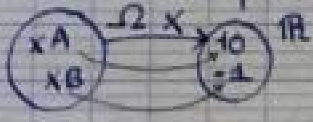
TD: ML

TD1 Exercice 1

1) 1) événement élémentaire = résultat de l'expérience
événement = pls événements élémentaires
univers = ens des événements élémentaires

2) $P : \mathcal{P}(\Omega)$ ens des parties de $\Omega \to [0,1]$
- $P(\Omega) = 1$ et $P(\emptyset) = 0$
- $P(x) \geq 0$
- $A \cap B = \emptyset$ alors $P(A \cup B) = P(A) + P(B)$

3) variable aléatoire : permet d'attribuer un résultat à un événement élémentaire



$X : \Omega \to \mathbb{R}$ → passerelle entre év et valeur
$X(\omega) = 10$
$P(X=10) = P(X^{-1}(10))$
$P(X \in [a,b]) = P(X^{-1}([a,b]))$

2) loi géométrique : $p^k(1-p)$

3) Les événements sont indépendants car $P(\text{Roi}) = 4/52$ et $P(\text{Pique}) = 13/52$ et $P\left(\begin{smallmatrix}\text{Roi de}\\ \text{Pique}\end{smallmatrix}\right) = 1/52$
Ils ne sont pas incompatibles car on peut tirer le roi de pique.
$P(\text{Roi} \cup \text{Pique}) = 4/52 + 13/52 - 1/52 = 16/52$

Exercice 2

1) L'univers est continu (les événements sont infini et non dénombrables)
Comment évolue la proba autour d'un point de Ω sachant que cette proba $= 0$?
⇒ densité de probabilité $= \int_a^b p(x)dx = P_X([a,b])$
densité jointe comme dans le cas discret
densité marginale : $P(X) = \int_{-\infty}^{+\infty} p(X,Y)\,dy$

2) $\mathbb{E}(X) = \int_{\mathbb{R}} x\,p(x)\,dx = \int_{\Omega} X(\omega)p(\omega)\,d\omega$

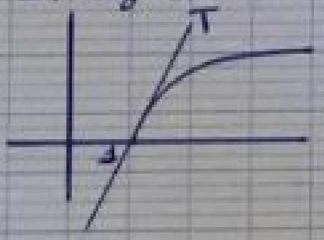
3) $\mathbb{E}(X|Y=y) = \int_{\mathbb{R}} x\,p(x|Y=y)\,dx$

4) $\mathbb{E}(X) \underset{|E| \to +\infty}{\leftarrow} \frac{1}{|E|} \sum_i x_i$ iid : tirages ind les uns des autres et choisis dans le même ensemble
→ loi des grands nombres

Exercice 3

2) 1) $H(p) \geq 0$ car $p_i \in]0;1] \Rightarrow \log p_i \leq 0 \Rightarrow -\log p_i \geq 0 \Rightarrow -\sum p_i \log p_i \geq 0 = H(p)$

2) $\log(x) \leq x - 1$



$T(x) \geq \log(x)$
$a\overset{1}{x} + b$
$\log(1)' = 1$

on a $T(1) = \overset{1}{x} + b = 0 \Leftrightarrow b = -1$
donc $T(x) = x - 1$

$-\sum p_i \log(p_i) \leq -\sum p_i \log(q_i)$
$\Leftrightarrow \sum p_i \log(p_i) - \sum p_i \log(q_i) = \sum p_i \log\left(\frac{p_i}{q_i}\right) \geq 0$
$\Leftrightarrow \sum p_i \log\left(\frac{q_i}{p_i}\right) \leq 0$
$\sum p_i \log\left(\frac{q_i}{p_i}\right) \leq \sum p_i \left(\frac{q_i}{p_i} - 1\right)$ car $\log(x) \leq x - 1$ ici $x = \frac{q_i}{p_i}$
$\leq \underbrace{\sum q_i}_{1} - \underbrace{\sum p_i}_{1} = 0$

$H(p) \leq -\sum p_i \log q_i$

3) avec $q_i = 1/m$, $-\sum p_i \log(1/m) = \sum p_i \log(m) = \log(m)$

3) 1) égalité montrable par récurrence en faisant un changement de variable

2) $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$
$\overset{\shortparallel}{\sum_{i=1}^{N}} x_i p_i \ldots$

3) $KL(p_i \| q_i) = \sum p_i \log \frac{p_i}{q_i} = -\sum p_i \log \frac{q_i}{p_i} \geq -\log\left(\sum p_i \frac{q_i}{p_i}\right)$
$\geq -\log(\underbrace{\sum q_i}_{1}) = 0$

6 n'est pas symétrique $\Rightarrow$ on peut le transformer en distance en prenant la moyenne entre $p \| q$ et $q \| p$

4) $\overset{\text{estimation ponctuelle}}{\hat{p}(x)} = 1/N \sum_{i=1}^{N} \delta(x - x_i)$ on essaie de l'approcher par p_θ

