UNIVERSITÉ CÔTE D'AZUR | CAMPUS VALROSE

ANNEE UNIVERSITAIRE 20__ / 20__

Niveau d'études (Ex : 2ème année) : _____

Épreuve de : _____

Note de l'épreuve /

(1) Le candidat doit inscrire ici : ses noms, prénoms, lieu et date de naissance, puis rabattre suivant le pointillé le coin de la copie et le coller.

Il est interdit au candidat de signer sa copie ou d'y inscrire un signe quelconque pouvant en indiquer la provenance.

Ne pas mettre de colle sur la partie grisée

(1) Nom : .....

Prénoms : .....

Né(e) à : .....

le : .....

Ouvrir ici ▲

Nombre d'intercalaires _____ A _____, le 16/01/ 20 24

loloee univ. fr

APPRENTISSAGE STATISTIQUE

Préambule

Apprentissage Statistique = Machine Learning (ML)
= Data Science

Historique

* 1930 - 1970 : Stat Inférentielle } Stat paramétrique
  - Expériences planifiés
  - Test de modèle.

* 1970 - 1980 : 1ère percée des stats non paramétriques
  - 1ère apparition des réseaux de neurones

* 1990 : - Jeux de données non planifiés.
  - Apprentissage des ML

* 2000 : Explosion du nombre de variables
(p = nbre de variable >>> n = nombre de données)

- Sélection d'un modèle
- Sélection de dimension.

*2010 : Explosion des ressources en données.

- Explosion des ressources en calcul

→ Retour des réseaux de neurones
→ Théorie de l'optimisation

II/ Estimation VS Apprentissage

* Estimation : On cherche à approcher au mieux le "vrai" modèle.

• Si $y = f(x) + \varepsilon$ → bruit

On cherche $\hat{f}$ très proche de f.

* Apprentissage : On cherche à prédire y.

- $y = f(x) + \varepsilon$.

On veut $\hat{y}$ très proche de y.

II/ Apprentissage Supervisé VS non Supervisé.

* Supervisé : Soit le couple (X, Y) tel :

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

On observe un échantillon iid, $(x_i, y_i)_{i=1}^{n}$

On veut prédire Y à partir de X

2 cas possibles.

Régression	Classification
$Y \in \mathbb{R}$ ($\mathbb{R}^d$)	$Y \in \mathbb{N}$ ($\mathbb{Z}$)

* Non Supervisé : On cherche à "comprendre" X à partir d'un échantillon $(x_i, y_i)_{i=1}$ iid.

- Estimation de densité
- Clustering
- Recherche de lien causal

APPRENTISSAGE SUPERVISE

Soit (X, Y) un couple de v.a

*But: prédire Y à partir de X.

*Matériel: des observations $\{(x_i, y_i)\}_{i=1}^{n}$

Choix possible de X et Y

X peut être	Y peut être
numérique	numérique ($\mathbb{R}$): Régression
qualitatif	qualitatif (ou discret): classification
mix des (2)	

I-1) Le problème basique

A partir des observations $\{(x_i, y_i)\}_{i=1}^{n}$,

On veut construire un prédicteur (machine) φ^n tel que $\hat{y}_{new} = \varphi^n(x_{new})$.

On apprend φ^n à partir des observations

Propriétés souhaitées

- Précision sur les données.

⇒ Les prédictions faites sur les données d'apprentissages doivent être proches des vraies valeurs (connues).

16.01.24

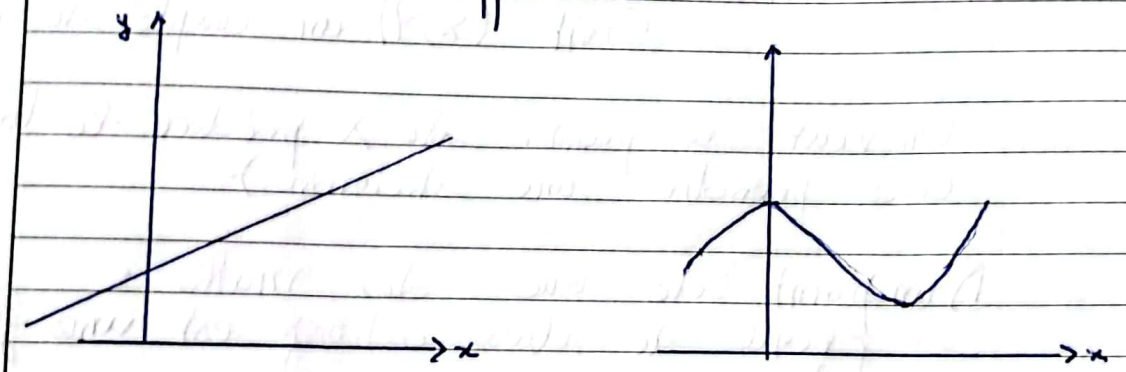
(2)

* Généralisabilité

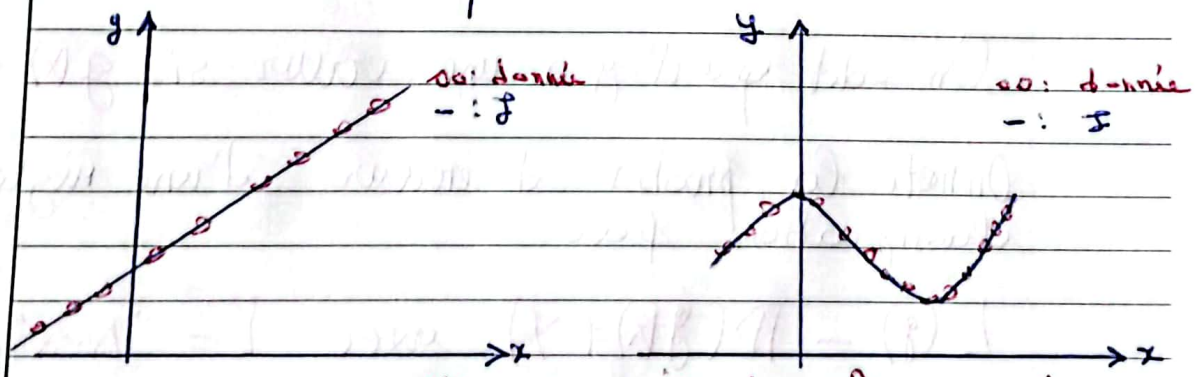
⇒ Les prédictions faites sur de nouvelles données doivent aussi être proches des vraies valeurs.

I.2) Sous apprentissage et Sur apprentissage

- Fonction à apprendre

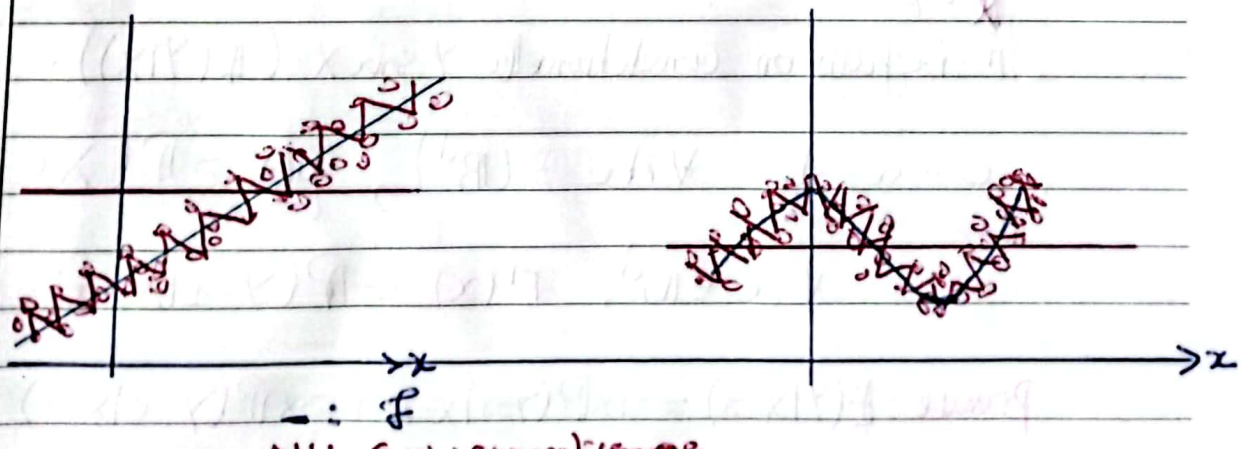


- Les données qu'on aimerait avoir



NB: on a jamais cela en pratique.

- Les données qu'on a :



- : f

NN: Sur apprentissage

- : Sous apprentissage.

Pour avoir de bonnes performances

* Avoir le meilleur algo (+)
* Avoir de meilleurs variables (+++)
* Avoir de meilleurs données (+++++)

CLASSIFICATION Binaire

I/ Généralités

Soit (X,Y) un couple de v.a.c $\mathbb{R}^d \times \{0,1\}$

On veut à partir de X prédire le label Y (i.e prendre une décision).

D'un point de vue des maths :
une {règle de classification / décision} est une fonction

$g : \mathbb{R}^d \longrightarrow \{0,1\}$.

On dit qu'il y a une erreur si $g(X) \neq Y$.

On note la proba d'erreur d'une règle de classification par :

$L(g) = \mathbb{P}(g(X) \neq Y)$ avec L = "Loss".

On note (μ, Γ) la loi du couple (X, Y)

$X \sim \mu$

Γ : régression conditionnelle Y sur X ($\mathbb{E}(Y|X)$).

c-à-d : $\forall A \in \mathcal{B}(\mathbb{R}^d)$, $\mu(A) = \mathbb{P}(X \in A)$

$\forall x \in \mathbb{R}^d$, $\Gamma(x) = \mathbb{P}(Y=1 | X=x) = \mathbb{E}(Y|X=x)$

Preuve : $\mathbb{E}(Y|X=x) = 1 \times \mathbb{P}(Y=1|X=x) + 0 \times \mathbb{P}(Y=0|X=x)$

16.04.24

Question ? Existe t il une règle meilleure que toutes les autres ? oui

I Règle de Bayes (Bonne règle)

On appelle règle de Bayes, la fonction g^* définie par:

$$g^*(x) = \begin{cases} 1, \text{ si } \mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x) \\ \qquad (\eta(x) > 1/2) \\ 0, \text{ sinon} \end{cases}$$

Théorème :

$\forall\, g : \mathbb{R}^d \longrightarrow \{0,1\}$ règle de décision

$$\mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y)$$

ie $L^* = L(g^*)$, $L^* = \inf_g (L(g))$.

