APPRENTISSAGE STATISTIQUE - TRAVAUX DIRIGÉS : ÉNONCÉS DES EXERCICES.

ENSAI - Mastères spécialisés - 2025/2026

ENSAI École nationale de la statistique et de l'analyse de l'information

1. Classifieur de Bayes

On considère un n-échantillon de v.a.i.i.d. $\mathscr{D}_n = \{Z_1, ..., Z_n\}$ avec $Z_i = (X_i, Y_i)$. Les X_i sont des observations issues d'une v.a. X, ce sont les données que l'on souhaite classer et qui formeront les variables explicatives. Les Y_i sont issues d'une v.a. Y et sont les catégories auxquelles appartiennent les X_i (on dit également étiquettes ou labels). L'objectif de l'apprentissage supervisé est de déterminer au mieux la catégorie Y à laquelle appartient la donnée X correspondante, à partir des seules observations de l'échantillon $Z_1, ...Z_n$.

On suppose que les v.a. X sont issues d'un espace $\mathbb{X}$, que les v.a. Y sont issues d'un espace $\mathbb{Y}$ et l'on se donne une loi de probabilité (inconnue) $\mathbb{P}$ sur l'espace $\mathscr{E} = \mathbb{X} \times \mathbb{Y}$. $\mathbb{P}$ est la loi de (X, Y) et également la loi jointe commune des (X_i, Y_i).

Une fonction de prédiction est un élément $g \in \mathscr{F} = \mathscr{F}(\mathbb{X}, \mathbb{Y})$ qui associe une étiquette à une observation. Pour mesurer la qualité de g, on définit différentes fonctions de perte $l : \mathbb{Y}^2 \longrightarrow \mathbb{R}_+$ telles que $l(Y, g(X))$ mesure l'écart entre la vraie valeur Y correspondant à X et la valeur $g(X)$ prédite à partir de la fonction g. Le risque de g est la valeur moyenne des réalisations de toutes les pertes possibles. Autrement dit,

$$R(g) = R_{\mathbb{P}}(g) = \mathbb{E}\left[l(Y, g(X))\right] \tag{1}$$

Le prédicteur de Bayes est l'élément $g^\star$ de $\mathscr{F}$ qui minimise la perte $R(g)$. C'est donc la fonction de prédiction optimale sachant les observations.

Dans cet exercice, nous nous limitons au problème de classification binaire, c'est à dire que Y ne peut prendre que deux valeurs : 0 ou 1. La fonction de perte naturelle associée est alors la fonction

$$l(Y, Y') = \mathbb{1}_{[Y \neq Y']} \tag{2}$$

On note

$$\eta(x) = \mathbb{P}[Y = 1 | X = x] = \mathbb{E}[Y | X = x] \tag{3}$$

et

$$g^\star(x) = \mathbb{1}_{[\eta(x) > 1/2]} \tag{4}$$

Nous allons montrer que $g^\star$ minimise l'erreur de classification binaire.

1°. Montrer que

$$\mathbb{P}[Y = g(X) | X = x] = \eta(x)\mathbb{1}_{[g(x)=1]} + (1 - \eta(x))\mathbb{1}_{[g(x)=0]}$$

2°. En déduire que

$$\mathbb{P}[Y \neq g^\star(X) | X = x] \leq \mathbb{P}[Y \neq g(X) | X = x] \tag{5}$$

pour toute fonction g et conclure.

3°. Montrer que le risque de Bayes $R^\star = R(g^\star)$ vérifie

$$R^\star = \mathbb{E}[\eta(X) \wedge (1 - \eta(X))] \tag{6}$$

$$= \frac{1}{2}\left(1 - \mathbb{E}[|2\eta(X) - 1|]\right) \tag{7}$$

avec $x \wedge y = \inf(x, y)$.

