

**Master ISF Apprentissage - Mido 2018-2019**  
Exam : Machine Learning in Finance<sup>1</sup> : Duration 2h

**Exercise 1 (QCM) : [10pts]**

1. If  $(X_i, Y_i)_{i \in \llbracket 1, n \rrbracket}$  is a learning sample, how is the calibration error generally defined for a classification problem ?
  - a)  $\frac{1}{n} \sum_{i=1}^{i=n} 1_{f(X_i) \neq Y_i}$
  - b)  $\frac{1}{n} \sum_{i=1}^{i=n} |f(X_i) - Y_i|$
  - c)  $E[|f(X_{n+1}) - Y_{n+1}|]$
  - d) none of the answers above
  
2. If  $(X_i, Y_i)_{i \in \llbracket 1, n \rrbracket}$  is a learning sample, which of the hypothesis always made on the variables ?
  - a) the  $X_i$  are all Gaussian
  - b) the  $(X_i, Y_i)$  have all the same laws
  - c) the  $X_i$  have all the same laws but not necessarily the  $Y_i$
  - d) the  $(X_i, Y_i)$  are all independent
  - e) none of the answers above
  
3. If  $R_n(f_n)$  is the calibration error for the optimal classifier  $f_n$  obtained for the learning sample  $(X_i, Y_i)_{i \in \llbracket 1, n \rrbracket}$  and if we note  $R(f_n) = E[1_{f_n(X_{n+1}) \neq Y_{n+1}}]$  the prediction error, which properties are always verified ?
  - a)  $R_n(f_n) = R(f_n)$
  - b)  $R_n(f_n) \leq R(f_n)$
  - c)  $E(R_n(f_n)) \leq R(f_n)$
  - d) none of the answers above
  
4. If there are  $k$  particular points in  $\mathbb{R}^d$  which can be classified in all possible ways by the family of classifiers  $\mathcal{F}_1$  and if it is not possible to classify them in all possible ways by the family of classifiers  $\mathcal{F}_2$  then which assertions is/are necessarily true :
  - a)  $VC(\mathcal{F}_1) = k$
  - b)  $VC(\mathcal{F}_2) < k$
  - c)  $VC(\mathcal{F}_1) > VC(\mathcal{F}_2)$
  - d) none of the answers above
  
5. If for two families of classifiers/machines  $\mathcal{F}_1$  and  $\mathcal{F}_2$  the error of calibration is the same on a learning sample, which machine does the Vapnik Chernovenkis inequality and SRM principle encourage to use to predict :

---

1. Pierre Brugière University Paris 9 Dauphine

- a)  $\mathcal{F}_1$  if  $VC(\mathcal{F}_1) > VC(\mathcal{F}_2)$   
 b) not  $\mathcal{F}_2$  if  $VC(\mathcal{F}_1) > VC(\mathcal{F}_2)$   
 c) none of the answers above
6. If  $x_1, x_2 \cdots x_n$  form a family of independent vectors, and if 0 is the null vector, in how many different ways is it possible to classify :  $0, x_1, x_2 \cdots x_n$  ?  
 a)  $2^n$   
 b)  $n - 1$   
 c)  $2^{n+1}$   
 d) none of the answers above
7. The inequality of Vapnik Chervonenkis enables, knowing the complexity of the family of classifiers/machine used and the error on calibration to :  
 a) define a confidence interval for  $R(f_n)$   
 b) calculate the exact value of  $R(f_n)$   
 c) define a boundary  $\epsilon < 1$  for  $R(f_n)$   
 d) none of the answers above
8. Which assertions is/are true ?  
 a) a high  $VC$  for a family of classifiers implies necessarily a bad quality of prediction  
 b) an infinite  $VC$  for a family of classifiers implies always a perfect calibration  
 c) the  $VC$  for a parametric family of classifiers is always close to the number of parameters of the family of classifiers  
 d) none of the answers above
9. Which assertions are true  
 a) It is unlikely that a machine which calibrates badly will predict accurately  
 b)  $VC$  gives no guarantee that a complex machine which calibrates well will predict accurately  
 c) SRM means Structural Risk Maximization  
 d) a machine which is very complex in a  $VC$  sense predicts always very accurately
10. In  $\mathbb{R}^d$  for  $w \neq 0$  let  $H_{w,b} = \{x \in \mathbb{R}^d, \langle w, x \rangle + b = 0\}$ . Which assertions are true :  
 a)  $\forall x \in \mathbb{R}^d d(x, H_{w,b}) = \frac{|\langle w, x \rangle + b|}{\|w\|^2}$   
 b)  $\forall x \in \mathbb{R}^d d(x, H_{w,b}) = \frac{|\langle w, x \rangle + b|}{\|w\|}$   
 c)  $\forall x \in \mathbb{R}^d d(x, H_{w,b}) = \frac{|\langle w, x \rangle - b|}{\|w\|}$   
 d) none of the answers above

