MASEF/104 19th January 2021

Exam : Machine Learning in Finance [1] : Time 1h30

Exercise 1. [**10pts**]

Select the correct answers

1. By assumption, the $(X^i, Y^i)_{i\in[\![1,n]\!]}$ in the learning sample, in supervised learning are such :
   a) the (X^i, Y^i) have all the same laws and are independent
   b) the (X^i, Y^i) have the same laws but may be dependent
   c) the X^i have all a normal distribution
   Answers : a

2. if $(X^i)_{i\in[\![1,n]\!]}$ is a learning sample without any labelling.
   a) we are in the framework of supervised learning
   b) we are in the framework of unsupervised learning
   c) generally we try to derive a structure for the X^i
   Answers : b,c

3. if $R_n(f_n)$ is the classification error and $R(f_n)$ is the prediction error then
   a) $R(f_n) < R_n(f_n)$
   b) $E(R(f_n)) \geq R_n(f_n)$
   c) $R(f_n) \geq E[R_n(f_n)]$
   d) $E(R(f_n)) \leq R_n(f_n)$
   Answers : c

4. in Vapnik Chervonenkis's formula
   a) $P\left(R(f_n) \leq R_n(f_n) + \phi_{n,\eta}\left(\frac{VC(\mathcal{F})}{n}\right)\right) \geq 1-\eta$
   b) $P\left(R_n(f_n) \leq R(f_n) + \phi_{n,\eta}\left(\frac{VC(\mathcal{F})}{n}\right)\right) \geq 1-\eta$
   c) $P\left(R_n(f_n) \leq R(f_n) + \phi_{n,\eta}\left(\frac{VC(\mathcal{F})}{n}\right)\right) \geq 1-\frac{\eta}{n}$
   a) $P\left(R(f_n) \leq R_n(f_n) + \phi_{n,\eta}\left(\frac{VC(\mathcal{F})}{n}\right)\right) \geq 1-\frac{\eta}{n}$
   Answers : a

5. in the Vapnik Chervonenkis's formula
   a) $\phi_{n,\eta}(t)$ is a function which is decreasing with t
   b) $\phi_{n,\eta}(t)$ is a function which is increasing with t
   c) $\phi_{n,\eta}(\frac{1}{n})$ tends to zero when n tends to ∞
   Answers : b,c

1. Pierre Brugière University Paris Dauphine PSL

6. in Structural Risk Minimisation with $\mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_k \cdots$ the classifier $f_{n,k}$ from class $\mathcal{F}_k$ is chosen if
   a) it minimizes over k, $R_n(f_{n,k}) + \phi_{n,\eta}(\frac{VC(\mathcal{F}_k)}{n})$
   b) it minimizes over k, $R_n(f_{n,k})$
   c) it minimizes over k, $R_n(f_{n,k}) - \phi_{n,\eta}(\frac{VC(\mathcal{F}_k)}{n})$
   Answers : a

7. if $x_1, x_2, \cdots x_n$ is a family of independent vectors then
   a) $x_1, x_2, \cdots x_n$ can be classified in all possible ways with hyperplanes
   b) 0 can be separated from $x_1, x_2, \cdots x_n$ with a hyperplane
   c) 0 can be separated from the convex envelope of $x_1, x_2, \cdots x_n$ with a hyperplane
   Answers : a,b,c

8. if a family of classifiers $\mathcal{F}$ is defined by d parameters $d \geq 1$ then
   a) $VC(\mathcal{F}) < d + 2$
   b) $VC(\mathcal{F}) = d + 1$
   c) there may be some cases for which $VC(\mathcal{F}) = +\infty$
   Answers : c

9. what is the distance between the point represented by the vector z and the hyperplane of equation $\langle w, x \rangle = b$
   a) $\frac{|b+\langle w,z \rangle|}{\|w\|^2}$
   b) $\frac{|\langle w,z \rangle - b|}{\|w\|}$
   c) $\frac{|\langle w,z \rangle - b|}{\|w\|^2}$
   Answers : b

10. what is the distance between the two hyperplanes of equations : $\langle w, x \rangle = b$ and $\langle -w, x \rangle + c = 0$
    a) $\frac{|b-c|}{\|w\|^2}$
    b) $\frac{|b+c|}{\|w\|}$
    c) $\frac{|b-c|}{\|w\|}$
    Answers : c

11. in $\mathbf{R}^d$
    a) the VC dimension of hyperplane classifiers of margin 1 is $d + 1$
    b) the VC dimension of hyperplane classifiers within a ball of radius 1 is $d + 1$
    c) the VC dimension of hyperplane classifiers of radius 1 and margin 0.1 is less than $d + 1$ when d is large
    Answers : a,b,c

12. in a C-SVM
    a) all the support vectors are always on the borders of the separating hyperplanes
    b) it is possible that some support vectors are not correctly classified
    c) it is impossible that some support vectors lay between the two separating hyperplanes
    Answers : b