4°. Montrer de façon plus générale que quelque soit la fonction f de $\mathbb{X}$ dans $\mathbb{R}$, $\eta(X)$ minimise l'erreur quadratique lorsque $f(X)$ prédit Y. C'est à dire, montrer que

$$\mathbb{E}\left[(\eta(X) - Y)^2\right] \leq \mathbb{E}\left[(f(X) - Y)^2\right] \tag{8}$$

5°. On prédit la réussite d'un étudiant à un examen en fonction du nombre d'heures X passées à travailler. $Y = 1$ signifie que l'étudiant réussit son examen. On suppose que

$$\eta(x) = \frac{x}{x + c} \tag{9}$$

où $c > 0$. Si X suit une loi uniforme sur $[0, 4c]$, calculer $R^\star$.

6°. On suppose que Y suit une loi de Bernoulli de paramètre $p \in]0, 1[$ et que la loi de X sachant $[Y = 0]$ est une loi uniforme sur $[0, 1/2]$, tandis que la loi de X sachant $[Y = 1]$ est une loi uniforme sur $[0, 1]$. Déterminer la loi marginale de X, sa fonction de répartition en fonction de p, sa densité f, puis calculer $\mathbb{E}[Y\mathbb{1}_{[X \leq x]}]$.

Démontrer que pour tout $x \in]0, 1[$,

$$\mathbb{E}[Y\mathbb{1}_{[X \leq x]}] = \int_0^x \eta(u) f(u) du \tag{10}$$

En déduire l'expression de $\eta(x)$. Déterminer la loi conditionnelle de Y sachant $[X = x]$ ainsi que la forme du prédicteur de Bayes.

7°. On dispose de deux variables aléatoires X et Y pour modéliser le comportement de clients. Y est une variable de Bernoulli de paramètre p, valant 1 si le client achète un article à une date donnée et 0 dans le cas contraire. X représente le nombre d'achats qu'il a déjà effectués dans le passé, durant un laps de temps donné. On suppose que la loi conditionnelle de X sachant $[Y = 1]$ est une loi de Poisson de paramètre 2θ tandis que la loi de X sachant $[Y = 0]$ suit une loi de Poisson de paramètre θ, $\theta > 0$. Déterminer l'expression de $\eta(x)$ ainsi que le prédicteur de Bayes.

2. Prédicteur optimal pour la régression au sens des MCO

Démontrer que le prédicteur optimal est

$$g^\star(x) = \mathbb{E}[Y | X = x] \tag{11}$$

3. Prédicteur de Bayes en classification binaire

On suppose que Y suit une loi de Benoulli de paramètre $p \in]0,1[$. On considère X, dont la loi conditionnelle sachant Y est définie comme suit : la loi de X sachant $[Y=0]$ est une loi uniforme sur $[0,1/2]$ et la loi de X sachant $[Y=01$ est une loi uniforme sur $[0,1]$.

1°. Quelle est la loi marginale de X ? Déterminer sa fonction de répartition en fonction de p.

2°. Déterminer la densité f de X par rapport à la mesure de Lebesgue.

3°. Pour tout $x \in [0,1]$, calculer $\mathbb{E}[Y\mathbb{1}_{[X\leq x]}]$.

4°. On pose $\eta^\star(x) = \mathbb{E}[Y|X=x]$. Montrer que pour tout $x \in]0,1[$,

$$\mathbb{E}[Y\mathbb{1}_{[X\leq x]}] = \int_0^x \eta^\star(u) f(u)\, du \tag{12}$$

et en déduire l'expression de $\eta^\star(x)$.

5°. Déterminer la loi conditionnelle de Y sachant X ainsi que la forme du classifieur de Bayes.

