Introduction à l'apprentissage supervisé

I. Des données et des prédictions

Observations : $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$

(Données d'entraînement
Training Data)

x_i : Inputs :
- Images
- Sons
- Vidéos
- Textes
- ...

y_i : Outputs :
- Classification binaire
  $\in \{0,1\}$ ou $\in \{-1,1\}$
- Classification multiclasse
  $\in \{0,1,\dots,k-1\}$ ou $\in \{(1,0,0,\dots,0), (0,1,0,\dots,0), \dots, (0,\dots,0,1)\}$
- Valeur $\in \mathbb{R}^d$

Objectif :

À partir d'un x observé ou non pendant l'entraînement, être capable de "prédire" le y qui lui correspond le mieux

$$\boxed{y \simeq f(x)}$$

⚠ - y peut être une fonction aléatoire de x
- f peut être complexe
- On n'observe qu'un nombre fini de données d'entraînt
- La dimension peut être élevée

II. Formalisation Mathématique

La modélisation se fait via la théorie des probabilités

$$\mathbb{P} \text{ sur } \mathcal{X} \times \mathcal{Y}$$

entrées sorties.

Hypothèse d'indépendance

$$(x,y), (x_1,y_1), (x_2,y_2), \dots, (x_n,y_n) \overset{iid}{\sim} \mathbb{P}.$$

Objectif:

Estimer $\mathbb{P}(y|x)$ la loi des conséquences conditionnellement aux causes.

Ce problème peut parfois être simplifié en estimer

$$f : \mathcal{X} \to \mathcal{Y}$$

telle que $\mathbb{P}(y \mid x) \approx \delta_{f(x)}$, ou de manière simplifiée, $y \approx f(x)$.

III. Evaluation des performances

Étant donné une fonction $f : \mathcal{X} \to \mathcal{Y}$, comment évaluer ses performances ?

1) Fonction de perte

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

avec pour interprétation $\ell(y, z)$ = erreur induite par la prédiction de z alors que la vérité était y.

Exemples

- Classification : $\ell(y, z) = \mathbb{1}_{y \neq z}$
- Régression : $\ell(y, z) = (z - y)^2$ ou $\ell(y, z) = |z - y|$

2) Risque

- Risque moyen

$$\mathcal{R}(f) := \mathbb{E}_{(x,y)}\left(\ell\left(y, f(x)\right)\right)$$

Exemple

En classification binaire

$$\mathcal{R}(f) = 0 \times \mathbb{P}(f(x) = y) + 1 \times \mathbb{P}(f(x) \neq y)$$
$$= \mathbb{P}(f(x) \neq y)$$

⚠ On aimerait choisir un f qui minimise $\mathcal{R}(f)$. Le problème est que $\mathcal{R}(\cdot)$ utilise la vraie distribution des données qui n'est pas connue ! Il faudra utiliser un estimateur.

3) Risque de Bayes - Prédicteur de Bayes.

Proposition - Définition :

Le risque moyen $\mathcal{R}(\cdot)$ est minimisé par le prédicteur de Bayes $f_* : \mathcal{X} \to \mathcal{Y}$ tq

$$\forall x' \in \mathcal{X},\ f_*(x') = \underset{z \in \mathcal{Y}}{\operatorname{argmin}} \underbrace{\mathbb{E}\left(\ell(y, z) \mid x = x'\right)}_{=: r(z \mid x')}$$

Le risque de Bayes est défini comme le risque du prédicteur de Bayes.

$$\mathcal{R}^* := \mathcal{R}(f_*)$$

Démonstration :

$$\mathcal{R}(f) - \mathcal{R}^* = \mathcal{R}(f) - \mathcal{R}(f_*)$$

$$= \int_{\mathcal{X}} \underbrace{\left(r(f(x')|x') - \min_{z \in \mathcal{Y}} r(z|x') \right)}_{\geq 0} d\mathbb{P}(x')$$

$$\geq 0 .$$

$\square$

Exemples :

- Classification

$$f_*(x') \in \underset{z}{\operatorname{argmin}}\, \mathbb{P}(y \neq z \mid x = x')$$

$$= \underset{z}{\operatorname{argmax}}\, \mathbb{P}(y = z \mid x = x').$$

- ~~Class~~ Régression :

$$f_*(x') \in \underset{z}{\operatorname{argmin}}\, \mathbb{E}\left((y-z)^2 \mid x = x' \right)$$

$$= \underset{z}{\operatorname{argmin}}\, \mathbb{E}\left((y - \mathbb{E}(y|x=x'))^2 \mid x = x' \right) + \left(z - \mathbb{E}(y|x=x') \right)^2$$

donc $f_*(x') = E(y \mid x = x')$

III. Paradigmes classiques d'apprentissage.

1) Minimisation du risque empirique.