$KL(\hat{p} \| p_\theta) = \sum_{x \in X} \hat{p}(x) \log \frac{\hat{p}(x)}{p_\theta(x)} = \underbrace{\sum_x \hat{p} \log \hat{p}(x)}_{-H(\hat{p})} - \sum_x \hat{p} \log p_\theta(x)$

$-\sum \hat{p} \log p_\theta(x) = -1/N \sum_x \sum_i \delta(x - x_i) \log p_\theta(x)$
$= -1/N \sum \log p_\theta(x)$

1) 1) 3^{16} valeurs différentes sont possibles

2) $\frac{P(D|V=v)}{P(R|V=v)}$ on ne peut pas calculer ça car on a pas tous les votes étiquetés (on a juste les votes agrégés par partie et pas les vecteurs individuels pour chaque sénateur)

$\Rightarrow$ On utilise Bayes

$$\frac{P(V=v|D)\,P(D)\,P(V=v)}{P(V=v|R)\,P(R)\,P(V=v)} = \frac{P(V=v|D)\,P(D)}{P(V=v|R)\,P(R)} \qquad P(D) = \frac{|D|}{|D|+|R|} \qquad P(R) = \frac{|R|}{|D|+|R|}$$

On ne peut pas estimer $P(V=v|D)$ car on a pas de vecteurs individuels. On est obligé de faire l'hypothèse d'indépendance.

Rmq : ici l'hypothèse d'ind n'est pas vraiment juste car les votes dépendent les uns des autres. Mais on a pas le choix de faire autrement

$$P(V=v|D) = \prod_{i=1}^{16} P(v_i|D) \qquad \text{On passe au log}$$

$$\log \frac{P(D|V)}{P(R|V)} = \sum \log P(v_i|D) - \sum \log P(v_i|R) + \log\left(\frac{|D|}{|R|}\right)$$

$$P(v_i = \text{OUI}|D) = 1 - P(v_i = \text{NON}|D) - P(v_i = \text{NSP}|D)$$

2) 1) $G \sim \text{Ber}\left(\frac{|h|}{|h|+|f|}\right)$

2) $p(t) = \underbrace{p(t|G=h)}_{\mathcal{N}(\mu_h,\,\sigma_h)}\,p(G=h) + \underbrace{p(t|G=f)}_{\mathcal{N}(\mu_f,\,\sigma_f)}\,p(G=f)$ $\Rightarrow$ mixture gaussienne

$$P(G=h|t) = \frac{p(t|G=h)\,p(G=h)}{p(t)}$$

3) classifieur bayésien $= \frac{p(G=h|t)}{P(G=f|t)} = \frac{P(t|G=h)\,p(h)}{P(t|G=f)\,p(f)}$

3) 5 paramètres : 2 par gaussienne (esp + variance)
1 param pour le rapport h/f

Exercice 4

1) 1) $p(x_1, \dots, x_m) = \prod_{i=1}^{m} P(x_i) = \prod_{i=1}^{6} p_i^{m_i}$
car indépendance

2) $m_k \to m p_k$ (loi des grands nb) donc $\prod_{i=1}^{6} p_i^{m_i} \to \prod_{i=1}^{6} p_i^{m p_i}$

$$= \prod_{i=1}^{6} 2^{m p_i \log p_i} = 2^{m \sum_{i=1}^{6} p_i \log p_i} = 2^{-m H(p)}$$

faible entropie = le nb de séquences est très diversifié donc bcp d'aléatoire
forte entropie = on tire presque tjrs la m̂ séquence

TD2 **Exercice 1**

$p(y)$: a priori
$p(x)$: évidence
$p(x|y)$: vraisemblance / a priori
$p(y|x)$: a posteriori

1) On se positionne dans un monde idéal.
X dans $\mathbb{R}^d$ et $Y = \{y_1, \dots, y_k\}$ $X = \{x_1, \dots, x_n\}$ $x_i \in \mathbb{R}^d$
Classifieur $X \to Y$ qui utilise des algo fondés sur Bayes
$\hat{y} = \arg\max_y p(y|x)$

2) Erreur : Proba de se tromper $= P(\hat{y} \neq y | x) = 1 - p(\hat{y} = y|x) = 1 - \max p(y|x)$ (car argmax)
utilité d'entraînement — spécialisation du 0 erreur

On ne va pas trouver mieux. Pour le prouver on utilise un raisonnement par l'absurde et on suppose qu'il existe un classifieur $\hat{y}_0$ tq $E_{Bay} > E_0$
$\Rightarrow$ contradiction.
$1 - \max p(y|x)$ $\quad 1 - p(y = \hat{y}_0 | x)$

3) $\lambda(y_j, y_i)$ coût de prédiction de y_j plutôt que y_i

__Dans le cas 0-1 :__ $\lambda(y_i, y_j) = \underbrace{\mathbb{1}}_{\text{indicatrice}}(y_i \neq y_j) = \delta(y_i \neq y_j) = \begin{cases} 1 & \text{si } y_i \neq y_j \\ 0 & \text{sinon} \end{cases}$

