Donnée et Apprentissage

Nicolas Vayatis
ENS Cachan

Table des matières

Notre professeur fait partie du labo CMLA (*Centre de Mathématiques et de Leurs Applications*) faisant parti du MLMDA (*Machine Learning and Massive Data Analysis*) étudiant principalement :

— Optimisation séquentielle et apprentissage actif

— Machine Learning sur des signaux temporels

— Processus de diffusion dans les réseaux avec des applications dans la santé publique (virus) et propagation d'information (virus informatique et réseaux sociaux)

Le laboratoire effectue principalement de la recherche (publications) et de la vulgarisation (pour les industries, les politiques, etc) et également des outils numérique (codes, portails, logiciel).

1 Introduction

Il existe trois grands problèmes avec les données :

— Classification (Apprentissage supervisé) : labels discret ($\rightarrow$ catégories)

— Régression (Apprentissage supervisé) : labels continus ($\rightarrow$ prédire un prix)

— Clustering ou segmentation (Apprentissage non supervisé $\rightarrow$ pas de labels)

Il existe également deux approches :

— L'approche statistique classique *paramétrique*

— L'approche par apprentissage *non paramétrique*

Principe : De par des données historiques $Z_i = (\underbrace{X_i}_{\text{mesures}} , \underbrace{Y_i}_{\substack{\text{label si} \\ \text{cadre supervisé}}})$, après un apprentissage A fournit une transformée $\hat{f}$ permettant une *règle de décision*, qui est le résultat de $\hat{f}(X_{n+1})$. La fonction $\hat{f}$ est testée, ce qui permet de juger ses performances avant son utilisation.

En statistique classique :

— On fixe la famille de lois qui génère les Z_i

— On en estime les paramètre de la loi, par exemple avec des méthodes de maximum de vraisemblance

— On fait du plug-in pour construire la règle de décision

En apprentissage :

— On fixe une structure de fonctions pour les règles de décision Par exemple

$$f_{\omega,\omega_0}(x) = \begin{cases} \text{"chat"} & \text{si} \quad \omega^T x + \omega_0 > 0 \\ \text{"chien} & \text{sinon} \end{cases}$$

— On fixe un critère de performance, par exemple :

$$\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{(f(X_i) = \text{"chien"} \wedge Y_i = \text{"chat"}) \vee (f(X_i) = \text{"chat"} \wedge Y_i = \text{"chien"}\}$$

— L'algorithme d'apprentissage doit minimiser $\hat{L}_n(f_{\omega,\omega_0})$ sur $\mathcal{F}$

Loi du mélange Soit X un vecteur aléatoire sur $\mathbb{R}^d$ suivant une loi de mélange. On considère K composantes de loi de densité f_k pour $1 \leq k \leq K$.

Un paramètre du mélange est $p = (p_1, ...p_K) \in \Delta_K \subset \mathbb{R}^K$ ou Δ_K est un simplexe de $\mathbb{R}^K$:

$$\Delta_K = \{p = (p_1, ..., p_K)^T \in \mathbb{R}_+^K : \sum_{k=1}^{K} p_k = 1\}$$

La densité du mélange est la loi de X :

$$f_X(x) = \sum_{k=1}^{K} p_k f_k(x) \qquad \forall x \in \mathbb{R}^d$$

Variables latentes (labels) On note $Y = (Y_1, ..., Y_k)^T$ avec $Y_K \in \{0, 1\}$, $\sum_{k=1}^{K} Y_k = 1$. Les Y_k sont des drapeaux exclusifs.

Quels Problème décisionnels ?

1. Clustering ou classification non supervisée :

1.1 Expectation Maximisation

Definition 1 (Maximum de vraisemblance)**.** *Soit $Z_1, ..., Z_n$ des variables aléatoires indépendantes et de même loi $z \mapsto f(z, \theta*)$, $\theta*$ paramètre inconnu. On appelle maximum de vraisemblance l'estimateur de $\theta*$ défini par*

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f(Z_i, \theta)$$

On utilise souvent la log-vraisemblance du modèle du mélange gaussien pour l'observation X :

2 Apprentissage Statistique

Soit X le vecteur des caractéristiques et Y la classification en sortie du modèle. On modélise

$$Y = f(X) + \epsilon$$

Avec ϵ une variation aléatoire.

Avec un f bien choisi on peut faire des prédictions sur de nouveaux points $X = (x_1, ..., x_p)$. On peut également comprendre quels composants $X = (X_1, X_2, ..., X_p)$ sont important pour expliquer Y, et lesquels sont hors de propos. Selon la complexité de f, il peut même être possible de comprendre dans quelle mesure chaque composant X_j de X affecte Y.

