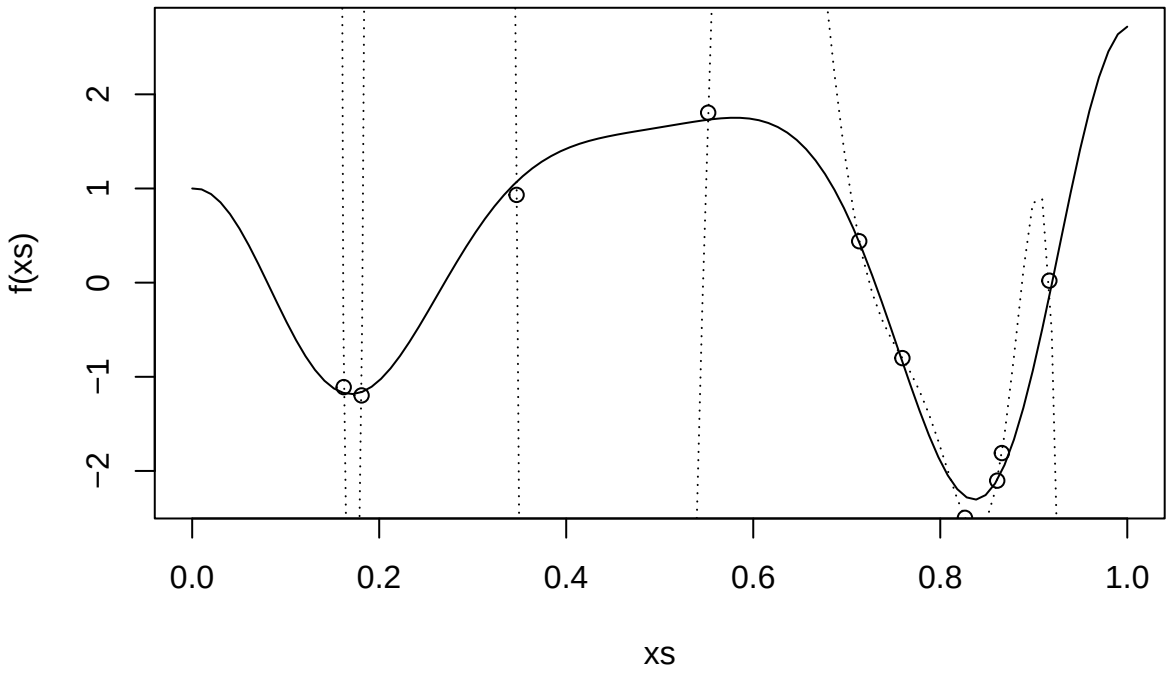
Soit un exemple de fonction non-linéaire, $f(x) = e^x \times cos(2\pi \times sin(\pi x))$, utilisée pour générer un jeu de données synthétique.

```
set.seed(1123)
# Image par f d'un échantillon uniforme sur l'intervalle [0,1], avec ajout d'un
# bruit gaussien de moyenne nulle et d'écart type 0.2
data = gendat(10,0.2)
plt(data,f)
```



Sur cet exemple, nous résolvons le système linéaire de Vandermonde pour découvrir un polynôme qui passe par chaque point du jeu de données.

```
# Résolution du système linéaire correspondant à la matrice de Vandermonde.
# coef contient les coefficients d'un polynôme qui passe par chaque point du jeu
# de données.
coef = polyreg1(data)
plt(data,f)
pltpoly(coef)
```



Il est improbable que ce polynôme, passant exactement par chaque observation, puisse offrir de bonnes capacités prédictives. Vérifier par exemple que, sur notre exemple synthétique, pour cinq points générés à partir de la fonction f et avec l'ajout d'un bruit gaussien (par exemple d'écart type 0.2), le polynôme découvert, de degré quatre, peut être très éloigné de la fonction génératrice. C'est un exemple du phénomène de sur-apprentissage. Pour limiter ce problème, nous cherchons à découvrir un polynôme de degré plus faible. Il ne passera pas exactement par toutes les observations, mais il prédira probablement mieux les étiquettes associées à de nouvelles observations.