Preuve :

Soit g une règle de décision

$$\mathbb{P}(g(X) \neq Y | X) = 1 - \mathbb{P}(g(X) = Y | X)$$

$$= 1 - \left(\mathbb{P}(Y=1, g(X)=1 | X) + \mathbb{P}(Y=0, g(X)=0 | X)\right)$$

$$= 1 - \left(\mathbb{1}_{g(X)=1}\, \mathbb{P}(Y=1|X) + \mathbb{1}_{g(X)=0}\, \mathbb{P}(Y=0|X)\right)$$

or $\eta(x) = \mathbb{P}(Y=1|X)$ et $\mathbb{P}(Y=0|X) = 1 - \eta(X)$.

D'où : $= 1 - \left(\mathbb{1}_{g(X)=1}\, \eta(X) + \mathbb{1}_{g(X)=0}\, (1-\eta(X))\right)$.

D'autre part, $\mathbb{P}(g(X) \neq Y|X) - \mathbb{P}(g^*(X) \neq Y|X)$

$$= \eta(x)\left(\mathbb{1}_{g^*(X)=1} - \mathbb{1}_{g(X)=1}\right) + (1-\eta(X))\left(\mathbb{1}_{g^*(X)=0} - \mathbb{1}_{g(X)=0}\right)$$

$\mathbb{1}_{g^*(X)=0} = 1 - \mathbb{1}_{g^*(X)=1}$

$$\text{D'où : } = \underbrace{(2\eta(X)-1)}_{A}\underbrace{\left(\mathbb{1}_{g^*(X)=1} - \mathbb{1}_{g(X)=1}\right)}_{B}$$

2 Cas possibles

* $A > 0 \Leftrightarrow g^*(x) = 1$
  ↳ $\eta(x) > \frac{1}{2}$
  $\Rightarrow B = 0$ ou $1 \Rightarrow AXB \geq 0$.

* $A < 0 \Rightarrow g^*(x) = 0$

  $\Rightarrow B = 0$ ou $-1 \Rightarrow AXB \geq 0$.

Dans tous les cas,

$\mathbb{P}(g(x) \neq Y | X) - \mathbb{P}(g^*(x) \neq Y | X) \geq 0$

d'où: $\mathbb{E}[\mathbb{P}(g(x) \neq Y | X) - \mathbb{P}(g^*(x) \neq Y | X)] \geq 0$

$\mathbb{E}(\mathbb{P}(g(x) \neq Y | X)) \geq \mathbb{E}(\mathbb{P}(g^*(x) \neq Y | X))$

$\mathbb{P}(g(x) \neq Y) \geq \mathbb{P}(g^*(x) \neq Y)$

D'où le résultat.

R_7:
$\mathbb{E}(\mathbb{P}(Y|X))$
$= \mathbb{P}(Y)$

Rq: On note L^* le risque de Bayes

$L^* = \mathbb{E}(\{\min(\eta(x), 1 - \eta(x))\})$

Problème : g^* dépend de la loi de (X,Y) qui est inconnue.

En la place on dispose d'un échantillon $S_n = \{(X_i, Y_i)\}_{i=1}^n$ iid de même loi que (X,Y)
On va chercher une règle $g_n(x, S_n)$ pour estimer Y à partir de X et S_n.

On note $g_n(x) = g_n(x, S_n)$

③ 16.01.24

Le processus de création de g_n est appelé apprentissage supervisé.

La proba d'erreur :

$$\mathcal{L}(g_n) = \mathbb{P}(g_n(X) \neq g_n(X, S_n) \mid S_n)$$

On va vouloir g_n qui minimise $\mathcal{L}$.

II| Consistance faible/universelle/forte

On veut approcher $\mathcal{L}^*$.

On voudrait que d'une certaine façon, $\mathcal{L}(g_n) \xrightarrow[n \to +\infty]{} \mathcal{L}^*$.

On parle de règle de consistance

Def 1 : Consistance faible/forte

Etant donnée une distribution de couple (X, Y), une règle g_n est dite faiblement consistante/convergente

* Si $\mathbb{E}[\mathcal{L}(g_n)] = \mathbb{P}(g_n(X) \neq Y) \xrightarrow[n \to +\infty]{} \mathcal{L}^*$

et fortement consistante

* Si $\mathcal{L}(g_n) \xrightarrow[n \to +\infty]{} \mathcal{L}^*$ p.s

① Consistance forte $\Rightarrow$ Consistance faible

② Consistance faible $\Leftrightarrow$ Consistance en proba.

⚠ Ces nombres sont valables à distribution de (X, Y) fixée.

Def 2 : Consistance universelle

g_n est universellement (fortement) consistante si elle l'est pour toute distribution de (X, Y)

Problème : Comment construire des règles vérifiant ces propriétés ?

II. 1) Outils probabilistes

* Théo de Borel Cantelli (adapté)

Soit X une v.a, X_n une suite de v.a. Si $\forall \varepsilon > 0$,

$$\sum_{i=1}^{n} \mathbb{P}(|X_i - X| > \varepsilon) < \infty \Rightarrow X_n \to X \text{ p.s}$$

(Ainsi, pour avoir une consistance d'une règle g_n, on va chercher à avoir des inégalités de type

$$\mathbb{P}(|\mathcal{L}_n - \mathcal{L}^*| > \varepsilon) \leq k e^{-nc\varepsilon^2}$$

(k, c des cte > 0)

II.3) Méthode Plug-in

Soit (X, Y) couple de v.a, μ loi de X

η : régression conditionnelle de $Y|X$.

$$g^*(x) = \begin{cases} 1 \text{ si } \eta(x) > 1/2 \\ 0 \text{ sinon} \end{cases}$$

L'idée du plug-in, c'est de remplacer $\eta(x)$ pour un estimateur $r_n(x)$.

$$g_n(x) = \begin{cases} 1 & \text{si } r_n(x) > 1/2 \\ 0 & \text{sinon} \end{cases}$$

L'espoir c'est que si $r_n(x)$ est proche de $r(x)$, alors g_n sera proche de g et L_n sera proche de L^*

*Théo:

Soit r_n un estimateur de r, g_n la règle associée, alors

$$0 \leq g_n - L^* \leq 2 \int_{\mathbb{R}^d} |r_n(x) - r(x)| \mu \, dx$$

Rq: La majoration par $2\int_{\mathbb{R}^d} |r_n(x) - r(x)| \mu \, dx$ n'est pas optimale.

Elle suffit pour montrer la constance mais ne permet pas d'étudier des vitesse de convergences.

En fait, on a pas vraiment besoin que Γ_n soit très proche de Γ sur tout $\mathbb{R}^d$, ce qui est critique c'est que : $\begin{cases} \Gamma_n(x) > 1/2 & \text{si } \Gamma(x) > 1/2 \\ \Gamma_n(x) < 1/2 & \text{si } \Gamma(x) < 1/2 \end{cases}$.

La zone critique est la zone où $\Gamma(x) \approx 1/2$.

Une hypothèse utile est souvent

$\mathbb{P}(\{|\Gamma(X) - 1/2| \leq t\}) \leq C t^{\alpha}$, $\forall 0 < t \leq t^*$ avec $\alpha = \text{Cte} > 0$.

II. 3) Théo de Stone (1977)

Le théo établit la consistance de certains estimateurs Γ_n de Γ.

On considère des estimateurs de type « moyenne locale », ie qui s'écrive

$\Gamma_n(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i$, $x \in \mathbb{R}^d$ avec

$(W_{n,1}^{(x)}, \dots, W_{n,n}^{(x)})$ un vecteur de poids tel que chaque $W_{n,i}^{(x)}$ est une fct mesurable de x et $(X_1, \dots, X_n)$, mais ⊥ de $Y_1, \dots, Y_n$.

Chp.

intro

Les paires (X_i, Y_i) dont X_i est proche de x sont plus intéressantes que les entiers.

Les poids devraient être plus importants pour les Y dont les X_i associés sont proches de x.

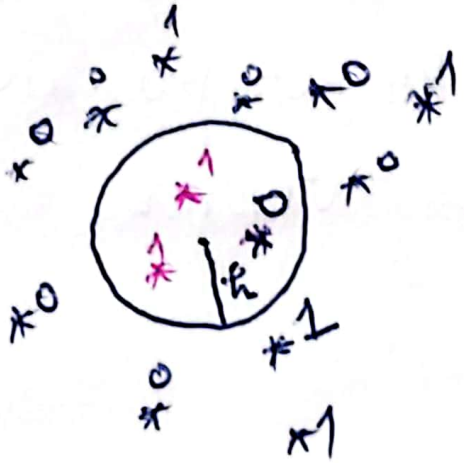
Ex: • Estimateur à noyau naïf

$$r_n(x) = \frac{\sum_{i=1}^n \mathbb{1}_{\{\|x - x_i\| \leq h\}} Y_i}{\sum_{i=1}^n \mathbb{1}_{\{\|x - x_i\| \leq h\}}} \quad \begin{matrix} h > 0 \\ h = \text{"fenêtre"} \end{matrix}$$

$$W_{n,i} = \frac{\mathbb{1}_{\{\|x - x_i\| \leq h\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{\|x - x_i\| \leq h\}}} = \begin{cases} \frac{1}{\#B(x,h)} & \text{si } x_i \in B(x,h) \\ 0 & \text{sinon} \end{cases}$$

En moyenne les Y_i dont les X_i sont dans une boule de centre x et de rayon h.

Ex:



$r_n(x) = 1 \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \frac{1}{3} * \cdots$

$r_n(x) = \frac{2}{3}$

* Estimation à noyau (Nadaraya-Watson)

$$\cdot W_{n,i} = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

$$\cdot r_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

où K = un noyau (une densité de proba)

Ex: Dans $\mathbb{R}$, K est un noyau Gaussien.

* Estimation des k plus proches voisins

- On fait la moyenne des k Y_i dont les x_i sont les plus proches de x.
- On peut prendre des poids ≥ 0 en fct des distances $\|x - X_i\|$, mais $W_{n,i} = 0$ si on est pas dans les k-plus proches voisins.