__Dans le cas asymétrique :__ → cas des classes "proches" (non binaire)
ex : chien plus proche de chat que camion $\lambda(\text{chien}, \text{chat}) < \lambda(\text{chien}, \text{camion})$
→ cas binaire
ex : tests médicaux / fraudes $\lambda(+,-) > \lambda(-,+)$

4) $R(y_i|x)$ en fonction de λ et $P(y|x)$
$R(y_i|x) = \underbrace{\mathbb{E}_{Y \sim P(\cdot|x)}[\lambda(y_i, Y)]}_{\text{espérance du coût}} = \sum_{y \in Y} p(y|x)\lambda(y_i, y)$

5) $D = \{(x_i, y_i)\}_{i=1}^{m}$ ens des données $\quad x_i \in X$, $y_i \in Y$
$R(f) = \int_{x, y \in X, Y} \lambda(\hat{y}, y)\, p(x, y)\, dx\, dy = \mathbb{E}_{XY \sim D}[\lambda(f(X), Y)]$
risque théorique dans le cas où on suppose qu'on connaît la distribution D — $f(x)$ — distribution de proba des données

Dans la pratique on suppose :
→ D est un ensemble iid suivant D la distribution
→ loi des grands nombres $Z = \lambda(f(X), Y)$
$E(Z) \approx 1/m \sum z_i = 1/m \sum \lambda(f(x), y) = R_D(f)$ risque empirique

6) $\lambda_{+-} \overset{\text{par définition}}{\triangleq} \lambda(x, y)$
Quel critère de décision faut-il prendre ?
→ on veut minimiser l'erreur
$E(\hat{y}) = \lambda(\hat{y}, +)P(y = +|x) + \lambda(\hat{y}, -)p(y = -|x)$
Décision : on prédit plus si $E(+) < E(-)$
$\lambda_{++}\, p(y = +|x) + \lambda_{+-}\, p(y = -|x) < \lambda_{-+}\, p(y = +|x) + \lambda_{--}\, p(y = -|x)$
__But__ : arriver à un critère du type $p(y = +|x) >$ seuil ← dans le cas 0-1
(vérifier que $\lambda^{(0-1)} \Rightarrow p(y = +|x) > 0{,}5$)
→ on va supposer que $\lambda_{+-} > \lambda_{--}$ et $\lambda_{-+} > \lambda_{++}$

5) x_i' est le plus proche voisin de x parmi $\{x_i\}_{i=1}^{m}$
Mq $(x_i')_{i\geq 1}$ converge vers x
Mq $\forall B(x,\varepsilon)$ $P(x_i' \in B) \underset{i \to +\infty}{\longrightarrow} 1$
Hyp : $p(x) > 0$ (densité)
$\rightarrow \int_B p(x)\,dx = P_x \forall B > 0$ et $P(X_{m'} \in B) > 0$ pour m'
$\rightarrow P(x_i' \in B) > 0$ pour $i' \geq m'$

6) $q_k(x) = P(y = k|x)$
$r(x, x'_m) = P(y \neq y'_m | x, x'_m)$
$= 1 - P(y = y'_m | x, x'_m)$
$= 1 - \sum_k \underbrace{P(y = k | x, x'_m)}_{q_k(x)\, q_k(x'_m)}$

7) $r(x, x'_m) \underset{m \to +\infty}{\longrightarrow} 1 - q_k(x)^2$ car $x'_m \underset{proba}{\longrightarrow} x$ $\quad q_k(x'_m) \underset{proba}{\longrightarrow} q_k(x)$

8) $r(x) = 1 - q_k(x)^2$

9) Mq $r(x) \leq r_f(x)\left(2 - \frac{K}{K-1} r_f(x)\right)$
En utilisant $(\sum u_i v_i)^2 \leq (\sum u_i^2)(\sum v_i^2)$
$r_f(x) = 1 - \max_y p(y|x) = 1 - p(z|x)$ pour $z = \arg\max_y p(y|x)$
avec v_i : $(\sum u_i)^2 \leq (\sum u_i^2)(\sum 1)$
$(\sum q_i)^2 \leq (\sum q_i^2) \times (K-1) \Leftrightarrow (1 - r_f)^2 \leq \dots$

TD 3 Exercice 1

1) f convexe $\Leftrightarrow f'' \geq 0 \Leftrightarrow f'$ est croissante
- $f(x) = x \cos(x)$ la fonction traverse + de 2 fois l'axe des abscisses $\Rightarrow$ pas convexe
- $g(x) = -\log(x) + x^2$ $\quad g''(x) = 1/x^2 + 2 \geq 0$ (2 fonctions convexes) $\Rightarrow$ convexe
- $h(x) = x\sqrt{x}$ $\quad h'(x) = 3/2\sqrt{x}$ $\quad h''(x) = 3/4\, x^{-1/2} \geq 0 \Rightarrow$ convexe
- $t(x) = -\log(x) - \log(10 - x) \rightarrow -1/x + \frac{1}{10-x} \rightarrow \frac{1}{x^2} + \frac{1}{(10-x)^2} > 0 \Rightarrow$ convexe