$\forall f,\ R(f) \approx \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) =: \hat{R}(f)$

"donc" $\underset{f}{\operatorname{argmin}}\, R(f) \approx \underset{f}{\operatorname{argmin}}\, \hat{R}(f)$

L'objet $\hat{R}(\cdot)$ s'appelle le risque empirique.

2) Méthodes génératives

Étape 1: Estimer $\mathbb{P}(x \mid y)$ et $\mathbb{P}(y)$

Étape 2: Déduire $\mathbb{P}(y \mid x)$ par la formule de Bayes

$$\mathbb{P}(y \mid x) \propto \mathbb{P}(x \mid y)\,\mathbb{P}(y)$$

3) Méthodes par moyennes locales

Idée : Approximer directement $f_*(x)$.

- **Classification** : $f_*(x') = \arg\max_j P(y=j \mid x=x')$

↳ Méthode des k-plus proches voisins.

Pour un x', prédire la classe majoritaire parmi les k plus proches voisins de x' dans le jeu d'apprentissage.

- **Régression** : $f_*(x') = \mathbb{E}(y \mid x=x')$

↳ Méthode des k-plus proches voisins.

Pour un x', prédire la valeur moyenne des y des k plus proches voisins de x' dans le jeu d'apprentissage.

4) Méthodes Probabilistes - Maximum de Vraissemblance.

Idée : Faire une hypothèse sur la classe de mesures de probabilités pertinentes pour le problème, et choisir celle qui maximise ~~de~~ les chances d'observer le jeu d'entrainement.

Si $\mathbb{P}(x, y \mid \underset{\substack{\uparrow \\ \text{paramètre de la loi}}}{\theta})$, $\theta \in \Theta$ est une paramétrisation de la loi,

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \; \mathbb{P}((x_1, y_1), \dots, (x_n, y_n) \mid \theta)$$

$$\underset{\text{indépendance}}{=} \underbrace{\mathbb{P}(x_1, y_1 \mid \theta) \times \dots \times \mathbb{P}(x_n, y_n \mid \theta)}_{=: \mathcal{L}((x_1, y_1), \dots, (x_n, y_n) \mid \theta)}$$

Fonction de vraisemblance (likelihood)

de manière équivalente,

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \; \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ln\left(\mathbb{P}(x_i, y_i \mid \theta)\right)}_{\text{log-vraisemblance (log-likelihood)}}.$$

IV. Exemples :

1) Régression linéaire par minimisation du risque empirique

Modèle : $y = \theta^T x + \varepsilon \leftarrow$ bruit.

$$\left(\begin{array}{c} \text{Perte quadratique} \\ + \text{Modèle linéaire} \end{array} \right) \Rightarrow \hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i)$$

2) Méthodes génératives, cas gaussien (en 1D)

$$y \sim \mathcal{B}\left(\tfrac{1}{2}\right) \qquad x|y \sim \mathcal{N}(y, 1)$$

i.e. $p(x|y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-y)^2}{2}\right)$

Dans ce cas simplifié, nous connaissons la loi de $x|y$ et la loi de y. Nous n'avons donc pas à l'estimer.

Par la formule de Bayes,

$$\mathbb{P}(y=0|x) \propto \frac{1}{2} \exp\left(-\frac{x^2}{2}\right)$$

$$\mathbb{P}(y=1|x) \propto \frac{1}{2} \exp\left(-\frac{(x-1)^2}{2}\right)$$

Ainsi, $\frac{\mathbb{P}(y=0|x)}{\mathbb{P}(y=1|x)} \geqslant 1$ ssi $\frac{(x-1)^2}{x^2} \geqslant 1$ ssi $\frac{1}{2} \geqslant x$

3) Maximum de vraisemblance - Régression logistique

Fonction sigmoïde : $\sigma(x) := \frac{1}{1+e^{-x}}$

Modèle logistique : $p(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$ $\quad y \in \{\cancel{0}, 1\}$

$$= \sigma(\theta^T x) = \sigma(y\theta^T x)$$

remarque

$$p(y=-1) = 1 - \sigma(\theta^T x)$$

$$= 1 - \frac{1}{1+e^{-\theta^T x}} = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}} = \frac{1}{1+e^{\theta^T x}}$$

$$= \sigma(y\theta^T x)$$

donc $p(y|x) = \sigma(y\theta^T x)$.

Maximum de vraissemblance:

$$\hat{\theta} \in \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ln\left(\frac{1}{1+e^{-y\theta^T x}}\right)$$

$$= \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ln\left(1+e^{-y\theta^T x}\right)$$