23.01.24

P) On considère les hypothèses suivantes

i) $\exists c_1 > 0$ tq pour toute fonction $f : \mathbb{R}^d \to \mathbb{R}$ (borélienne) avec $\mathbb{E}(|f(X)|) < \infty$.

$$\mathbb{E}\left(\sum |W_{n,i}(X)| \cdot \|f(X_i)\|\right) \leq c_1 \mathbb{E}(f(X)), \quad \forall n \geq 1.$$

ii) $\exists c_2 > 0$ tq $\mathbb{P}(\sum |W_{n,i}(X)| \leq c_2) = 1, \forall n \geq 1$

iii) $\forall \epsilon > 0, \sum |W_{n,i}(X)| \mathbb{1}_{\{\|X_i - X\| > \epsilon\}} \xrightarrow[n \to +\infty]{\mathbb{P}} 0$

iv) $\sum W_{n,i}(X) \xrightarrow[n \to +\infty]{\mathbb{P}} 1$

v) $\max_{1 \leq i \leq n} |W_{n,i}(X)| \xrightarrow[n \to +\infty]{\mathbb{P}} 0.$

• Si i) – iv) sont vérifiés pour toute distribution de X, alors Γ_n est universellement L^p consistant ($\forall p \geq 1$).
i.e $\forall p \geq 1$, $\mathbb{E}(|\Gamma_n(x) - \Gamma(x)|^p) \longrightarrow 0$

Rq: C'est presque une condition nécessaire et suffisante.

Discutons des hypos

i) hypothèse technique pour éviter de disposer que Γ est continue.

ii) et iv) disent que la somme des poids est bornée et tend $\underline{P} > 1$.

iii) $\Rightarrow$ que le poids global des valeurs hors d'un voisinage de $X \longrightarrow 0$.

Asymptotiquement, l'estimateur ne dépend que des données proches de lui-même (Moyenne locale)

v) $\Rightarrow$ que tous les poids deviennent petits quand n grandit

($\Rightarrow$ une x donnée ne peut décider seule)

Cas Classique

$(W_{n,1}(x), \dots, W_{n,n}(x))$ est un vecteur de $\mathbb{P}$.

(ie. $W_{n,i} \geq 0$, $\sum W_{n,i} = 1$).

ii) et iv) sont vérifiées.

Les autres sont une condition nécessaire et suffisante pour la CV-gence.

- On va utiliser ce théo pour notre problème de classification

Rappel : $\hat{g}(x) = \begin{cases} 1 \text{ si } r(x) > 1/2 \\ 0 \text{ sinon} \end{cases}$

Théo de Stone pour la classificat°

Si les hypo du théo de stone sont vérifiées pour toute distribution de (X,Y), alors $\hat{g}$ est universellement consistante.

ie $\forall (\mathcal{L}(\hat{g})) \longrightarrow \mathcal{L}^*$ $\forall$ distribut° (X,Y)

REGLES NON-PARAMETRIQUES ET DATA-SPLITTING

I | Règle des k plus proches voisins (kppv)

Soit (X,Y) couple de v.a $\to \mathbb{R}^d \times \{0,1\}$

$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ iid.

Pour x fixé ($x \in \mathbb{R}^d$), on réordonne l'échantillon $\mathcal{D}_n(x) = \{(X_{(i)}^{(x)}, Y_{(i)}^{(x)})\}$ fct° des distances croissantes $\|X_{(i)} - x\|$.

On a : $\|X_{(1)} - x\| \leq \|X_{(2)} - x\| \leq \dots \leq \|X_{(n)} - x\|$

Def 1 : Soit $(W_{n,1}, \dots, W_{n,n})$ un vecteur de poids de somme 1. La règle des kppv s'écrit $\forall x \in \mathbb{R}^d$.

$$\hat{g}_n(x) = \begin{cases} 1 \text{ si } \sum W_{n,i} \mathbb{1}_{\{Y_i = 1\}} > \sum W_{n,i} \mathbb{1}_{\{Y_i = 0\}} \\ 0 \text{ sinon} \end{cases}$$

Rq : On prend $W_{n,i} \forall \ominus$ ent.

$\Rightarrow$ On ~~vote~~ fait un vote pondéré des plus proches voisins.

Le choix classique est de prendre

$\{W_{n,i}\}_{i=1}^{n} = \{\underbrace{\frac{1}{k}; \frac{1}{k}; \dots; \frac{1}{k}}_{k \text{ fois}}; 0; \dots 0; 0\}$

On obtient la règle classique des k ppv.

$$\hat{g}_{n,k}(x) = \begin{cases} 1 & \text{si } \sum \mathbb{1}_{\{Y_i = 1\}} > \sum \mathbb{1}_{\{Y_i = 0\}} \\ 0 & \text{sinon} \end{cases}$$

⚠ On privilégi k: impaire

I. c) Théo de CVgence

* Théo de Stone

Si $k_n \xrightarrow[n \to +\infty]{} +\infty$ et $\frac{k_n}{n} \xrightarrow[n \to \infty]{} 0$, alors pour toute loi (X, Y), $\mathbb{E}[\mathcal{L}(g_{n,k})] \to \mathcal{L}^*$

C'est la consistance faible universelle.

- $k_n \to +\infty \Rightarrow$ On prend suffisament de voisins pour ne pas rester dans un voisinage temps de x.

Cela évite le sur-apprentissage.

- $\frac{k_n}{n} \to 0 \Rightarrow$ On prend relativement peu de voisins pour ne pas considéra un voisinage temps grand

Cela évite le sous-apprentissage.

⚠ Difficulté : Calibrer k.

Théo I.3.1 / $\lim_{n \to +\infty} \mathcal{L}(g_n, R) = \mathcal{L}_{RNN}$.

On connaît L_{RNN} en fct de Γ (R : inconnu)

Théo I.3.2 /

$$L^* \leq \dots \leq L_{(2k+1)nn} \leq L_{(2k-1)nn} \leq L_{3nn} \leq L_{1nn} \leq 2L^*$$

Cela est intéressant parce que ça dit que si $L^* \approx 0$, les règles R_{nn} sont bonnes.

Théo I.3.3 : Pour toute distribution de (X,Y) et $k \geq 3$

- $L_{Rnn} \leq L^* \left(1 + \frac{cte}{\sqrt{k}}\right)$ à peu près !

30.01.24

Solutions des paramètres
Data Splitting V-fold

Toutes les méthodes précédentes dépendent des paramètres à calibrer (k, h, ...).

I/ Risque Empirique

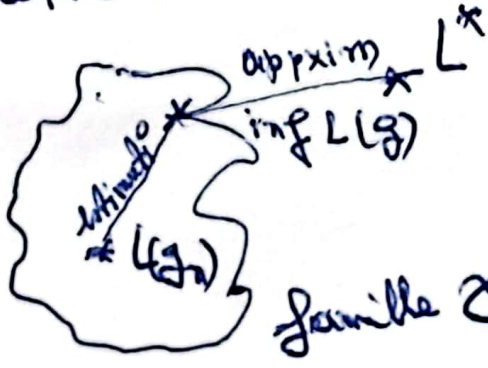
On décompose l'erreur

$$L(g_n) - L^* = \underbrace{\left(L(g_n) - \inf_{g \in \mathcal{C}} L(g)\right)}_{(1)} + \underbrace{\left(\inf_{g \in \mathcal{C}} L(g) - L^*\right)}_{(2)}$$

$\mathcal{C}$ une famille de règles possible (par exemple $\mathcal{C} = k_{PPV}$)

(1) = erreur d'estimation

(2) = erreur d'approximation
↳ erreur structurelle que l'on connaît en se restreignant à une famille de règles.
Elle est indépendante du choix du paramètre.



on se focalise sur l'erreur d'estimation.

On voudrait choisir g_n^* la meilleur règle dans $\mathcal{L}$:

$$g_n^* \in \underset{g \in \mathcal{L}}{\operatorname{argmin}}\, L(g).$$

Cela est impossible parce qu'on ne connait pas la loi de (X, Y) et donc on ne peut pas calculer $L(g)$.

A la place, on minimise le risque empirique :

$$\hat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{g(X_i) \neq Y_i\}}$$

approximation (Monte-Carlo)

$$\hat{g}_n^* \in \underset{g \in \mathcal{L}}{\operatorname{argmin}}\, \hat{L}_n(g)$$

I.1) Cas des Kppv

$g_{n,k}$ règle des kppv $\{(X_{(i)}^{(x)}, Y_{(i)}^{(x)})\}_{i=1}^{n}$ échantillon ordonné selon les distances croissantes $\|X_i - x\|$.

$$g_{n,k}(x) = \begin{cases} 1 & \text{, si } \sum_{i=1}^{k} \mathbb{1}_{\{Y_i = 1\}} > \sum_{i=1}^{k} \mathbb{1}_{\{Y_i = 0\}} \\ 0 & \text{sinon.} \end{cases}$$

$g^*_{n,k} \in \operatorname{argmin} \hat{L}_n(g)$

$\Leftrightarrow k^* \in \operatorname{argmin}_{k \in \{1,\dots,n\}} \hat{L}(g_{n,k})$

Ainsi, $g^*_{n,k} = g_{n,k^*}$.

Pourquoi ça ne marche pas ?

- On cherche à optimiser l'estimateur sur les données qui ont servi à la ~~construction~~ à construire.
- $\hat{L}_n$ n'est pas consistant car $\hat{L}_n$ et $g_{n,k}$ dépendent des mêmes données.
- $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{g_{n,k}(X_i) \neq Y_i\}} \longrightarrow \mathcal{L}^*$ si $g_{n,k} \perp\!\!\!\perp (X_i, Y_i)_{i=1}^n$
- On sur-apprend.

II/ Validation croisée

II-1) Hold-out (Data-splitting)

13·02·24 — APP - STAT

II. 3 | Consistance

* Data splitting

On veut calculer / contrôler

$$L(g_n^*) - \inf_{g_m \in \mathcal{L}_m} L(g_m)$$

↓ taille de A_m

$$L(g_n^*) = P(g_n^*(X) \neq Y \mid S_n)$$

$$L(g_m) = P(g_m(X) \neq Y \mid A_m) \quad (S_n)$$

$A_m \perp S_n$

Il est facile de montrer que

$$L(g_n^*) - \inf_{g_m \in \mathcal{L}_m} (L(g_m)) \leq 2 \sup_{g_m \in \mathcal{L}_m} |\hat{L}_{m,\ell}(g_m) - L(g_m)|$$

on contrôle facilement les choses

si $\#\mathcal{L}_m$ est fini.

$$P\left\{\sup_{g_m \in \mathcal{L}_m} |\hat{L}_{m,\ell}(g_m) - L(g_m)| > \varepsilon \mid S_n\right\} \leq 2 \times \#\mathcal{L}_m \, e^{-2\ell\varepsilon^2}$$

si $m = \ell = \frac{n}{2}$.

$$\mathbb{E}\left(L(g_m^*) - \inf_{g_m \in \mathcal{C}_m} L(g_m)\right) \leq 2\sqrt{\frac{\log(2e\,\#\mathcal{C}_m)}{n}}$$

Lorsque $\#\mathcal{C}_m$ est infini, on utilise la théorie de Vapnik-Chervonenkis

Dimension V.C

I/ Erreur Estimation / Approximation

Soit $\mathcal{C}$ une famille de règles et On veut choisir une règle dans $\mathcal{C}$.