13. which of these inequalities is correct :
    a) $\sup_{z\in\mathcal{Z}}\left[\inf_{y\in\mathcal{Y}} g(y,z)\right] \leq \inf_{y\in\mathcal{Y}}\left[\sup_{z\in\mathcal{Z}} g(y,z)\right]$
    b) $\sup_{z\in\mathcal{Z}}\left[\inf_{y\in\mathcal{Y}} g(y,z)\right] \geq \inf_{y\in\mathcal{Y}}\left[\sup_{z\in\mathcal{Z}} g(y,z)\right]$
    Answers : a

14. which of the following assertions are true :
    a) if the KKT conditions are satisfied the primal and dual problems have the same value
    b) if the primal and dual problems have the same value the KKT conditions are satisfied
    c) the KKT conditions enable to calculate some of the parameters of the separating hyperplanes
    Answers : a,b,c

15. which of the following assertions are true for the Gaussian Kernel K_σ and associated transformation ϕ_σ :
    a) all the transformed points $\phi_\sigma(x_i)$ are on sphere or radius 1
    b) for any sample $\{(x_i, y_i)\}_{i\in[\![1,n]\!]}$ it is possible to find σ such that the classification with the Gaussian Kernel K_σ is perfect
    c) a notion of margin can be associated to Gaussian Kernel classifiers
    Answers : a,b,c

16. single class SVMs can be used
    a) to estimate the support of a distribution
    b) to define different clusters amongst a sample
    c) to solve a regression problem
    Answers : a,b

17. if $K(.,.)$ is a kernel with values in $\mathbf{R}$ and $(H, \langle .,. \rangle_{RK})$ is a Reproducing Kernel Hilbert Space for K then
    a) for all $f \in H$, $\langle K(x,.), f\rangle_{RK} = f(x)$
    b) $K(x_1,x_2)^2 \leq K(x_1,x_1)K(x_1,x_2)$
    c) for all $f \in H$, $f(x)^2 \leq K(x,x)\langle f, f\rangle_{RK}$
    Answers : a,b,c

18. in the discount factor curve construction problem studied in the course
a) the regularisation term is linked to a scalar product on the function space
b) the regularisation term transforms a non parametric problem into a parametric problem
c) without the regularisation term the problem would not be well defined
Answers : a,b,c

19. in the discount factor curve construction problem studied in the course
a) the spline functions are of degrees 2
b) the spline functions have knots on the cash flow dates of the instruments used as inputs
c) the coefficients of the spline functions are the solutions of a ridge regression problem
d) the coefficients of the spline functions are the solutions of a Lasso regression problem
Answers : b,c

20. in the Smith and Nelson model the regularisation term is of the form
a) $\int f''^2(s)ds$
b) $\int f''^2(s) + \sigma^2 f'^2(s)ds$
c) $\int f''(s) + \sigma^2 f'(s)ds$
Answers : b

Exercise [10pts]

Let $\{(x_i, y_i)\}_{i\epsilon[\![1,n]\!]}$ be a learning sample derived from (X, Y) with $X \in \mathbf{R}^d$ and $Y \in \{-1, 1\}$. Let's consider the C-SVM defined by

$$(P_C) \begin{cases} \inf\limits_{w,b,\xi} \|w\|^2 + C \sum\limits_{i=1}^{i=l} \xi_i \\ y_i[\langle w, x_i\rangle + b] \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

1. **[2pts]** Show that if (w^*, b^*, ξ^*) is a solution of (P_C) then necessarily : $\forall i \in [\![1, n]\!], \xi_i^* = max(0, 1 - y_i[\langle w^*, x_i\rangle + b^*])$
Proof : if it was not the case there would be an index i such that

$$\xi_i^* > max(0, 1 - y_i[\langle w^*, x_i\rangle + b^*])$$

then if we define ξ^{**} by : $\begin{cases} \forall i \neq j, \xi_i^{**} = \xi_i^* \text{ and} \\ \xi_i^{**} = max(0, 1 - y_i[\langle w^*, x_i\rangle + b^*]) \end{cases}$
then $(w^{**}, b^{**}, \xi^{**})$ would be a better solution than (w^*, b^*, ξ^*), which would lead to a contradiction Q.E.D.

From now on we assume that for all $C > 0$, (P_C) has a unique solution in (w, b, ξ) reached for (w_C^*, b_C^*, ξ_C^*) and we consider the problem (Q_λ) defined by :

$$Q_\lambda : \inf_{w,b} \sum_{i=1}^{n} max(0, 1 - y_i[\langle w, x_i \rangle + b]) + \lambda \|w\|^2$$

2. **[1pt]** if we assume that for all $\lambda > 0$ (Q_λ) has a unique solution in (w, b) reached for $(\tilde{w}_\lambda, \tilde{b}_\lambda)$ show that :
$\begin{cases} \tilde{w}_\lambda = w^*_{\frac{1}{\lambda}} \text{ and} \\ \tilde{b}_\lambda = b^*_{\frac{1}{\lambda}} \end{cases}$
Proof : According to the reasoning in 1, the solutions of (P_C) are to be searched amongst (w, b, ξ) such that

$$\xi_i = max(0, 1 - y_i[\langle w, x_i \rangle + b])$$

therefore the solutions of (P_C) solve

$$\inf_{w,b} \|w\|^2 + C \sum_{i=1}^{i=l} max(0, 1 - y_i[\langle w, x_i \rangle + b])$$

which has the same argument solutions as (Q_λ) for $\lambda = \frac{1}{C}$. Q.E.D