4. Prédicteur de Bayes en classification binaire

On considère le problème de classification binaire avec $\mathbb{Y} = \{0,1\}$ et $\mathbb{X} = \mathbb{N}$. On suppose que $(X_i, U_i)_{i\in\mathbb{N}}$ est une suite de vecteurs aléatoires i.i.d. telle que pour chaque i, X_i suit une loi de Poisson de paramètre θ, U_i suit une loi de Poisson de paramètre γ, X_i et U_i sont indépendantes. L'étiquette Y_i de chaque exemple X_i est déterminée par

$$Y_i = \mathbb{1}_{[X_i+U_i\geq\lambda]}. \tag{13}$$

Les paramètres θ, γ, λ sont inconnus. Les variables U_i ne sont pas observables. On ne dispose que d'un échantillon (X_i, Y_i) pour $i = 1,...,n$ et on veut en déduire une règle de classification.

1°. Déterminer la loi marginale de Y_1 en fonction de (θ,γ) lorsque $\lambda = 1$.

2°. Calculer la fonction de régression $\eta^\star(x)$.

3°. Déterminer la forme du classifieur de Bayes.

4°. En utilisant le résultat de la question précédente, proposer une méthode de classification qui tire profit de la forme de la loi des observations (X_i, Y_i).

5. Minimisation de l'erreur stochastique

Soit $g^\star$ le prédicteur optimal minimisant le risque moyen sur un dictionnaire $\mathscr{G}$.
Soit g_n le minimiseur du risque empirique à partir de l'échantillon $\mathscr{D}_n$ sur un $\mathscr{G}$.
Ces deux fonctions dépendent de $\mathscr{G}$, même si $\mathscr{G}$ n'apparaît pas dans la notation.

On cherche à quantifier à quel point le risque de prédiction de $R(\widehat{g}_n)$ est éloigné du risque de prédiction minimal $R(g^\star)$.

1°. Expliquer pourquoi $R(g^\star) \leq R(\widehat{g}_n)$ et $R_n(g^\star) \geq R_n(\widehat{g}_n)$.

2°. Montrer que

$$R(\widehat{g}_n) - R_n(\widehat{g}_n) \leq \max_{g\in\mathscr{G}} |R(g) - R_n(g)|$$
$$R(g^\star) - R_n(g^\star) \leq \max_{g\in\mathscr{G}} |R(g) - R_n(g)|$$

3°. En déduire que :

$$0 \leq R(\widehat{g}_n) - R(g^\star) \leq 2\max_{g\in\mathscr{G}} |R(g) - R_n(g)|$$

4°. Interpréter les inégalités précédentes en termes de biais et de fluctuation ou variance.

6. Inégalité oracle pour un dictionnaire fini

On considère une fonction de perte l à valeurs dans $[0,1]$ et un dictionnaire fini $\mathscr{G}$ contenant M fonctions. On pose $\delta \in]0,1[$.

On considère à nouveau un n-échantillon de v.a.i.i.d. $\mathscr{D}_n = \{Z_1,...,Z_n\}$ avec $Z_i = (X_i, Y_i)$. Les X_i sont des observations issues d'une v.a. X, les Y_i sont issues d'une v.a. Y et sont les catégories auxquelles appartiennent les X_i. On suppose que X est issu d'un espace $\mathbb{X}$ et Y d'un espace $\mathbb{Y}$. $\mathbb{P}$ définie sur $\mathscr{E} = \mathbb{X}\times\mathbb{Y}$ est la loi inconnue de (X,Y) .

$g \in \mathscr{G} \subset \mathscr{F}(\mathbb{X},\mathbb{Y})$ est une fonction de prédiction. Le risque moyen est

$$R(g) = \mathbb{E}\left[l(Y, g(X))\right] \tag{14}$$

et $g^\star$ est le risque minimal optimal sur $\mathscr{G}$ qui minimise la perte $R(g)$.

Soit enfin $\widehat{g}_n$ le minimiseur du risque empirique R_n calculé à partir de $\mathscr{D}_n$.

Le but de cet exerice est de démontrer qu'avec probabilité supérieure à $1-\epsilon$,

$$R(\widehat{g}_n) - R(g^\star) \leq \sqrt{\frac{2}{n}\ln\left(\frac{2M}{\epsilon}\right)} \tag{15}$$

1°. Commencer par interpréter et discuter ce résultat.