Eq:
$L(g) = \mathbb{P}(g(X) \neq Y)$ soit proba de

$\inf_{g \in \mathcal{C}} \{\mathbb{P}(g(X) \neq Y)\}$.

On minimise le risque empirique

$$g_n^* \in \operatorname{argmin} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(X_i) \neq Y_i}}_{L_n(g)}$$

(en utilisant une procédure de validation croisée)

$$L(g_m^*) = \mathbb{P}(g_m^*(X) \neq Y \mid S_n)$$

on espère que

$$L(g_m^*) \simeq \inf_{g \in C} L(g) \quad [\text{vrai si } \#C \text{ est fini}]$$

- Rappel : $L(g_m^*) - L^*$

$$= \underbrace{\{L(g_m^*) - \inf_{g \in C} L(g)\}}_{\text{Terme d'estimation}} + \underbrace{\{\inf_{g \in C} L(g) - L^*\}}_{\text{Terme d'approximation}}$$

- On se concentre sur le terme d'estimation

Lemme (Important)

1) $L(g_n^*) - \inf_{g \in C} L(g) \leq 2 \sup_{g \in C} |\hat{L}_n(g) - L(g)|$

2) $|\hat{L}_n(g_n^*) - L(g_n^*)| \leq \sup_{g \in C} |\hat{L}_n(g) - L(g)|$

Preuve

2) $\Rightarrow$ trivial

1) $|L(g_n^*) - \inf_{g \in C} L(g)|$

$$\leq \underbrace{|L(g_n^*) - \hat{L}_n(g_n^*)|}_{①} + \underbrace{|\hat{L}_n(g_n^*) - \inf_{g \in C} L(g)|}_{②}$$

ona :

① $\leq \sup_{g \in C} |\hat{L}(g) - L(g)|$

$$② = \left| \inf_{g \in C} \hat{L}_n(g) - \inf_{g \in C} L(g) \right| \underset{(*)}{\leq} \sup_{g \in C} \left| \hat{L}_n(g) - L(g) \right|$$

Exo

Prouver (*) : $\left| \inf_{x \in A} (B(x)) - \inf_{x \in A} (g(x)) \right| \leq \sup_{x \in A} |f(x) - g(x)|$

Ça prouve (*).

Indication :

$f(x) - g(y) \leq |f(y) - g(y)| \leq \sup_x |f(x) - g(x)|$

de même, $g(y) - f(y) \leq$ ——— $\leq \sup_x |f(x) - g(x)| \quad \forall y$

$\Rightarrow f(x) \leq$ ——— $\forall y$

$g(y) \leq$ ——— $\forall y$

Prendre l'$\inf_x$ intelligement

II. Classe C de taille finie.

Tout repose sur l'inégalité de Hoeffding.

13.02.24 | APP - STAT

Théo II.1

$X_1, \dots X_n$ v.a bornées $\forall i \in [\![1,n]\!] \; (a_i, b_i)$

$P(X_i \in [a_i, b_i]) = 1$. On pose $S_n = \sum_{i=1}^{n} X_i$

$\forall \varepsilon > 0$

① $P\{S_n - E(S_n) \geq \varepsilon\} \leq e^{\frac{-2\varepsilon^2}{\sum (b_i - a_i)^2}}$

② $P(S_n - E(S_n) \leq -\varepsilon) \leq e^{\frac{-2\varepsilon^2}{\sum (b_i - a_i)^2}}$

③ $P[|S_n - E(S_n)| \geq \varepsilon] \leq 2\, e^{\frac{-2\varepsilon^2}{\sum (b_i - a_i)^2}}$

Preuve

① Admis

② Découle de ① (Exo).

③ découle de ① et ② (Exo).

Rq : $X_1, \dots, X_n$ n'ont pas besoin d'avoir la même loi.

Exemple fondamental (Exo)

$X_1, \dots, X_n$ iid $\sim B(p)$, alors

$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$

On a :

$X \sim \mathcal{B}(p) \Rightarrow [a_i, b_i] = [0,1]$,

$p = \mathbb{E}(X) = \mathbb{E}\left(\frac{S_n}{n}\right)$.

Soit $\varepsilon > 0$.

$\left|\frac{S_n}{n} - p\right| \geq \varepsilon \Leftrightarrow |S_n - np| \geq n\varepsilon$

~~$\Leftrightarrow \mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq n\varepsilon)$~~

$\Leftrightarrow \mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq n\varepsilon) \leq 2e^{-\frac{2\varepsilon^2 n^2}{\sum (b_i - a_i)^2 = n}}$

$\Leftrightarrow \mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq n\varepsilon) \leq 2e^{-2n\varepsilon^2}$.

Théo II.2

Si $\#\mathcal{C} \leq N$, alors $\forall \varepsilon > 0$

$\mathbb{P}\left(\sup_{g \in \mathcal{C}} |\hat{L}_n(g) - L(g)| > \varepsilon\right) \leq 2Ne^{-2n\varepsilon^2}$.

Preuve

- Loi de $Z_i = \mathbb{1}_{\{g(x_i) \neq y_i\}}$?
- Doeblin sur $\sum Z_i$ (exemple fondamental)
- $\mathbb{P}\left(\sup_{g \in \mathcal{C}} |\hat{L}_n(g) - L(g)| > \varepsilon\right) = \mathbb{P}(|\hat{L}_n(g) - L(g)| > \varepsilon)$

$\leq \sum_{g \in \mathcal{C}} \mathbb{P}(|\hat{L}_n(g) - L(g)| > \varepsilon)$ $\forall g \in \mathcal{C}$

Ona: $Z_i \sim \mathcal{B}(\underbrace{\mathbb{P}(g(X) \neq Y)}_{L(g)})$

$S_n = \sum Z_i$

$\frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{g(x_i) \neq y_i\}} = \hat{L}_n(g)$

$p = L(g)$.

D'où $\frac{S_n}{n} - p = \hat{L}_n(g) - L(g)$. Donc,

$\forall g, \mathbb{P}(|\hat{L}_n(g) - L(g)| > \varepsilon)$

$= \mathbb{P}(|\frac{S_n}{n} - p| > \varepsilon) \leq 2e^{-2n\varepsilon^2}$

avec S_n somme de Bernoulli $\mathcal{B}(p)$

. Aussi, $\mathbb{P}(\sup_{g \in C} |\hat{L}_n(g) - L(g)| > \varepsilon)$

$= \mathbb{P}(\forall g \in C, |\hat{L}_n(g) - L(g)| > \varepsilon)$

$\leq \sum_{g \in C} \mathbb{P}(|\hat{L}_n(g) - L(g)| > \varepsilon)$

$\leq \sum_{g \in C} 2e^{-2n\varepsilon^2} \leq 2Ne^{-2n\varepsilon^2}$

(parce que $\# C \leq N$)

NB: Cela est bien car nous pouvon intégrer en B.

Problème: gérer le cas $\#\mathcal{C} = \infty$

⇒ Théorie de V. C

III. La dimension V. C

III. 1) Passer de $\sup_{g \in \mathcal{C}}$ à $\sup_{A \in \mathcal{A}}$

• ν la loi du couple (X, Y).

$\forall A \in \text{Boréliens}(\mathbb{R}^d) \times \{0,1\}$

$\nu(A) = \mathbb{P}((X,Y) \in A)$

• ν_n la mesure empirique

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{(X_i, Y_i) \in A\}}$$

$$L(g) = \mathbb{P}(g(X) \neq Y) = \mathbb{P}((X,Y) \in \{(x,y) : g(x) \neq y\})$$

$$= \nu(\underbrace{\{(x,y) : g(x) \neq y\}}_{A_g})$$

$$\hat{L}_n(g) = \nu_n(A_g)$$

On peut réécrire

$$A_g = (\{x : g(x) = 1\} \times \{0\}) \cup (\{x : g(x) = 0\} \times \{1\})$$

On note : $\mathcal{A} = \{A_g, g \in \mathcal{C}\}$

$$\sup_{g \in \mathcal{C}} |L_n(g) - L(g)| = \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)|.$$

On se ramène à l'étude de la convergence de la mesure empirique

Propriétés.

1) $\forall A$ mesurable

$$|\nu_n(A) - \nu(A)| \xrightarrow[n \to +\infty]{P.S} 0 \quad (LGN)$$

2) $P(|\nu_n(A) - \nu(A)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}$ (Hoeffding)

3) Si $\mathcal{A}$ est de # finie (cardinal fini),

$$P(\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \varepsilon) \leq 2 \#\mathcal{A}\, e^{-2n\varepsilon^2}$$

(Théo II.2)

4) Si $\mathcal{A}$ est l'ensemble de tous les boréliens de $\mathbb{R}^d \times]0,1[$ (ENORME)

alors $\sup_{A \in \mathcal{A}} |V_n(A) - V(A)|$ ne peut CV pour $\forall V$.

Conclusion : Il faut contrôler la taille de $\mathcal{A}$

- On veut $\mathcal{A}$ grande pour l'erreur d'approximation
- On veut $\mathcal{A}$ pas trop grande pour contrôler $\sup_{A \in \mathcal{A}} |V_n(A) - V(A)|$

III.2) Coefficient de fractionnement et dimension V.C.

On se donne $\mathcal{A}$ une famille d'ensemble de $\mathbb{R}^d$, $Z_1, \dots, Z_n$ iid $\in \mathbb{R}^d$

$$V(A) = \mathbb{P}(Z \in A)$$

$$V_n(A) = \frac{1}{n} \sum \mathbb{1}_{\{Z_i \in A\}}$$

On veut des résultats sur la convergence de $\sup_{A \in \mathcal{A}} |V_n(A) - V(A)|$

Def :)

Soit $\mathcal{A}$ une famille d'ensemble mesurable de $\mathbb{R}^d$ et $(z_1, \dots, z_n) \in (\mathbb{R}^d)^n$

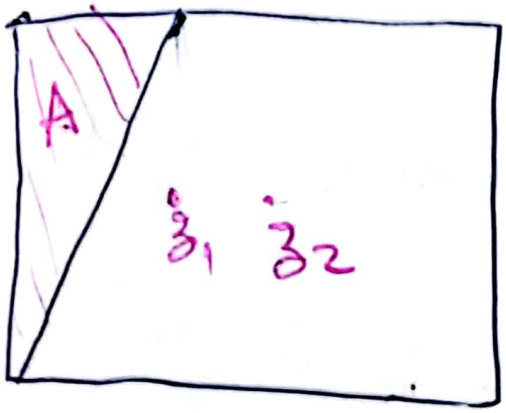
$N_{\mathcal{A}}(z_1, \dots, z_n)$ cardinal de $\{\{z_1, \dots, z_n\} \cap A, A \in \mathcal{A}\}$

Lorsque $N_{\mathcal{A}}(z_1, \dots, z_n) = 2^n$, on dit que $\mathcal{A}$ $\mathcal{A}$ pulvérise $\{z_1, \dots, z_n\}$.