2) $\nabla_x f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix}$ $\quad$ ici $f(x) = 2x_1 + x_2^2 + x_2 x_3$
$\nabla f = \begin{pmatrix} 2 \\ 2x_2 + x_3 \\ x_2 \end{pmatrix}$ $\quad \nabla_x f$ est de même dimension que x

3) $\nabla_x (f(x) + g(x)) = \begin{pmatrix} \frac{\partial (f+g)}{\partial x_1} \\ \vdots \\ \frac{\partial (f+g)}{\partial x_d} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x_1} + \frac{\partial g}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} + \frac{\partial g}{\partial x_d} \end{pmatrix} = \nabla f + \nabla g$

- $\nabla_x t f(x) = \begin{pmatrix} \frac{\partial t f}{\partial x_1} \\ \vdots \\ \frac{\partial t f}{\partial x_d} \end{pmatrix} = \begin{pmatrix} \frac{t \partial f}{t \partial x_1} \\ \vdots \\ \frac{t \partial f}{t \partial x_d} \end{pmatrix} = t \nabla f$

Exercice 2

On veut calculer la densité

1) estimation de la densité p_B de X pour B de volume V

On a $D = \{x_i\}_{i=1}^{n}$ $P(X \in B)$

$P(X \in B) = \int_{x \in B} p(x)\,dx$ + à savoir

hyp : si B est suffisament petit, $p(x) \simeq p_B$ (passage à la limite) (constante) $P(X \in B) \approx \int p_B\,dx = p_B V$

$X \in B$ loi de Bernouilli succès ou échec

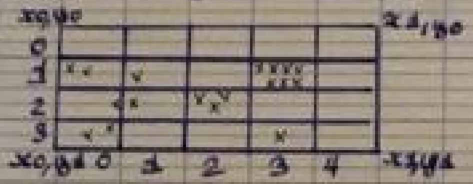
Si X_i iid et de même loi que X, $\sum X_i \sim \text{Binomiale}(p(X \in B), m)$

$\mathbb{E}(\sum_{i=1} X_i) = m\,p(X \in B) \approx \underbrace{\sum_{i=1} \mathbb{1}(x_i \in B)}_{k}$

$\rightarrow P(X \in B) \simeq k/m$

$\rightarrow p_B V = k/m \quad p_B \simeq k/mV$

2) $X \in \mathcal{X}$ On a $X_0 = \{x_i\}_{i=1}^{m}$ $p(X)$?



résolution Δ

en x : $m_x = \left[\frac{x_1 - x_0}{\Delta}\right]^{+}$ nb d'intervalles

$m_y = \left[\frac{y_1 - y_0}{\Delta}\right]^{+}$

$\Delta x = \frac{x_1 - x_0}{m_x} \qquad \Delta y = \frac{y_1 - y_0}{m_y}$

On suppose que la densité est constante dans les hypercubes = rectangles

Pour chaque rectangle π_{ij} on estime $P_{ij} = \frac{k_{ij}}{mV} = \frac{k_{ij}}{m \Delta x \Delta y}$

3) Pb de la méthode des histo : dépend de la manière dont on découpe les données

$\phi(x) \qquad p(x) \approx 1/m \sum \phi(x - x_i)$

ϕ
- → symétrique la direction n'a pas d'importance
- → doublement monotone avec un max en 0 plus un point est proche, plus il est important
- → $\int_x \phi(x)\,dx = 1$ densité de probabilité

On utilise souvent un facteur d'échelle $\phi'(x) = \phi\left(\frac{x}{\sigma}\right)$ pour avoir une résolution plus fine ou plus grossière

$\sigma \in \mathbb{R}_+^{p}$

le voisinage est plus grand et la résolution est plus fine — le voisinage est plus petit et la résolution est plus grossière

cas 1D : $\int_x \phi'(x)\,dx = \int_x \phi\left(\frac{x}{\sigma}\right)dx$

→ changement de variable $y : x \to \sigma x$

$= \int \phi(y) \left|\det \frac{dy}{dx}\right| = \sigma d$

$\Leftrightarrow \phi_\sigma(x) = \frac{1}{\sigma d}\,\phi\left(\frac{x}{\sigma}\right)$

Exercice 3

1) Parzen :
- → prendre un k suffisament grand
- → classe choisie = classe majoritaire de l'hypercube centré en x
- → prend en compte la localité des points

K-nn :
- → classe choisie = classe majoritaire parmi les k plus proches voisins
- → avec un k majoritaire, prédit la classe majoritaire

2) et 3) voir feuille (en bleu)

4) voir feuille (en gris)