2°. Rappeler l'expression de l'inégalité de Hoeffding.

3°. En utilisant cette inégalité, en déduire l'inégalité recherchée.

7. Classifieur non binaire

Soient $\mathbb{Y} = \{a_1,...,a_K\}$ et $\mathbb{X}$ sous ensemble mesurable de $\mathbb{R}^d$. On considère le cadre d'apprentissage supervisé à partir d'un échantillon $\mathscr{D}_n = \{(X_i, Y_i), i = 1,..,n\}$ de loi commune $\mathbb{P}$ sur $\mathbb{R}^d \times \mathbb{Y}$. On note ν la mesure de Lebesgue sur $\mathbb{R}^d$ et $\delta_\cdot$ la mesure de Dirac sur $\mathbb{Y}$. On dit que la fonction $f : \mathbb{R}^d \times \mathbb{Y}$ à valeurs dans $\mathbb{R}_+$ est la densité de $\mathbb{P}$ par rapport à la mesure produit

$$\left(\sum_{k=1}^K \delta_{a_k}\right) \otimes \nu, \tag{16}$$

si pour toute fonction h de $\mathbb{R}^d \times \mathbb{Y}$ dans $\mathbb{R}$ mesurable et bornée, on a :

$$\mathbb{E}[h(X,Y)] = \sum_{k=1}^{K} \int_{\mathbb{R}^d} h(x,a_k) f(x,a_k) dx \tag{17}$$

1°. Montrer que pour tout classifieur $g : \mathbb{X} \longrightarrow \mathbb{Y}$, on a :

$$\mathbb{P}[g(X) \neq Y] = 1 - \int_{\mathbb{R}^d} f(x, g(x)) dx \tag{18}$$

2°. En déduire que si $\mathbb{P}$ est une loi à densité, alors le classifieur oracle est donné par :

$$g^\star(x) = \arg\max_{a \in \mathbb{Y}} f(x,a) \tag{19}$$

3°. Soit $K = 2$ avec $a_1 = 0$ et $a_2 = 1$. Montrer que le classifieur oracle coïncide avec le classifieur oracle binaire du cours.

4°. On suppose maintenant que $\mathbb{Y} = \{0,1\}$ et que $f(x,k)$ est la densité de la loi gaussienne de moyenne μ_k et matrice de covariance Σ. Déterminer la forme du classifieur de Bayes et montrer qu'il coïncide avec la règle de classification linéaire de Fisher. Proposer un estimateur simple de $g^\star$ dans ce contexte, basé sur $\mathscr{D}_n$.

8. Consistance universelle uniforme

Cet exercice est difficille, inutile d'essayer de le faire, vous n'y arriverez jamais. On reprend les hypothèses et notations vues en classification binaire (c.f. exercice 1).

1°. On suppose que $\mathbb{X}$ est fini, de cardinal K. Quel est le cardinal de $\mathscr{F}(\mathbb{X},\mathbb{Y})$?

2°. Rappeler la borne de risque obtenue pour un dictionnaire fini. Peut-on déduire que $\widehat{g}_n$ est uniformément et universellement consistant ?

3°. On suppose que $\mathbb{X}$ est infini et que $K = K_n$ dépend de la taille de l'échantillon. Démontrer que si K_n / n tend vers 0 quand n tend vers l'infini, alors $\widehat{g}_n$ est uniformément et universellement consistant.