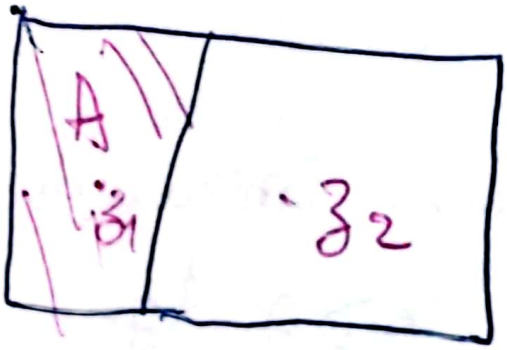
Exemple :

Soit $\mathbb{R}^2$, $\mathcal{A}$ = ensemble des 1/2 espaces de $\mathbb{R}^2$

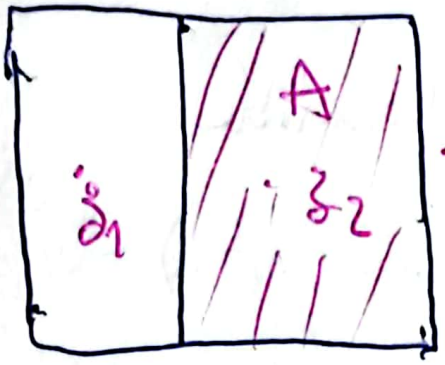
$\mathbb{R}^2$



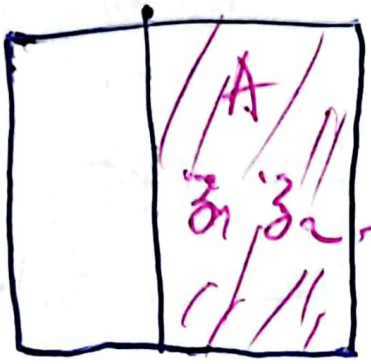
$\rightarrow \{\{z_1, z_2\} \cap A\} = \phi$



$\rightarrow \{\{z_1, z_2\} \cap A\} = \{z_1\}$



$\rightarrow \{\{z_1, z_2\} \cap A\} = \{z_2\}$



$\rightarrow \{\{z_1, z_2\} \cap A\} = \{z_1, z_2\}$

Ainsi, $\mathcal{N}(A\{z_1, z_2\}) = 2^2 = 4$

L'ensemble des 1/2 espaces pulvérise $\{z_1, z_2\}$.

Def 2

Le coeff de pulvérisation de $(A, \mathcal{C})$

$$\pi_A(n) = \max_{(z_1, \dots, z_n) \in (\mathbb{R}^d)^n} \mathcal{N}_{\mathcal{C}}(z_1, \dots, z_n)$$

Def 3

On une famille d'ensemble (mesurables) le plus grand entier $k \geq 1$ tq $S_{\mathcal{C}}(k) = 2^k$ est appelé dimension V.C

il est noté $\sqrt{_A}$.

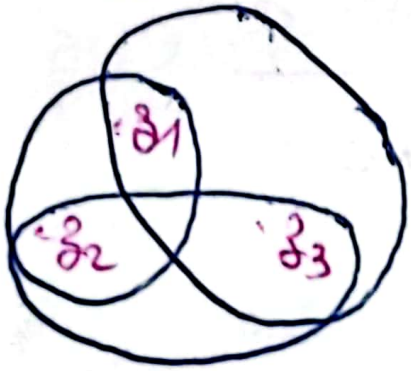
si $S_A(k) = 2^k \quad \forall k > 1$, $\sqrt{_A} = \infty$.

Rq :

- Plus $\sqrt{_A}$ est grande, et plus on va pouvoir pulvériser en grand nombre de points.
- Pulvériser n points, c'est être capable d'isoler n'importe quelle sous-partie de ces n points.

Exemple 1 $\mathcal{A}$ = disque de $\mathbb{R}^2$.

$\rightarrow S_A(2) = 2^2 = 4$



$\rightarrow S_A(3) = 2^3 = 8$

$\cdot z_1$ $\cdot z_2$

$\rightarrow S_A(4) < 16$.

$\cdot z_4$ $\cdot z_3$ Donc, $v_A = 3$

Exemples : A = ellipse de $\mathbb{R}^2$

$\cdot z_1$ $\cdot z_2$

z_5 $\cdots$ z_4

$S_A(4) = 16$

Rq : $S_A(5) < 32$.

Donc $v_A = 4$.

III. 3| Exemple important

1) si A est de Cardinal fini

$S_A(n) \leq \#A$, $\forall n \geq 4$.

pour définition $S_A(V_A) = 2^{V_A}$.

(car $V_A = \max_k$ tq $S_A(k) = 2^k$)

$\Rightarrow 2^{V_A} \leq \#A$

$\Rightarrow V_A \leq \dfrac{\ln(\#A)}{\ln(2)} = \log_2(\#A)$

2) dans $\mathbb{R}$, $A = \{]-\infty; x], x \in \mathbb{R}\}$

$S_0(0) = \emptyset \longrightarrow \{\emptyset, \{z_1\}\}$

$n = 2$.

~~$\{\emptyset, \{z_1\}, \{z_2\}\}$~~ $\{\emptyset, \{z_1\}, \{z_1, z_2\}\}$

$\Rightarrow S_0(2) < 4$

$\Rightarrow V_A = 1$.

$S_A(n) = n + 1$

3) dans $\mathbb{R}$, $A = [a, b]$, $a, b \in \mathbb{R}$.

$z_1, \dots z_n \Rightarrow \emptyset$

$\{z_1\}, \{z_2\}, \dots \{z_n\}$

$\{z_1, z_2\}, \{z_2, z_3\} \dots \{z_{n-1}, z_n\}$

$\{z_1, z_2, z_3\}, \{z_2, z_3, z_4\}, \dots$

1

$\rightarrow n$

$n - 1$

n

$\vdots$

$S_n(1) = 2^1 = 2$

$S_n(2) = 2^2 = 4$

$S_n(3) = 7 < 2^3 = 8$

$V_A = 2$

4) Résultat dans $\mathbb{R}^d$:

. $\mathcal{A} = \{]-\infty; x_1] \times \dots]-\infty; x_d] \}$

$V_{\mathcal{A}} = d$

. $\mathcal{A} = \{$ rectangles de $\mathbb{R}^d \}$, $V_{\mathcal{A}} = ?$

. $\mathcal{A}$ = famille des 1/2 espaces linéaires

$\{ x : a_1 x_1 + a_2 x_2 + \dots + a_d x_d + b \geq 0 \}$

$V_{\mathcal{A}} \leq d+1$.

20/02/24 : IV : Convergence uniforme de la mesure empirique

IV.1 : Le théorème de Vapnik-Chervonenkis

Théo : IV.1 (1971)

Thm

ν une mesure de proba sur $\mathbb{R}^d$,

$$\nu_n = \frac{1}{n} \sum_{i=0}^{n} \mathbb{1}_{\{X_i \in A\}},$$

$\mathcal{A}$ une famille d'ensemble mesurables.

$\forall n \geq 1$, $\varepsilon > 0$,

$$\mathbb{P}\left\{ \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \varepsilon \right\} \leq 8 S_{\mathcal{A}}(n) e^{\frac{-n\varepsilon^2}{32}}$$

©

Rq : ① Si $\mathcal{A}$ est de $S_{\mathcal{A}}(n) \leq \#\mathcal{A}$ et donc la série $S_{\mathcal{A}}(n) e^{\frac{n\varepsilon^2}{32}}$ est une série convergente ($\sum_{n=1}^{+\infty} < \infty$).

et donc $\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| \xrightarrow{p.s} 0$

② Si $S_{\mathcal{A}}(n)$ est un polynôme en n de degré fini, la série converge.

$\left[S_{\mathcal{A}}(n) = \sum_{i=0}^{d} a_i n^i \right)$.

③ Si $S_{\mathcal{A}}(n) = 2^n$, la série ne converge pas.

IV.2 Application

* Si $\mathcal{A} = \{]-\infty ; z] \ , z \in \mathbb{R} \}$, $S_{\mathcal{A}}(n) = n+1$

$$\mathbb{P}\left\{ \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \varepsilon \right\} \leq 8(n+1) e^{-\frac{n\varepsilon^2}{32}}$$

$$\mathbb{P}\left\{ \sup_{z \in \mathbb{R}} |F_n(z) - F(z)| > \varepsilon \right\}.$$

Convergence uniforme de la fonction de répartition empirique (Glivenko - Cantelli)

$$\sup |F_n - F| \longrightarrow 0 \quad p.s.$$

* Cas général

Lemme IV.1

X une v.a positive tq $\forall \varepsilon > 0$

$$\mathbb{P}(X > \varepsilon) \leq c\, e^{-2n\varepsilon^2}$$

Alors $\mathbb{E}[X^2] \leq \dfrac{\log(ce)}{2n}$

$$\mathbb{E}(X) \leq \sqrt{\frac{\log(ce)}{2n}}$$

Preuve

$$\mathbb{E}(X^2) = \int_0^{+\infty} \mathbb{P}(X^2 > \varepsilon)\, d\varepsilon$$

$$= \int_0^{u} \underbrace{\mathbb{P}(X^2 > \varepsilon)}_{\leq 1}\, d\varepsilon + \int_u^{+\infty} \mathbb{P}(X^2 > \varepsilon)\, d\varepsilon$$

$$\leq u + \int_u^{+\infty} \mathbb{P}(X > \sqrt{\varepsilon})\, d\varepsilon$$

$$\leq u + \int_u^{+\infty} c\, e^{-2n\varepsilon}\, d\varepsilon$$

Il faut trouver u qui minimise cette borne.

On trouve u tel que $\mathbb{E}(X^2) \leq \dfrac{\log(ce)}{2n}$.