3. **[2pt]** (independent from the rest) determine a function $f : \mathbf{R}^d \longrightarrow [-1, 1]$ solution of :

$$\arg\min_f E(max(0, 1 - Yf(X)))$$

Proof : here $max(0, 1 - Yf(X)) = 1 - Yf(X)$ and
$\arg\min_f E(1 - Yf(X)) = \arg\max_f E(Yf(X)) = \arg\max_f E(g(X)f(X))$
where $g(X) = E[Y|X]$. So, as f takes value in $\{-1, 1\}$ the maximum is reached by taking $f(X) = 1$ when $g(X) > 0$ and $f(X) = -1$ otherwise. Now, $g(x) = P(Y = 1|X = x] - P(Y = -1|X = x]$ so we can define f by
$\begin{cases} f(x) = 1 \text{ if } P(Y = 1|X = x) > P(Y = -1|X = x) \text{ and} \\ f(x) = -1 \text{ otherwise.} \end{cases}$.

Let $P_{w,b}$ be defined by

$$P_{w,b}(Y = 1|X = x) = \sigma(\langle w, x \rangle + b)$$

where $\sigma : \mathbf{R} \longrightarrow \mathbf{R}^+$ is the sigmoid function defined by

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

4. **[0.5pt]** show that $P_{w,b}(Y = -1|X = x) = \sigma(-\langle w, x \rangle - b)$
Proof :
Results from $1 - \sigma(z) = \sigma(-z)$ as $1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} = \frac{1}{1+e^z}$

5. **[0.5pt]** write the expression of the log likelihood of the sample under $P_{w,b}$, i.e write the expression of : $\ln\big(\prod_{i=1}^{n} P_{w,b}(Y = y_i|X = x_i)\big)$ as a function of the parameters w, b and the observations (x_i, y_i).
Proof : $\ln\big(\prod_{i=1}^{n} P_{w,b}(Y = y_i|X = x_i)\big) = \sum_{i=1}^{n} \ln\big(\sigma\big(y_i[\langle w, x_i\rangle + b]\big)\big)$
$= -\sum_{i=1}^{n} \ln\big(1 + exp(-y_i[\langle w, x_i\rangle + b)\big)$

6. **[1pt]** show that the problem (LR_β) of maximising the log likelihood of the sample under $P_{w,b}$ penalised by a cost $\beta > 0$ on the value of $\|w\|^2$ can be written as

$$(LR_\lambda) : \inf_{w,b} \sum_{i=1}^{i=l} l(y_i, \langle w, x_i\rangle + b) + \beta\|w\|^2$$

where $l(.,.)$ is a function that you will explicit
Proof : $\sup_{w,b} \ln\big(-\sum_{i=1}^{n} \ln\big(1 + exp(-y_i[\langle w, x_i\rangle + b)\big)$
$= -\inf_{w,b} \ln\big(\sum_{i=1}^{n} \ln\big(1 + exp(-y_i[\langle w, x_i\rangle + b)\big)$
so, with the penalisation term the problem to solve is
$-\inf_{w,b} \ln\big(\sum_{i=1}^{n} \ln\big(1 + exp(-y_i[\langle w, x_i\rangle + b)\big) + \lambda\|w\|^2$
so, $l(y, z) = ln(1+e^{-yz})$. (which is called a loss function even if contrarily to some other loss functions it does not vanish when $y = z$)

7. **[1pt]** what relationship do you see between solving the problem (LR_β) (which is a penalised logistic regression problem) and solving the problem (P_C) which is a C-SVM problem ?
Solution : the problem (P_C) is equivalent to $(Q_{\frac{1}{C}})$ which, if $\beta = \frac{1}{C}$ differs only from problem (LR_β) by the loss function which is $max(0, 1-yz)$ instead of $ln(1 + e^{-yz})$. So a logistic regression problem corresponds to a C-SVM problem for which the loss function would be modified.

8. **[1pt]** After running a C-SVM if you were asked to define a probability for a point $x \in \mathbf{R}^d$ to belong to a class $y \in \{-1, 1\}$ what formula would you think of, based on the parameters w_C^* and b_C^* you have calculated ?
Solution : $P_e(Y = y_i|X = x_i) = \sigma(\langle w_C^*, x_i\rangle + b_C^*)$ seems natural and is called Platt's probability

9. **[1pt]** What is the main difference you see between a penalised logistic regression and a C-SVM in terms of the number of points useful to solve the optimisation problem and the robustness to outliers.
Solution : in a penalised logistic regression all points will influence the

choice of the optimal parameters even outliers, while in a C-SVM many points have no influence on the determination of the optimal parameters and can be disregarded.