9. Minimiseur du risque empirique en classification et régression pour une méthode à partition

Soit $\mathscr{A} = (A_1, ..., A_M)$ une partition de $\mathbb{X}$ et $\mathscr{G}(\mathscr{A}) \subset \mathscr{F}(\mathbb{X},\mathbb{Y})$ l'ensemble des fonctions constantes sur chaque $A_m \in \mathscr{A}$.
Soit $\widehat{g}_n = \widehat{g}_n(.,\mathscr{A})$ le minimiseur du risque empirique sur le dictionnaire $\mathscr{G}(\mathscr{A})$ et N_m le nombre d'exemples appartenant à A_m :

$$\widehat{g}_n \in \arg\min_{g \in \mathscr{G}(\mathscr{A})} R_n(g) = \arg\min_{g \in \mathscr{G}(\mathscr{A})} \frac{1}{n} \sum_{i=1}^{n} l(Y_i, g(X_i))$$

et $N_m = \sum_{i=1}^{n} \mathbb{1}_{A_m}(X_i)$

1°. Montrer que pour le problème de régression,

$$\begin{aligned} \widehat{g}_n(x,\mathscr{A}) &= \sum_{m=1}^{M} \overline{Y}_{A_m} \mathbb{1}_{A_m}(x) \\ &= \sum_{m=1}^{M} \left(\frac{1}{N_m} \sum_{i=1}^{n} Y_i \mathbb{1}_{A_m}(X_i) \right) \mathbb{1}_{A_m}(x) \end{aligned}$$

2°. Montrer que pour le problème de classification binaire, $\forall m = 1, ..., M$, $\forall x \in A_m$,

$$\widehat{g}_n(x,\mathscr{A}) = \begin{cases} 1 & \text{si } \overline{Y}_{A_m} > 1/2 \\ a_m & \text{si } \overline{Y}_{A_m} = 1/2 \\ 0 & \text{si } \overline{Y}_{A_m} < 1/2 \end{cases}$$

10. Non consistance de k-NN pour $k = 1$

On pose $\mathbb{X} = [0,1]$ et $\mathbb{Y} = \{0,1\}$. Soit $\mathbb{P}_X$ la loi marginale des X_i. X et Y sont des va génériques de même loi que les X_i et Y_i. On suppose que

$$\eta^\star(x) = \mathbb{P}[Y = 1 | X = x] \equiv \frac{3}{4}, \ \forall x \in \mathbb{X}. \tag{20}$$

L'objectif des questions suivantes est de calculer le risque du classifieur oracle $g^\star$ ainsi que celui du classifieur k-NN avec $k = 1$. On verra que ce dernier ne dépend pas de la taille de l'échantillon et est strictement plus grand que le risque de l'oracle.

1°. Montrer que pour toute fonction g de $\mathbb{X}$ dans $\{0,1\}$,

$$R(g) = \mathbb{E}[\eta^\star(X)] + \mathbb{E}[g(X)(1 - 2\eta^\star(X)] \tag{21}$$

2°. En déduire que si $\eta^\star \equiv 3/4$, alors le classifieur oracle de Bayes est donné par $g^\star \equiv 1$ et que son risque vaut $R(g^\star) = 1/4$.

3°. Montrer que pour toute fonction g,

$$R(g) = \frac{3}{4} - \frac{1}{2} \int_{\mathbb{X}} g(x) \mathbb{P}_X(dx) \tag{22}$$

4°. Soit $\widehat{g}_n(x)$ le classifieur au sens des plus proches voisins (ppv). Posons $Z_i = \mathbb{1}_{[X_i \text{ est le ppv de x}]}$. Montrer que

$$\mathbb{E}[\widehat{g}_n(x)] = \sum_{i=1}^{n} \mathbb{E}[Y_i Z_i] \tag{23}$$

5°. Montrer que Y_i et Z_i sont indépendantes et que $\sum_{i=1}^{n} Z_i = 1$.

6°. En déduire que

$$\mathbb{E}[\widehat{g}_n(x)] = \frac{3}{8} \tag{24}$$

7°. Considérer le cas du minimiseur du risque $\widehat{g}_{3,n}$ pour l'algorithme des 3-ppv. Montrer que son risque moyen $\mathbb{E}[R(\widehat{g}_{3,n})]$ est égal à $21/64$.