Rq : $\mathbb{E}(f(X)) \geq f(\mathbb{E}(X))$ si f est convexe

$\mathbb{E}(f(X)) \leq f(\mathbb{E}(X))$ si f est concave

C'est l'inégalité de JENSEN

$Z = X^2$, $f(x) = \sqrt{x}$ concave

$\mathbb{E}(X) = \mathbb{E}(\sqrt{X^2}) \leq \sqrt{\mathbb{E}(X^2)}$

Corollaire IV.1 (IMPORTANT)

$$\mathbb{E}\left(\sup_{A \in \mathcal{A}} |V_n(A) - V(A)| \right) \leq 8 \sqrt{\frac{\log(8 e S_{\mathcal{A}}(n))}{n}}$$

IV.3 Coefficient de pulvérisation et dimension Vapnick - C

$S_{\mathcal{A}}(n)$ = Coef de pulvérisati

$= \max_{(z_1, \dots, z_n) \in (\mathbb{R}^d)^n} N_{\mathcal{A}}(z_1, \dots, z_n)$

$V_{\mathcal{A}}$ = dimension V.C de $\mathcal{A}$

= le plus grand k tq

$S_{\mathcal{A}}(k) = 2^k$

$= \infty$ si $S_{\mathcal{A}}(k) = 2^k \ \forall k$.

Propriété (Importante)

* Si $V(\mathcal{A})$ $V_{\mathcal{A}} < \infty$, $S_{\mathcal{A}}(n) \leq (n+1)^{V_{\mathcal{A}}}$

$\Rightarrow$ soit $S_{\mathcal{A}}(n) = 2^n \quad \forall n \geq 1 \quad (V_{\mathcal{A}} = \infty)$

soit $S_{\mathcal{A}}(n) \leq (n+1)^{V_{\mathcal{A}}}, \forall n \geq 1 \quad [V_{\mathcal{A}} < \infty)$

$\Leftrightarrow$ soit $S_{\mathcal{A}}(n) = 2^n$, soit il est polynomial en n.

Résultat final

- si $V_{\mathcal{A}} < \infty$, $\exists C > 0$ tq $\mathbb{E}\left(\sup_{A \in \mathcal{A}} |V_n(A) - V(A)|\right) \leq C\sqrt{\frac{V_{\mathcal{A}} \log(n)}{n}}$

- si $V_{\mathcal{A}} < \infty$, on a une convergence en $\sqrt{\frac{\log(n)}{n}}$

05/03/24 App stat ~~Maths~~ 9

II/ Problème de Robustesse

III. 1 Existence d'adversaires

Principal contributeur. Ian Goodfellow

2013. "Intriging properties of Neural Networks"

Possible de perturber très fortement les prédiction d'un réseau en perturbant très faiblement les données à classer.

La perturbation dépend de la données à classer.

La perturbation dépend de la donnée d'entrée.

2017: "Universal Adversarial perturbation" → perturbation dépend de la donnée d'entrée.

05.03.24 **IV. Robustification**

La principale méthode : Goodfellow 2015
« Adversarial training »

Principe général :

On considère une fct de coût J

(pour $J = \hat{L}_{on}$), $J(w, x, y)$
poids, entrée, label.

au lieu de minimiser J en w, on va regarder une version perturbée.

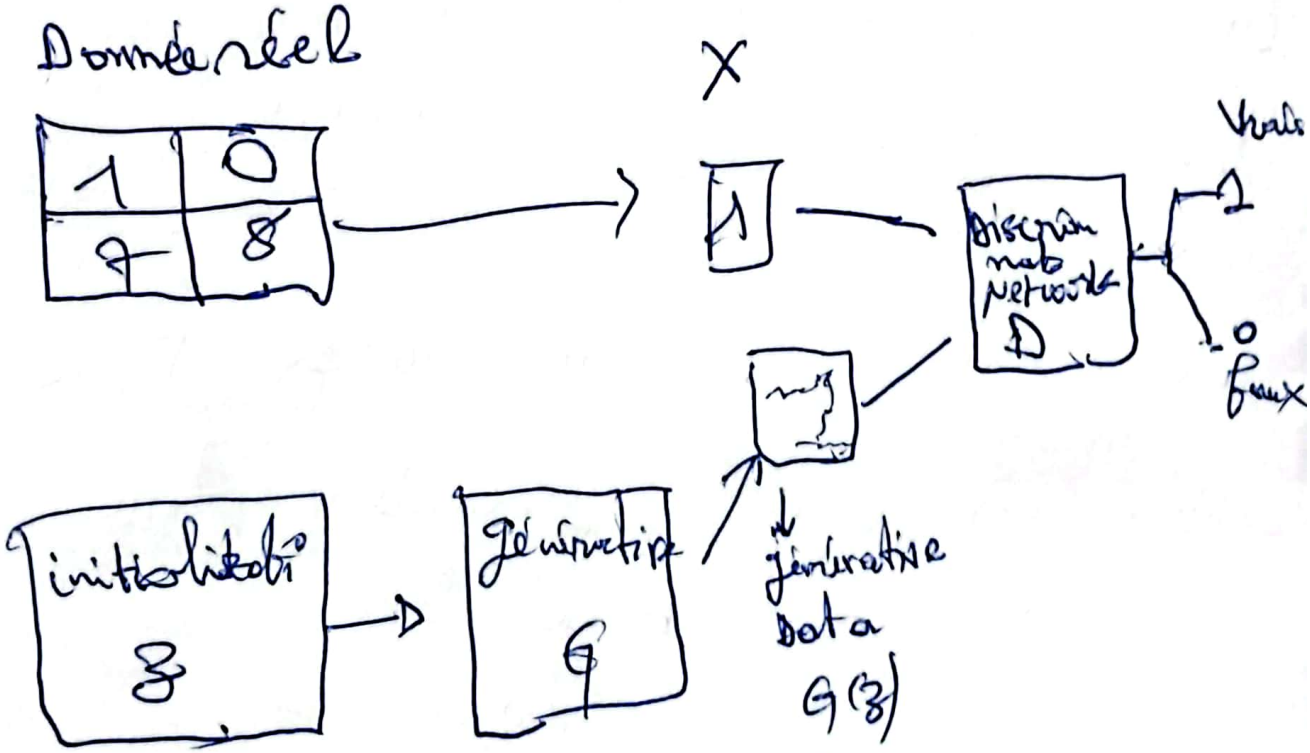
$\tilde{J}(w, x, y) = \alpha J(w, x, y) + (1-\alpha) J(w, \text{perturb}(x), y)$
$x + v$.

$\alpha \in [0, 1]$.

en essayant de prendre des perturbation qui « embêtent » le plus le réseau.

N.1. Generative adversarial Network

Le but est de générer des fausses données "vraies possible" les plus



La fct cible $V(G,D)$

$$V(G,D) = \mathbb{E}_x[\log(D(x))] + \mathbb{E}_z[\log(1-D(G(z)))]$$

On optimise le F.A.N en cherchant $\min_G \max_D V(G,D)$

En pratique

1) On part d'un jeu de données réelles et on entraîne D (on vise $D(x) = 1$).

2) On génère avec G et on entraîne D (on vise $D(G(z)) = 0$)

$\max_D$

3) On fige D et on entraîne G (on vise $D(G(z)) = 1$)

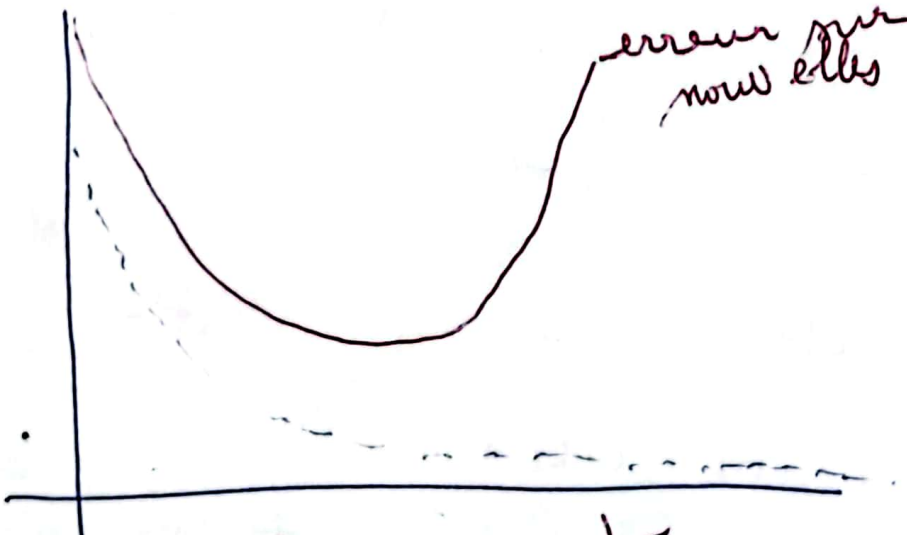
$\min_G$

On itère (1) + (2), et 3.

Au stade de sur-apprentissage

On part d'un échantillon $\{(x_i, y_i)\}_{i=1}^n$ et on veut construire une règle g_n $\mathbb{R}^d \to \mathbb{R}$ minimisant $\mathbb{P}(g_n(X) \neq Y)$

On a vu que selon la complexité de la feuille de règles (et le choix des paramètres) on peut aller du sous-apprentissage (trop simple et on prédit toujours même chose) au sur-apprentissage (temps spécifique aux données d'apprentissage)



Pour le projet coder noyau naif et moyenne mobile.

$$\frac{1}{\#B(x,h)} \sum_{x_i \in B(x,h)} \mathbb{1}_{\{y_i = y\}}$$

Mardi 2/04/24 pas cours

Apprentissage flou

Réf : * Michel Grabisch, 2020
* Louis Lacoste, "Eléments de logique floue"
* Fuzzy models for pattern recognition

Intro

La théorie des ensembles flous développée en 1965 par Lotfi Zadeh. C'est une théorie de l'algèbre abstraite qui a pour but de représenter l'imprécision relative à l'appartenance à certaines classe d'objets. Cette théorie sert de fondement à la logique floue.

Très rapidement en apprentissage on s'est aperçu que la notion de classes

(notamment pour la reconnaissance de flous) pouvoit bénéficier de cette théorie.

En effet, une classe est souvent représentée comme un groupe d'individus partageant certaines similitudes. Ces similitudes peuvent ± fortes entre les individus d'une même classe, et un individu peut présenter des similitudes avec des individus d'autres classes.

Ainsi, l'appartenance d'un individus ne peut être réduite à une seule classe mais "distribuée" sur plusieurs classe. Cela rejoint le fonctionnement des ensembles flous.

Cependant, si des ensembles flou peuvent sembler constituer un cadre naturel pour la notion de classes, force est de constater que les algo classiques considèrent implicitement les classes comme nettes.