$k \to \infty \Rightarrow$ classe majoritaire

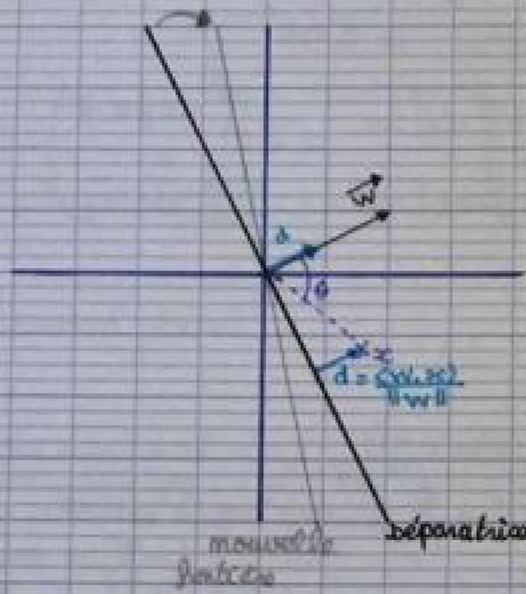
TD4 Exercice 1

$f_w(x) = \sum_{i=1}^{d} x_i w_i + b = \langle x, w \rangle + b$

1) $MSE(f_w(x), y) = (f_w(x) - y)^2$ si $y \in \mathbb{R}^d$: $\|f_w(x) - y\|^2$

2) $L(f_w(x), y) = \max(0, -y\, f_w(x))$
- $xy > 0$ et $f_w(x) > 0$ ou $y < 0$ et $f_w(x) < 0$ $\}$ $-y\, f_w(x) < 0 \Rightarrow L = 0$
- sinon $L > 0$

4)



$\langle w, x \rangle = \|w\| \|x\| \cos(\theta)$

$w' = w + y x$

$\langle w', x \rangle = \langle w + y x, x \rangle = \langle w, x \rangle + y \|x\|^2$

5) w^1, w^2 sont les mêmes que $W = (2, 1)$ car ils sont colinéaires
w^3 est le symétrique de $w \Rightarrow$ inversement des classes

6) $\nabla_w L = \begin{cases} -y\, f_w(x) < 0 \to L = 0 \to 0 \\ -y(\langle w . x \rangle + b) > 0 \to -y x \end{cases}$

$w' = w - \varepsilon \nabla_w L = w + \begin{cases} \varepsilon y x \leftarrow L > 0 \\ 0 \leftarrow L = 0 \end{cases}$

$\max(0, x - y\, f_w(x))$

$y\, f_w(x) = 0,5 \quad -y\, f_w(x) = -0,5 \to L = 0 \quad L = -0,5$

$y\, f_w(x) > \alpha \Rightarrow L = 0$

7) Si $w = 0$, alors il resant toujours à 0 donc on pose $\max(0, \alpha - y\, f_w(x))$

$y \cdot f_w(x) = 0.5$

$-y \cdot f_w(x) = 0.5 \to L = 0$

8) Batch : $\nabla L(f(X), X) = 1/N \sum_{i=1}^{N} \nabla L(f(x^i), y^i)$

stochastique : $\to$ on mélange ts les points de X
for (x_i, y_i) :
$w \leftarrow w - \varepsilon \nabla L(f(x_i), y_i)$

mini-batch : $(X_1, \dots, X_D)$
for X_i :
$w \leftarrow w - \varepsilon \nabla L(f(x_i), y_i)$

- $\beta_i \xi_i = 0$
- $1 - \xi_i - f_w(x_i) y_i \leq 0$
- $\alpha_i (1 - f_w(x_i) y_i - \xi_i) = 0$
- $\beta_i \xi_i = 0$
- $f_w(x_i) y_i < 1 \Rightarrow \xi_i > 0$
- $\beta_i = 0 \Rightarrow \alpha_i = K$

$\Rightarrow f_w(x_i) y_i > 1 \Rightarrow \xi_i = 0 \Rightarrow \alpha_i(1 - f_w(x_i) y_i) = 0 \Rightarrow \alpha_i = 0$

$\alpha_i = K \Rightarrow \beta_i = 0 \Rightarrow \xi_i \geq 0 \rightarrow$ point dans la marge ou mal classé

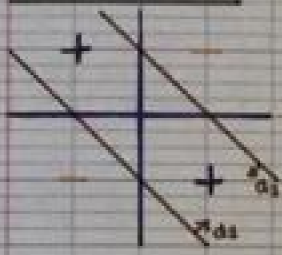
$\alpha_i = 0 \Rightarrow \beta_i = K \Rightarrow \xi_i = 0 \rightarrow$ point bien classé