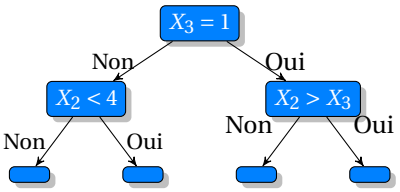
8°. Pour le cas général du prédicteur $\widehat{g}_{k,n}$, considérer $V_1, ..., V_k$ des vaiid de loi de Bernoulli de paramètre $3/4$. Montrer que

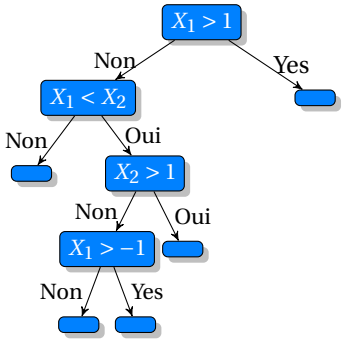
$$\mathbb{E}[\widehat{g}_{k,n}(x)] = \mathbb{P}[\overline{V}_k > 1/2] \tag{25}$$

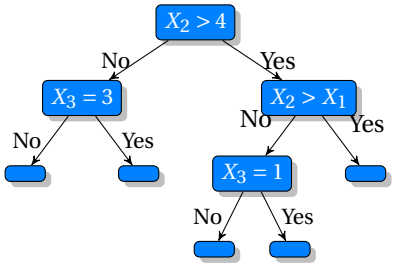
en déduire que cette espérance tend vers 1 lorsque k tend vers l'infini et que le risque espéré tend vers le risque de l'oracle, c'est à dire $1/4$.

11. Arbres de décision

On considère le problème de classification binaire avec $\mathbb{Y} = \{0,1\}$ et $\mathbb{X} \subset R^2 \times \{1,2,3\}$. Soit $\mathscr{D}_n = (Z_1, ..., Z_n)$ n vaiid de loi $\mathbb{P}$. On suppose que l'on dispose de 3 arbres de décision suivants :







qui fournissent les prédicteurs $\widehat{g}_1$, $\widehat{g}_2$ et $\widehat{g}_3$. On note $\mathscr{A}_1$, $\mathscr{A}_2$ et $\mathscr{A}_3$ les partitions de $\mathbb{X}$ engendrées par ces trois arbres.

1°. Lesquel des 3 prédicteurs a le moins de risque de sur-apprendre?

2°. Soit $\mathscr{G}_i$ l'ensemble des prédicteurs constants par morceaux sur la partition $\mathscr{A}_i$ et soit $\mathscr{G}$ la réunion des trois. On note $\widehat{g}_n$ le minimiseur du risque empirique sur $\mathscr{G}$:

$$\widehat{g}_n \in \operatorname*{argmin}_{g\in\mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i\neq g(X_i)} = \operatorname*{argmin}_{g\in\mathscr{G}} R_n(g) \tag{26}$$

Montrer que

$$\widehat{g}_n \in \operatorname*{argmin}_{g\in\{\widehat{g}_1,\widehat{g}_2,\widehat{g}_3\}} R_n(g) \tag{27}$$

4°. On observe un échantillon de taille $n = 10$ comportant les valeurs suivantes :

X_1	0.1	-0.3	1.2	1.8	-0.9	0.5	-1.1	1.4
X_2	-0.4	1.1	0.2	-0.3	-0.5	0.6	-0.7	-0.6
X_3	1	3	1	2	2	2	1	1
Y	0	0	1	0	1	1	1	0

X_1	-1.3	1.01
X_2	0.9	0.81
X_3	1	3
Y	1	0

Déterminer les valeurs de $\widehat{g}_1$ et $\widehat{g}_3$ sur chacune des feuilles.

12. Convexification du problème de minimisation du risque empirique

Soit $g : \mathbb{R} \longrightarrow \mathbb{R}_+$ et $G : \mathbb{R} \longrightarrow [0,1]$ la densité et la fonction de répartition de la loi gaussienne $\mathscr{N}(0,1)$. On définit la fonction de perte $\phi(u)$ part la formule

$$\phi(u) = g(u) + uG(u),\ \forall u \in \mathbb{R}. \tag{28}$$

1°. ϕ est-elle monotone? Convexe?