Les méthodes Bayésiennes donnent bien une proba d'appartenir à une classe, mais cette notion de proba (associé à la notion de fréquence d'apparti d'un évènement lorsqu'on répète une expérience) est fondamentalement différente de la notion d'appartenance à une classe (sur laquelle repose la théorie "flou").

Exemple : L l'ensemble des liquides

$\mathcal{L}$ le sous-ensemble pour les liquides portables

$\overline{\mathcal{L}}$ le sous-ensemble pour les poison.

On est dans le désert, sans eau depuis longtemps et on trouve 2 bouteilles A et B. Il faut choisir laquelle boire.

On a 2 infos :

* Le degré d'appartenance de B à $\mathcal{L}$ est 0,9 (échelle = [0, 1])

* La proba que $A \in \mathcal{L}$ est 0,9.

Si on choisit B, on sait que le liquide est "pratiquement" portable (ne contiendra pas les substances les plus toxiques),

Si on choisit A, on a 9 chance sur 10 d'avoir un liquide parfaitement

préférable, mais l'absence de moyen

II - Ensemble flous - Relation flous

II.1) Ensembles flous (1965)

Dans le cadre de la théorie des ensembles conventionnelle, les éléments appartiennent ou non (de manière binaires) à des ensembles.
Dans le cadre de la théorie flous, l'appartenance à un ensemble n'est pas si tranchée.
Pour représenter ça mathématiquement, en introduisant une fonction "d'appartenance" d'un élément à un ensemble, les ensembles flous sont définis de la

manière suivante.

Def II.1 : Ensemble flou

Un ensemble flou A est un ensemble, caractérisé par une fonction d'appartenance $f_A : E \longrightarrow [0,1]$, où E est un espace quelconque.

Si $f_A(x)$ est proche de 1, x a un fort degré d'appartenance à A

si $f_A(x)$ est proche de 0, x a un faible degré d'appartenance à A.

Si $f_A : E \longrightarrow \{0,1\}$, on retombe sur la théorie conventionnelle.

Def II.2 : Ensemble vide

A (ensemble flou) est vide si $f_A = 0$

Def II. 3 égalité

2 ensemble A et B sont égaux si $f_A = f_B$ ou $\forall x, f_A(x) = f_B(x)$.

Def III: 4 complémentaire

A ensemble flou, $\overline{A}$ son complémentaire défini par $f_{\overline{A}}(x) = 1 - f_A(x)$.

Def II-5 : inclusion

A, B flous : A est inclu dans B ($A \subset B$) si $f_A(x) \leq f_B(x)$ $\forall x$

Théo classique

si $x \in A$ alors $x \in B$

$x \in A \Rightarrow f_A(x) = 1 \Rightarrow f_B(x) \neq f_A(x) \Rightarrow \underbrace{f_B(x) = 1}_{\Rightarrow x \in B}$.

Def II. 6 Union

A, B flous $\quad C = A \cup B$ est un ensemble flou. Donc $f_C(x) = \max(f_A(x), f_B(x))$ $\forall x \in E$.

Thé. classique

- Si $x \in A, x \notin B$, $f_A(x) = 1, f_B(x) = 0, f_C(x) = 1$
- Si $x \in B, x \notin A$, $f_A(x) = -, f_B(x) = 1, f_C(x) = 1$
- Si $x \in A, x \in B$, $f_A(x) = 1, f_B(x) = 1, f_C(x) = 1$
- Si $x \notin A, x \notin B$, $f_A(x) = 0, f_B(x) = 0, f_C(x) = 0$

Proposition II. 1

$A \cup (B \cup C) = (A \cup B) \cup C$

$A \subset A \cup B$

$B \subset A \cup B$

- $f_{A\cup(B\cup C)}(x) = \max\{f_A(x), f_{B\cup C}(x)\}$

$= \max\{f_A(x), \max\{f_B(x), f_C(x)\}\}$

$= \max\{f_A(x), f_B(x), f_C(x)\}$

$= \max\{\max\{f_A(x), f_B(x)\}, f_C(x)\}$

$= f_{(A\cup B)\cup C}(x)$.

- $f_{A\cup B}(x) = \max\{f_A(x), f_B(x)\} \geq f_A(x)$

$\Rightarrow A \subset A\cup B$.

Def. II.7 intersection

A, B étant deux ens. $C = A\cap B$ définie par

$f_C(x) = \min\{f_A(x), f_B(x)\}$

P II. 2

- $A \cap (B \cap C) = (A \cap B) \cap C$
- $A \cap B \subset A$
- $A \cap B \subset B$.

} Exo.

P. II. 3

- $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$

~~$C \in (A \cup B) = (C \cap A) \cup (C \cap B)$~~

~~$C \in$~~

- $C \cap (A \cup B) = (C \cap A) \cup (C \cap B)$
- $C \cup (A \cap B) = (C \cup A) \cap (C \cup B)$

} Exo

II. 2) Relation floues

$D \subset \mathbb{R}^d$, $k > 1$ (cache conventionnel pour commencer)

"Une partition de D en k

clusters / Classes est donnée par des fonctions indicatrices

$\mu_1, \ldots, \mu_k$

$\mu_i(x) = \begin{cases} 1 & \text{si } x \in \text{classe } i, \ i \in \{1, \ldots, k\} \\ 0 & \text{sinon.} \end{cases}$

Une relation ("dure" / "nette") r dans D est définie comme une fonction

$r : D \times D \longrightarrow \{0, 1\}$

On dit que x et y ($\in D$) partagent une relation si $r(x, y) = 1$.

formalisation matricielle

$D = \{x_1, \ldots, x_n\} \in (\mathbb{R}^d)^n$

$\mu_i(x_j) = \mu_{ij} \quad \begin{cases} i \in \{1, \ldots, k\} \\ j \in \{1, \ldots, n\} \end{cases}$

$R(x_i, x_j) = \mu_{ij}$ $\begin{cases} i \in \{1, \dots, n\} \\ j \in \{1, \dots, n\} \end{cases}$

M_k = ensemble de toutes les matrices de partition U.

$\begin{cases} dim(U) = k \times n \\ U[i,j] = \mu_{i,j} \in \{0, 1\} \\ \sum_{i=1}^{k} \mu_{i,j} = 1 \quad \forall j \quad (\text{chq donnée} \in \text{à une unique classe}). \\ \sum_{j=1}^{n} \mu_{i,j} > 0 \quad \forall i \quad (\text{pas de classe vide}.) \end{cases}$

Pour $U \in M_k$, on associe une matrice de relation $R = \{r_{j,l}\}$ avec $dim(R) = n \times n$ définie par :

$r_{j,l} = \begin{cases} 1 \text{ si } \mu_{i,j} = \mu_{i,l} = 1, \text{ pour un certain } i \\ 0 \text{ sinon.} \end{cases}$

*

Exemple 3 données ($n=3$)
2 classe ($k=2$)

$$U = \begin{pmatrix} x_1 & x_2 & x_3 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

$x_1 \in$ classe 2
$x_2 \in$ ——— 2
$x_3 \in$ ——— 1

μ_{22} : 2ème classe, 2ème ligne ; μ_{23} : 2ème ligne

→ Matrice de relation R associée

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

NB: se poser la qtié si x_1 et x_2 sont ensemble.

* Cache flou

Def VI.8 Matrice flou

$U \in M_{f,k}$ (ensemble des partitions flous à k classe)

si $\dim(U) = k \times n$ (n = nbr de données à classifier)

$U[i,j] = \mu_{ij} \in [0,1]$ (proba = 1 si la donnée j ressemble à la classe i)

$$\sum_{i=1}^{k} \mu_{i,j} = 1$$

$$\sum_{j=1}^{n} \mu_{ij} > 0$$

On peut définir une matrice de relations floue associée à U.

Def II.3 - Matrice de relation floue

$U \in M_{k,n}$, la matrice R_f associée est définie pour

$\begin{cases} 1 \le j \le n \\ 1 \le l \le n \end{cases}$ par

$R_f[j,l] = R_{jl} = \max_{1 \le i \le k} (\min(\mu_{ij}, \mu_{il}))$

Exemple 3 données ($n=3$)
2 colonnes ($k=2$).

$$U = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{pmatrix} x_1 & x_2 & x_3 \\ 0,3 & 0,2 & 0,9 \\ 0,7 & 0,8 & 0,1 \end{pmatrix}$$

x_1 ~~Ensemble~~ plutôt

$\mu_{2,1}$

↳ degré d'appartenance de x_1 à la classe 2.

x_1 ressemble plutôt à la classe 2

x_2 ——— 2

x_3 ——— 1

$* = R_{1,2} = \max\{\min\{\mu_{1,1}, \mu_{1,2}\}, \min(\mu_{2,1}, \mu_{2,2})\}$ = 0,7

$$R_j = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{pmatrix} x_1 & x_2 & x_3 \\ \boxed{0,7} & 0,7 & 0,3 \\ 0,7 & 0,8 & 0,2 \\ 0,3 & 0,2 & 0,9 \end{pmatrix}$$

x_1 et x_2 plutôt ensembles

x_1 et x_3 ——— pas ensemble

x_2 et x_3 plutôt pas ensemble.

III - Classification floue

III-1. Kppv

$S = \{x_1, \dots, x_n\}$ dont on connaît les labels (pas flou)

$\nu_{i,j} \in \{0, 1\}$

$\nu_{j,\ell} \in \{0, 1\}$

$S' = \{x'_1, \dots, x'_m\}$ dont les labels sont inconnus.