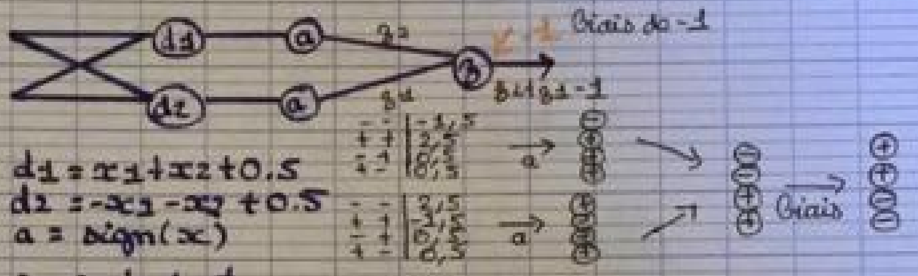
$0 < \alpha_i < K \Rightarrow 0 < \beta_i < K \Rightarrow \xi_i = 0 \Rightarrow f_w(x_i) y_i = 1 \rightarrow$ point sur la marge

$f_w(x) = \langle w . x \rangle + b = \langle \sum_i \alpha_i y_i x_i , x \rangle + b = \sum_i \alpha_i y_i \langle x , x_i \rangle + b$

10) $\langle w_x , x_i \rangle + b = y_i$

TD6 Exercice 1



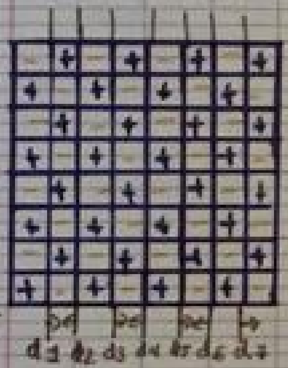


$d_1 = x_1 + x_2 + 0.5$

$d_2 = -x_1 - x_2 + 0.5$

$a = \text{sign}(x)$

$g = g_2 + g_1 - 1$



$> 0 \Rightarrow$ pair de droites
$< 0 \Rightarrow$ impair de droites } pour les lignes puis les colonnes

Exercice 2

3) $\frac{1}{N} \sum_{i=1}^{N} (f_w(x^i) - y^i)^2 \xrightarrow[m \to +\infty]{} R^*(f_w) = \iint (f_w(x) - y)^2 p(x,y)\, dx\, dy$
$= \mathbb{E}[(f_w(x) - y)^2]$

4) $R(f_w)(x) = \int_y (f_w(x) - y)^2 p(y|x)\, dy$

$\frac{\partial R}{\partial f_w} = 2 \int_y (f_w(x) - y)\, p(y|x)\, dy = 0$

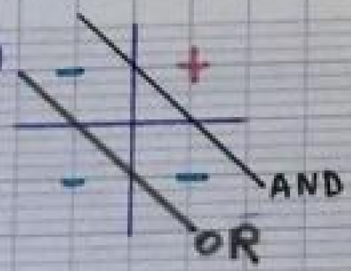
$\Leftrightarrow f_w^*(x) \int p(y|x)\, dy - y \int p(y|x)\, dy = 0$

$\Leftrightarrow f_w^*(x) \int p(y|x)\, dy = \int_y y\, p(y|x)\, dy$

$\Leftrightarrow f_w^*(x) = \mathbb{E}[y|x]$

ça on dit et
[illegible]
$\sum_y p(y|x) = 1$

5) $\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \begin{matrix} f_1(x) \\ f_2(x) \\ \vdots \\ \\ f_k(x) \end{matrix}$ $\quad f_k(x) = \sum_{i \neq k} 0 . p(y_i|x) + 1 . p(y_k|x)$

3) 

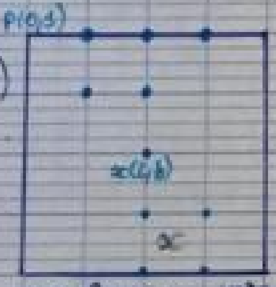
$w = (-1, 1, 1)$

Exercice 3

1) $x \in \mathbb{R}^2 \to (x_1, x_2)$
$f_w(x) = \langle x . w \rangle$ $\quad w \in \mathbb{R}^d$

2) $\phi(x) = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2) \in \mathbb{R}^6$
$w \to \mathbb{R}^6$
$f_w(\phi(x)) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2$
si $w_4 w_5 > 0 \to$ ellipse
si $w_4 w_5 < 0 \to$ hyperbole
sinon $\to$ parabole

3) Oui c'est plutôt intéressant (on a plus d'expressivité mais on garde des frontières "propres")
↳ pas le même coût (on ne peut pas le comparer entre familles de fonc° diff.)
$\Rightarrow$ ça nous intéresse le taux de bonne classification

4) 

On peut les retrouver en annulant w_3, w_4 et w_5

$$\phi(x) = \begin{pmatrix} e^{-\frac{\|x - p_{1,1}\|^2}{\sigma^2}} \\ \vdots \\ e^{-\frac{\|x - p_{m,m}\|^2}{\sigma^2}} \end{pmatrix} = \begin{pmatrix} S(x, p_i, j) \\ \vdots \end{pmatrix}$$

$$f_w(\phi(x)) = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{i,j} \, \phi_{ij}(x)$$
$$= \sum \sum w_{i,j} \, e^{-\frac{\|x - p_{ij}\|^2}{\sigma^2}}$$

$\Rightarrow$ on classe x uniquement en fonction de son voisinage.