2°. On définit $\psi : [0,1] \longrightarrow \mathbb{R}$ par

$$\psi(p) = \inf_{u\in\mathbb{R}} \left[p\phi(-u) + (1-p)\phi(u)\right], \tag{29}$$

pour $u \in [0,1]$. Déterminer les valeurs de ψ en $0, 1$ et $1/2$.

3°. Montrer que

$$\psi(p) = \inf_{u\in\mathbb{R}} (\phi(u) - pu). \tag{30}$$

4°. Soit $Q :]0,1[\longrightarrow \mathbb{R}$ la fonction quantile de la loi gaussienne standard. Elle est définit par :

$$G(Q(p)) = p \text{ et } Q(G(x)) = x,\ \forall p \in]0,1[, x \in \mathbb{R}. \tag{31}$$

En utilisant la question précédente, montrer que $\psi(p) = g(Q(p))$.

5°. Vérifier que $Q(1/2) = 0$ et en déduire que $\psi'(1/2) = 0$.

6°. Vérifier que $\psi(G(x)) = g(x)$ pour tout x réel. En déduire les expressions de $\psi'(G(x))$ et $\psi''(G(x))$ pour tout x réel.

7°. Déterminer les constantes c et γ du lemme de Zhang et écrire l'inégalité qui relie l'excès de risque de classification à l'excès du ϕ-risque pour la fonction ϕ définie en début d'exercice.

13. Conditions de consistance pour le minimiseur du ϕ-risque

Le minimiseur du ϕ-risque empirique est défini par

$$\widehat{h}_n = \operatorname*{argmin}_{h\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^{n} \phi(-Y_i h(X_i)). \tag{32}$$

Il dépend de $\mathscr{H}$. La consistance de ce classifieur peut être obtenue en utilisant le résultat suivant :

Soit $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction convexe telle que $u(\phi(u) - \phi(-u)) \geq 0$ pour tout $u \in \mathbb{R}$. Soit ψ définie par

$$\psi(p) = \inf_{u\in\mathbb{R}} \left[p\phi(-u) + (1-p)\phi(u)\right]. \tag{33}$$

S'il existe $\gamma \in [0,1]$ et $c > 0$ tels que

$$|1-2p| \leq c[\phi(0) - \psi(p)]^{\gamma},\ \forall p \in [0,1], \tag{34}$$

alors quelque soit la fonction de prédiction h,

$$R(\operatorname{sgn}(h)) - R(g^\star) \leq c[A(h) - A(h^\star]^{\gamma}. \tag{35}$$

Le but de cet exercice est d'étudier la fonction ψ et de vérifier que la perte de Boosting $\phi(u) = e^u$ vérifie les conditions de ce théorème.

1°. Montrer que $\phi(0) = \psi(1/2)$.

2°. Vérifier l'inégalité $\psi(p) = \phi(0)$ et en déduire que la condition (34) équivaut à

$$\psi(1/2) - \psi(p) \geq \left(\frac{2}{c}\left|\frac{1}{2} - p\right|\right)^{1/\gamma}. \tag{36}$$

3°. Montrer que $\psi(p) = \psi(1-p)$. En déduire que si l'inégalité (36) est satisfaite pour tout $p \in [0, 1/2]$, alors elle l'est pour tout $p \in [0,1]$.

4°. Montrer que ψ est une fonction concave et que si $\psi'(1/2)$ existe, alors $\psi'(1/2) = 0$. En déduire que si ψ est deux fois différentiable alors

$$\psi(1/2) - \psi(p) \geq \left(p - \frac{1}{2}\right)^2 \inf_p |\psi''(p)|. \tag{37}$$

5°. Soit ψ une fonction deux fois continûment différentiable sur $[0, 1/2]$ avec

$$\sup_{u \in]0,1/2]} \psi''(u) = -a < 0. \tag{38}$$

Montrer que la condition (34) est remplie avec $\gamma = 1/2$ et $c = \sqrt{8/a}$. Trouver γ et c qui correspondent à la perte de Boosting $\phi(u) = e^u$.