$1 \le i \le$ nbr de classe $= c$

On calcul $\nu'_{i,j}$, $1 \le j \le m$

$$\nu'_{i,\ell} = \frac{\sum_{j=1}^{k} \nu_{i,j} \times \frac{1}{\|x'_\ell - x_j\|^{2/(\lambda-1)}}}{\sum_{j=1}^{k} \frac{1}{\|x'_\ell - x_j\|^{2/(\lambda-1)}}}$$

$1 \le i \le c$

λ est un paramètre $(1 < \lambda < \infty)$ qui détermine le degré de flou de la classification

Exemple : 3 données, 2 classe 2ppv, $\lambda=2$

$$U = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

$x_1 = (0,0)$
$x_2 = (1,1)$
$x_3 = (4,4)$

2 nouvelles données $x'_1 = (0,1)$
$x'_2 = (4,1)$

Les 2ppv de x'_1 : x_1 et x_2 $\begin{pmatrix} \|x'_1 - x_1\| = 1 \\ \|x'_2 - x_2\| = 1 \end{pmatrix}$

——— x'_2 : x_2 et x_3 $\begin{pmatrix} \|x'_2 - x_2\| = 3 \\ \|x'_2 - x_3\| = 3 \end{pmatrix}$

$$\mu'_{1,1} = \frac{\mathring{\mu}_{1,1} \times \frac{1}{\|x'_1 - x_{(1)}\|^2} + \mathring{\mu}_{1,2} \times \frac{1}{\|x'_1 - x_{(2)}\|^2}}{\frac{1}{\|x'_1 - x_{(1)}\|^2} + \frac{1}{\|x'_1 - x_{(2)}\|^2}}$$

$$\mu'_{1,1} = 0$$

$$\mu'_{1,2} = \frac{\mathring{\mu}_{1,2} \times \frac{1}{\|x'_2 - x_2\|^2} + \hat{\mu}_{1,3} \times \frac{1}{\|x'_2 - x_3\|^2}}{\frac{1}{\|x'_2 - x_2\|^2} + \frac{1}{\|x'_2 - x_3\|^2}} =$$

$N'_{4,2} = \frac{1}{9} \times$

$$N'_{12} = \frac{1/9}{1/9 + 1/9} = \frac{1}{2} = 0.5$$

$$U' = \begin{pmatrix} 0 & 0.5 \\ 1 & 0.5 \end{pmatrix}$$

Create U

```
label = Data[, Nbcol] # recupere le label.

unique(c(0,0,1,1,1,2,2))
  ↳ (0,1,2)

# recupere le nbr de label
Nblabel = length(uniquelabel)
# Def une matrice contenant que des 0
U = matrix(0, ncol=nbData, nrow = Nblabel)
# remplir la matrice
for (i in 1:Nblabel){
  U[i,] = as.numeric(label == uniquelabel[i])
}
```

Créer fuzzyknn :

```
fuzzyknn = function(x, Data, k, lambda, U = 0){
Nb col = ncol(Data)
nb data = nrow(Data)
Data = Data[order(Data[, Nb col]), ]
donnees = Data[, -nb col]
Label = Data[k, nb col]
Uniquelabel = Unique(Label)
Nblabel = Length(uniquelabel)
if (U==0){ U = create U (Data) }
D = ... # calcul des distances ||X-x||

ODk # indice des kppv.
opp # le vecteur dans lequel on
      va mettre les degrés d'appar
      (evidence de x à chq groupe
for(i in 1:Nblabel ){ # applique la
                        formule }
return (opp)
```

si $D = (1, 0.8, 2, 0.6)$

$ODk = (4, 2, 1)$

ordre ⇒ donne les indice du plus grand (nbr) ou plus petit (nbr)

---

fuzzypred = (_____ d = 4)

hard pred = (_____ d = 1,05)

Afficher les

Comparer avec plusieurs valeurs de λ.

26.03.24

App. STAT

Apprentissage séquentiel

I/ introduction

En stat, l'analyse séquentielle est une analyse où la taille de l'échantillon n'est pas connue à l'avance.

A la place, les données sont évalués au fur et à mesure qu'elle sont recoltés.

— Jusqu'à présent on avait un échantillon $S_n = \{(x_i, y_i)\}_{i=1}^{n}$ et on construisait une règle de décision $\hat{g}_n$ à partir de S_n.

Si on récupère de nouvelles données, il faut tout refaire.

L'approche séquentielle est différente:

On veut pouvoir actualiser $\hat{g}_{n+1}$

à partir de $\hat{g}_n$ et (x_{n+1}, y_{n+1})

& intérêts.

* (1) gain Computationnel : On peut avoir un faible Coût d'actualisati. Il peut être ~~itérativement~~ intéressant de traiter des jeux de données de manière séquentielle même si on les a depuis le début.

* (2) gain d'espace de stockage : une fois $\hat{g}_n$ calculé, on va essayer de se passer de S_n.

II - kppv

II-1) De quoi a-t-on besoin ?

* Retour au cas "classique"

$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, $x \in \mathbb{R}^d$, k fixé, m label

On réordonne l'échantillon $\mathcal{D}_n = \{(X_{(i)}(x), Y_{(i)}(x))\}_{i=1}^n$ et on affecte à x, la classe la plus représenté dans ses kppv.

$$g_n(x) = \underset{j \in \text{Label}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{k} \mathbb{1}_{\{Y_{(i)}(x) = j\}} \right\}$$

* Si une nouvelle donnée (X_{n+1}, Y_{n+1}) arrive, de quoi a-t-on besoin pour actualiser $\hat{g}_n$?

Localement, On a besoin de savoir si X_{n+1} est dans les kppv de x. Pour cela, il faut pouvoir ~~Comparer~~ Comparer $\|X_{n+1} - x\|$ à $\|X_{(1)} - x\|$, $\|X_{(2)} - x\|$, ..., $\|X_{(k)} - x\|$
(d_1) (d_2) (d_k)
à l'étape précédente

⇒ Il faut stocker $d_{1,\dots} d_k$

si $\| X_{n+1} - x \| \geq d_k$, on change rien.

si $\| X_{n+1} - x \| < d_k$, il faut intégrer X_{n+1} dans les kppv, actualiser $\{d_1, \dots, d_k\}$, et il faut recalculer

$$\sum_{i=1}^{q_k} \mathbb{1}_{\{Y_{(i)}(x) = j\}} = n_j .$$

⇒ On va garder $\{n_1, \dots n_m\}$ et

$Y_{(1)}, \dots Y_{(k)}$

II.2) kppv séquentielle

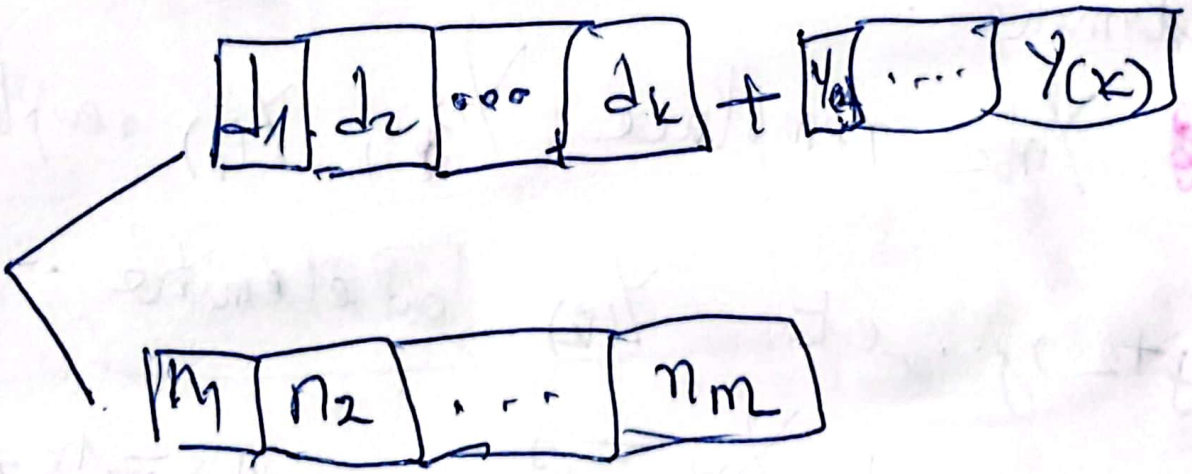
Etape 1 : On a

$S_n = \{(X_i, Y_i)\}_{i=1}^n$, x, k fixés,

n classes ($\{1, \dots m\}$).

On calcule les distances
$d_1 \leq d_2 \leq \dots \leq d_n$ et on garde
$\{d_1, \dots, d_k\}$, q $Y_{(1)}, \dots, Y_{(k)}$

On obtient



On peut calculer
$\hat{g}_n(x) \in \underset{j \in \{1, \dots, m_Y\}}{\arg\max}\ n_j$

Etape 2 : Nouvelle donnée (X_{n+1}, Y_{n+2})
si $\| X_{n+2} - x \| \geq d_k \Rightarrow$ on fait rien.

si $\|x_{n+1} - x\| \in [d_i, d_j[$

$(0 < i < j \leq k, \; d_0 = \infty)$

On remplace d_j par $\|x_{n+1} - x\|$, d_{j+1} par d_j ; d_{j+2} par d_j, d_{j+2} par d_{j+1} ... et d_k est éliminé.

② y_{n+1} remplace $y_{(i)}$, $y_{(j)}$ remplace y_{j+1}, ... et $y_{(k)}$ est éliminé.

③ si $\begin{cases} y_{n+1} = \rho \\ y_{(k)} = t \end{cases} \Rightarrow \begin{cases} n_s = n_s + 1 \\ n_b = n_{t-1} \end{cases}$

g_{n+1} & $\operatorname{argmax}\{n_j\}$

* ~~Avantage~~ : peu de chose à calculer, peu de chose à stocker.

* inconvénient : K est figé
Coole MR

① kppv "classiq" sur un 1er nbr de
donnée (Etape 1)

② appliquer une fct° qui
actualise
↳ f[k, distance, kLabels,
nn, x, NomData,)

EXAMEN 24 [0?] 24

Règles d'association

* Apprentissage non supervisé.

*→ Influence causale

• Si quelqu'un achete du pain
et du beurre est-ce qu'il va
aussi acheter du lait.

Sac	Contenance
1	Beurre, pain, lait,
2	pain, viande
⋮	⋮ ⋮ ⋮
n	Poisson, pain

2 critères : "Support", "confidence"

* support : % d'observation qui contiennent les éléments de la règle.

100 sac, 30 avec au moins lait, pain, beurre.

Support {lait, pain, Beurre} = 0.3 (30%)

* confidence : % d'observation contenant y sachant qu'il y avait X.

26.03.24 App. STAT (3)

100 sac, 33 contiennent Pain et beurre et dans les 33, 30 contiennent du lait.

Confiance : $\frac{30}{33} \times 0.91$

⇒ Règles intéressantes : gain support + gain confiance

Un algo : "A priori" (1997)

Etape 1 : On cherche des sous ensembles de support assez élevé.

Etape 2 : Dans ces sous-ensembles, on cherche les règles de grande confiance.

Exemple : Critère { support > 0.4
confidence > 0.7

1	A, B, C
2	A, B, C
3	B, D, E

Etape 1 :

table 1 → utile pour éliminer les combinaisons de taille supérieure.

A	0.67
B	1
C	0.67
D	0.33
E	0.33

proportion d'apparition.

support

table 2

{A, B}	0.67
{B, C}	0.67
{A, C}	0.67

Taille 2

{A,B,C}	0.67

Etape 2

taille 2

confidence ? 0.7

A → B	$\frac{2}{2} = 1$
B → A	$2/3 = 0.67$
A → C	1
C → A	1
B → C	0.67
C → B	1

2 qui contient A1 et pour C1 A 2, 2 contient B

2 fois