TD5 Exercice 1

1) frontière de décision : $f_w(x) = \langle x . w \rangle + b \in \mathbb{R}$
$f_{\alpha w}(x) = \langle \alpha x . w \rangle + \alpha b = \alpha(\langle x . w \rangle + b)$

1) $x_A = x_B + \gamma y^i \frac{w}{\|w\|}$ (projeté) ↳ translation du point
$x_B = x_A - \gamma y^i \frac{w}{\|w\|}$

$\langle x_B, w \rangle = 0$ car B est sur la frontière de décision
$\Rightarrow \langle x_A - \gamma y^i \frac{w}{\|w\|}, w \rangle + b = 0$
$\Rightarrow \langle x_A, w \rangle + b - \gamma y^i \frac{\langle w . w \rangle}{\|w\|} = 0$
$\Rightarrow f_w(x_A) = \gamma y^i \frac{\|w\|^2}{\|w\|}$
$\Rightarrow \gamma = \frac{y^i f_w(x_A)}{\|w\|}$

7) $R(f_w(x)) = \int_y (f_w(x)-y)^2 \, p(y|x)\, dy$

$= E_{y|x}((f_w(x)-y)^2)$

$= E_{y|x}((f_w(x) - f^*(x) + f^*(x) - y)^2)$

$= E_{y|x}((f_w(x) - f^*(x))^2 + (f^*(x)-y)^2 + \underbrace{2(f_w(x)-f^*(x))(f^*(x)-y)}_{①})$

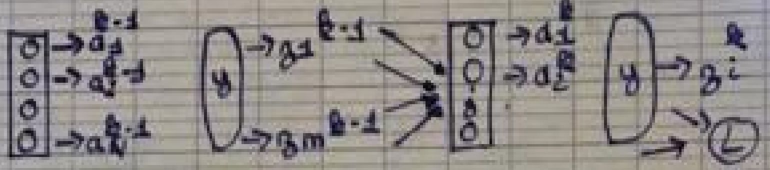
① $E_{y|x}((f_w(x)-f^*(x))(f^*(x)-y)) = (f_w(x)-f^*(x))\, E_{y|x}(f^*(x)-y)$

$= (f_w(x)-f^*(x))\,\underbrace{(f^*(x) - E_{y|x}(y))}_{0}$

$= 0$

$$R(f_w(x)) = E_{y|x}(\underbrace{(f_w(x)-f^*(x))^2}_{\text{variance}} + \underbrace{(f^*(x)-y)^2}_{\text{biais}})$$

Exercice 3



coût MSE
$= \sum (y_k - f(Wx))^2$
$= (g(a_j^k) - y_j)^2$

$z_i^k = g^k\left(\sum_{j=1}^{N} w_{ji}^k \, z_j^{k-1}\right)$

$\frac{\partial L}{\partial w_{ij}^k} = \frac{\partial L}{\partial a_j^k}\frac{\partial a_j^k}{\partial w_{ij}^k}$

$\frac{\partial L}{\partial a_j^k} = \frac{\partial (g(a_j^k) - y_j)^2}{\partial a_j^k} = 2(g(a_j^k) - y_j)\, g'(a_j)$

$\frac{\partial a_j^k}{\partial w_{ij}^k} = z_i^{k-1}$

$\delta_i^k = \frac{\partial L}{\partial a_i^k} = \sum_j \frac{\partial L}{\partial a_j^{k+1}} \frac{\partial a_j^{k+1}}{\partial a_i^k}$

$= \sum \delta_j^{k+1} \frac{\partial w_{ij}^{k+1} g(a_i^k)}{\partial z^k}$

$= \sum_j \delta_j^{k+1} w_{ij}^{k+1} g'(a_i^k)$

$\frac{\partial L}{\partial w_{ij}^k} = \frac{\partial L}{\partial a_j^k}\frac{\partial a_j^k}{\partial w_{ij}^k} = \delta_i^k z_i^{k-1} = \left(\sum_h \delta_h^{k+1} w_{jh}^{k+1} g'(a_h^k)\right)(z_i^{k-1})$

TD7 Exercice 1

$K_1(x,y) = \langle \phi_1(x), \phi_1(y)\rangle$

$K_2(x,y) = \langle \phi_2(x), \phi_2(y)\rangle$

$X \xrightarrow{\phi} E \qquad \text{Kernel}(x,y) = \langle \phi(x), \phi(y)\rangle$

a) $cK_1(x,y) \overset{?}{=} \langle \phi_3(x), \phi_3(y)\rangle$ on pose $\phi_3(x) = \sqrt{c}\,\phi_1(x)$

$\langle \phi_3(x), \phi_3(y)\rangle = \langle \sqrt{c}\,\phi_1(x), \sqrt{c}\,\phi_1(y)\rangle$