Existe-t-il une meilleure $f(X)$? Oui, l'espérance conditionnelle de Y sachant X soit $\mathbb{E}(Y|X = x)$. Cette fonction $f(x) \mapsto \mathbb{E}(Y|X = x)$ est appelé *fonction de régression*. C'est elle qui vérifie le minimum du critère *des moindres carrés :*

$$R_{(X,Y)}(f) = \mathbb{E}[\big(Y - f(X)\big)^2]$$

Theorem 1. *$f(x) \mapsto \mathbb{E}(Y|X = x)$ réalise le minimum de $R_{(X,Y)}(f) = \mathbb{E}[\big(Y - f(X)\big)^2]$.*

$\epsilon = Y - f(x)$ est *l'erreur irréductible* : même lorsque l'on connait $f(x)$, on fera toujours des erreurs de prédiction !

Soucis : on ne connait pas la loi des données : il faut *l'estimer*. Pour ce faire, on prend une fenêtre et $\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$ où $\mathcal{N}(x)$ est un voisinage de x.

La malédiction de la dimension En prenant des intervalles fixes, en grande dimension il se passe "des choses bizarres" $\rightarrow$ les points deviennent très loin les uns des autres. Autrement dit, le volume nécessaire pour garder la même quantité de points à l'intérieur augmente fortement.

3 Régression Linéaire

Definition 2. *Un modèle linéaire de dimension p est un $p+1$-uplet de paramètres notés β_k. La fonction linéaire modélisée est alors :*

$$f_L(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

On peut également poser $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ *et ainsi*

$$f_{\beta,\beta_0} = {}^t\beta.x + \beta_0$$

Le critère empirique est alors

$$\hat{R}_n(f_{\beta,\beta_0}) = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - (\beta_0 + {}^t\beta X_i)^2\right)$$

Remarque : On peut également estimer de cette manière n'importe quelle fonction polynomiale :

$$\begin{aligned} f(X) &= \beta_0 + {}^t\beta.X \\ &= {}^t\tilde{\beta} \cdot \begin{pmatrix} 1 \\ X \end{pmatrix} \end{aligned}$$

Que l'on peut généraliser en

$$f(X) = {}^t\tilde{\beta}.g(X)$$

Avec par exemple $g(X) = \begin{pmatrix} 1 \\ X \\ X^2 \\ X^3 \end{pmatrix}$ pour un polynôme de degrés 3.

Problèmes possibles :

- Sur-apprentissage : apprendre des spécificités liées aux données et non à la structure implicite sous-jacente que l'on cherche à estimer
- Sous-apprentissage : ne pas apprendre suffisamment de caractéristique de la part des données

Il faut souvent faire des choix :

- Prédiction vs interprétabilité
- Sur vs sous apprentissage
- Parcimonie vs boite noire
- Monitorage de la règle de décision le dans temps (Réutilisabilité ?)

Comment estimer l'efficacité d'un modèle ? On peut découper en deux les données d'apprentissage, une partie pour l'apprentissage et l'autre pour le test.

Choix variance-bias Supposons que l'on veut adapter un modèle $\hat{f}(x)$ à des données d'apprentissages `Tr`, et soit (x_0, y_0) une observation tirée depuis la population. Si le vrai modèle est $Y = f(X) + \epsilon$ avec $f(x) = \mathbb{E}(Y|X = x)$ et ϵ un bruit indépendant de X, alors

$$\mathbb{E}\Big(Y - \hat{f}(X)\Big)^2 = \mathbb{V}\big(\hat{f}(X)\big) + [\text{Bias}(\hat{f}(X))]^2 + \mathbb{V}(\epsilon)$$

Avec $\text{Bias}(\hat{f}(x_0)) = \mathbb{E}(\hat{f}(X)) - f(X)$

Estimation des paramètres par les moindres carrés L'objectif est de minimiser $(y_i - \beta_0\beta_i x_i)^2$

Definition 3 (RSS)**.**

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1.x_1)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Definition 4 (t-statistique et p-value)**.**

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

$\mathbb{P}(|X| > |t|)$ *est une p-value.*

Problème	Supervisé	Non Supervisé	Objectif	Critère
Régression	$Y \in \mathbb{R}$		Y	Erreur de prédiction/Moindre carrées
Classification	$Y \in \{0, 1\}$		Y	Erreur de prédiction / Taux d'erreur
Clustering		X, Y non observé	Y	pas de critère clair
Scoring	$Y \in \{0, 1\}$		Y	AUC, courbe Roc, courbes ROC, courbe Précision/Rappel

4 Apprentissage non supervisé

Le principe est de découvrir des informations pour mieux visualiser les données ou de trouver des sous-groupes au sein de ces données.

Le principe est difficile car il n'existe pas vraiment de critère pour définir un "bon" cluster. Cependant l'apprentissage non supervisé "marche bien" $\rightarrow$ classifier les personnes ou détecter automatiquement les tags.