14. SVM : résolution du problème d'optimisation cas linéairement séparable

1°. Écrire le problème d'optimisation primal lié aux machines à vecteur de support, en gardant les notations du cours.

2°. En déduire le lagrangien correspondant.

3°. Préciser les conditions KKT de Karush-Kuhn-Tucker.

4°. Déterminer le problème dual associé.

5°. Décrire la méthode de résolution du problème et la forme de la solution.

15. SVM : problème d'optimisation cas non linéairement séparable

1°. Écrire le problème d'optimisation primal lié aux machines à vecteur de support, dans le cas d'une marge souple.

2°. En déduire le lagrangien correspondant.

3°. Préciser les conditions KKT de Karush-Kuhn-Tucker.

4°. Déterminer le problème dual associé.

5°. Décrire la méthode de résolution du problème.

6°. Que devient le problème dual lorsqu'on utilise un noyau non linéaire ? Expliquer pourquoi w et b n'interviennent plus.

16. Adaboost

On observe un échantillon $\mathscr{D}_n = \{(x_i, y_i)\}_{i=1\ldots n}$ avec $y_i \in \{-1, +1\}$. On considère une famille de classifieurs faibles $\mathscr{H}$ à valeurs dans $\{-1, +1\}$.

On rappelle l'algorithme AdaBoost (version binaire) pour un nombre d'itérations M :

1. Initialiser $w_i(0) = 1/n$ pour tout i.
2. Pour $m = 1, \ldots, M$:
   - Apprendre $h_m \in \mathscr{H}$ minimisant l'erreur pondérée

$$\varepsilon_m = \mathbb{P}[h(x_i) \neq y_i] = \sum_{i=1}^{n} \frac{w_i(m)}{||w||_1} \mathbb{1}_{[h(x_i) \neq y_i]}.$$

   - Poser $\alpha_m = \frac{1}{2}\ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right)$.
   - Mettre à jour $\forall\, i = 1, .., n$:

$$w_i(m+1) = w_i(m) \exp\left(\alpha_m \mathbb{1}_{[y_i \neq h(x_i)]}\right).$$

3. Le classifieur final est

$$\widehat{h}_M(x) = \operatorname{sgn}\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right).$$

1°. Rappeler l'expression de la perte Boosting et du ϕ-risque empirique noté $A_n(h)$ associé à cette perte Boosting.

2°. Démontrer l'inégalité $\mathbb{1}_{[y_i h(x_i) \leq 0]} \leq \exp(-y_i h(x_i))$.

3°. Montrer que le choix de (h_m, α_m) ci-dessus réalise, à chaque itération, le minimum de $A_n(h)$.

17. Bagging

On considère M échantillons indépendants $\mathscr{D}_{n,m}$, de taille n, chacun étant utilisé pour entraîner un estimateur h_m de régression au sens des moindres carrés et l'on note h_M leur moyenne empirique.

1°. Étudier comment varie le biais et la variance de h_M quand M augmente.

On ne dispose maintenant que d'un seul échantillon $\mathscr{D}_n$. Á partir de cet échantillon, on construit M échantillons $\mathscr{D}_{n,m}$ de taille n en effectuant un tirage uniforme avec remise dans $\mathscr{D}_n$. On entraîne un estimateur de régression au sens des moindres carrés h_m pour chacune de ces échantillons bootstrap et on note à nouveau h_M leur moyenne empirique.

2°. Comparer, en termes de biais et de variance, le comportement de h_M construit par bootstrap à celui de h_M dans le cas idéal où l'on dispose de M échantillons indépendants.

3°. Expliquer en quoi cette procédure peut permettre de réduire l'erreur de généralisation d'un estimateur de régression instable (par exemple un arbre de décision). Comment appelle-t-on cette méthode ?