$= c\langle \phi_1(x), \phi_1(y)\rangle$

$= c\,K_1(x,y)$

b) $K_4(x,y) \overset{?}{=} K_1(x,y) + K_2(x,y) = \langle \phi_1(x), \phi_1(y)\rangle + \langle \phi_2(x), \phi_2(y)\rangle$

on pose $\phi_4(x) = \begin{pmatrix}\phi_1(x)\\ \phi_2(x)\end{pmatrix}$ $\quad \langle \phi_4(x), \phi_4(y)\rangle = \left\langle \begin{pmatrix}\phi_1(x)\\ \phi_2(x)\end{pmatrix}, \begin{pmatrix}\phi_1(y)\\ \phi_2(y)\end{pmatrix}\right\rangle$

$= \langle \phi_1(x), \phi_1(y)\rangle + \langle \phi_2(x), \phi_2(y)\rangle$

2) $\gamma^* = \min \frac{|y_i f_w(x_i)|}{\|w\|}$
Si on souhaite maximiser, il faut minimiser $\frac{|y_i f_w(x_i)|}{\|w\|}$

Il faut que tous les points soient bien classés.

2) 1) ≥ 1 car le point le proche est 1 et il faut que ts les points soient $\geq 1 \Rightarrow \exists i$ tq $f_w(x_i)y_i = 1$
On veut que ts les points soient ≥ 1
Ça ne doit pas être 0 car sinon on ne considère que la frontière et pas la marge $\Rightarrow$ on retourne dans le cas du perceptron

2) Quand on a un pb avec $\operatorname{argmin} C(w)$ (fonction de coût) et des contraintes tq $g_1(w)=0$, $g_2(w)=0 \cdots g_n(w)=0$ et $c_1(w)\leq 0 \cdots c_N(w) \leq 0$.

Au point solution : les gradiants sont colinéaires. (sinon on pourrait continuer à descendre dans la descente de gradient et trouver une autre solution)

$\Rightarrow$ On a le Langrangien $L(w, \beta_1 - \beta_N, \alpha_1 - \alpha_N) = C(w)$ (fonction de coût) $+ \beta_1 g_1(w) + \cdots + \beta_N g_N(w) + \alpha_1 c_1(w) + \cdots + \alpha_N c_N(w)$
$= \min_w \max_{\alpha_i, \beta_i} L(w,\beta,\alpha) \Leftrightarrow \max_{\alpha,\beta} \min_w L(w,\beta,\alpha)$

3) Les contraintes sont très difficiles à respecter. (les contraintes d'égalités) demandent bcp moins de corrections à faire).
$\Rightarrow$ si les contraintes ne sont pas respectées, la quantité max sera non bornée.

On pose le lagrangien du SVM. On ajoute m contraintes (car on a m exemples)
$L(w,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1} \alpha_i (1 - (\langle x_i . w\rangle + b) y^i)$
$\Rightarrow$ On dérive par rapport à w
$\nabla w_L = W - \sum_{i=1}^N \alpha_i y_i x_i = 0 \quad \Rightarrow w^* = \sum_{i=1}^N \alpha_i y_i x_i$
$\frac{\partial L}{\partial b} = \sum \alpha_i y_i = 0$

On va remplacer w et b
$L(w,\alpha) = \frac{1}{2}\|\sum_{i=1}^N \alpha_i y_i x_i\|^2 + \sum_{i=1}^N \alpha_i (1 - (\langle x_i, \sum_{i=1}^N \alpha_i y^i b) - \sum^i \sum^j \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle$
$= \frac{1}{2}\langle \sum \alpha_i y_i x_i, \sum x_j y_j \langle x_i x_j\rangle + \sum \alpha_i N - (\sum y^i \alpha_i b) N -$
$\sum_i \sum_j \alpha_i \alpha_j y^i y_j \langle x_i, x_j\rangle$
$L(w^*, \alpha) = -\frac{1}{2}\sum_i\sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle + \sum^i \alpha_i$

On veut maximiser par rapport à α

7) $\operatorname{argmin}_w \frac{1}{2}\|w\|^2 + W\sum_{i=1}^N \xi_i$ tq $\forall i \; f_w(x_i)y \geq 1 - \xi_i$, $\xi_i \geq 0$ est au min du lagrangien
$L(w,\alpha,\beta) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^N \alpha_i (1 - (\langle x_i, w\rangle + b) y^i - \xi_i) + W\sum_{i=1}^N \xi_i - \sum_{i=1}^N \beta_i \xi_i$
$\frac{\partial L}{\partial \xi_i} = -\alpha_i + W - \beta_i = 0 \Rightarrow W = \alpha_i + \beta_i \Rightarrow \alpha_i \leq W$

8) Ça s'annule dans l'optimisation du lagrangien donc on a $\max_\alpha L(w^*,\alpha) = -\frac{1}{2}\sum\sum \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle + \sum \alpha_i$

9) • $\nabla L = 0$
• $1 - \xi_i - f_w(x_i) y_i < 0$
• $\alpha_i, \xi_i \geq 0$
• $\alpha_i (1 - (\langle x_i, w\rangle + b) y_i - \xi_i) = 0$