11. Let  $w \in \mathbb{R}^d \setminus \{0\}$ . Let  $H_{w,b} = \{(x, y) \in \mathbb{R}^{d+1}, \langle w, x \rangle + b - y = 0\}$ . Which assertions are true :
- $d(H_{w,b_1}, H_{w,b_2}) = \frac{|b_2+b_1|}{\|w\|_d}$
  - $d(H_{w,b_1}, H_{w,b_2}) = \frac{|b_2-b_1|}{\sqrt{1+\|w\|_d^2}}$
  - $d(H_{w,b_1}, H_{w,b_2}) = \frac{|b_2-b_1|}{\|w\|_d}$
  - none of the answers above
12. If you can classify perfectly a learning sample with the hyperplanes of margin  $\Delta_1$  around  $H_{w_1,b_1} = \{x \in \mathbb{R}^d, \langle w_1, x \rangle + b_1 = 0\}$  or of margin  $\Delta_2$  around  $H_{w_2,b_2} = \{x \in \mathbb{R}^d, \langle w_2, x \rangle + b_2 = 0\}$  then you have a good reason to choose  $H_{w_1,b_1}$  if :
- $\Delta_1 < \Delta_2$
  - $\Delta_1 > \Delta_2$
  - $b_1 > b_2$
  - $b_2 > b_1$
13. If you can classify perfectly a learning sample with two classes with an hyperplane  $H$  of margin  $\Delta$  and if we call  $\mathcal{C}_1$  and  $\mathcal{C}_2$  the convex envelopes of the two classes of points then :
- for sure  $d(\mathcal{C}_1, \mathcal{C}_2) \geq \Delta$
  - for sure  $d(\mathcal{C}_1, \mathcal{C}_2) > \Delta$
  - $H$  of maximum margin  $\implies (H \cap \mathcal{C}_1 \neq \emptyset \text{ and } H \cap \mathcal{C}_2 \neq \emptyset)$
  - none of the answers above
14. If  $\mathcal{F}$  is a family of classifiers of  $\mathbb{R}^{10^6}$  of diameters 1 of hyperplanes of margin 0.1 then :
- $VC(\mathcal{F}) = 10^6$
  - $VC(\mathcal{F}) = 10^6 + 1$
  - $VC(\mathcal{F}) \approx 100$
  - $VC(\mathcal{F}) \approx 10$
15. According to the minimax theorem for any function  $g : \mathcal{Y} \times \mathcal{Z} \longrightarrow \mathbb{R}$
- $\inf_{y \in \mathcal{Y}} \left[ \sup_{z \in \mathcal{Z}} g(y, z) \right] \leq \sup_{z \in \mathcal{Z}} \left[ \inf_{y \in \mathcal{Y}} g(y, z) \right]$
  - $\sup_{z \in \mathcal{Z}} \left[ \inf_{y \in \mathcal{Y}} g(y, z) \right] \leq \inf_{y \in \mathcal{Y}} \left[ \sup_{z \in \mathcal{Z}} g(y, z) \right]$
  - $\sup_{z \in \mathcal{Z}} \left[ \inf_{y \in \mathcal{Y}} g(y, z) \right] = \inf_{y \in \mathcal{Y}} \left[ \sup_{z \in \mathcal{Z}} g(y, z) \right]$
16. When solving a SVM for a classification  $\{-1, 1\}$  which assertions are true :
- The Primal and Dual problems have the same solution if and only if the KKT conditions are added to the constraints of the dual problem
  - The Primal and Dual problems have the same solution because of the

particular nature of the problem

c) The KKT conditions are automatically satisfied because of the particular nature of the problem

d) none of the answers above.

17. When solving a C-SVM for a classification  $\{-1, 1\}$  which assertions are true :

a) all the support vectors are necessarily classified correctly

b) some support vectors may be classified incorrectly

c) the support vectors are necessarily on the border of the maximum margin hyperplane

d) if  $0 < \alpha_i < C$ ,  $x_i$  is a support vector on the margin of the maximum margin hyperplane.

18. Among the following functions from  $\mathbb{R}^d \times \mathbb{R}^d$  to  $\mathbb{R}$  which ones are Kernel

a)  $\exp(-\frac{\|x-y\|^2}{2}) + \exp(-\frac{\|x-y\|^2}{4})$  (because sum of Kernels implies strictly positive as well)

b)  $\exp(-\|x\| - \|y\|)$  (criteria of strict positiveness easy to verify)

c)  $\langle x, y \rangle$  (because of the form  $\langle \phi(x), \phi(y) \rangle$ ).

19. If  $\phi(\cdot)$  is the transformation associated to the kernel  $K(x, y) = \exp(-\frac{\|x-y\|^2}{2})$  by the relationship  $\langle \phi(x), \phi(y) \rangle = K(x, y)$  among these properties which ones are true for the image points  $\phi(x_i)$  from a learning sample.

a) the  $\phi(x_i)$  are necessarily on a sphere of radius 1

b) the  $\phi(x_i)$  for sure can be separated from 0 by an hyperplane

c) the  $\phi(x_i)$  for sure are independent if the  $x_i$  are distinct

20. If  $K(\cdot, \cdot)$  is a kernel and  $(H, \langle \cdot, \cdot \rangle_{RK})$  is a Reproducing Hilbert Space for  $K$  then which of the following assertions are true :

a) for all  $f \in H$ ,  $\langle K(x, \cdot), f \rangle_{RK} = f(x)$

b)  $K(x_1, x_2)^2 \leq K(x_1, x_1)K(x_1, x_2)$

c) for all  $f \in H$ ,  $f(x)^2 \leq K(x, x)\langle f, f \rangle_{RK}$

### Exercise : [13pts]

Let  $\{(x_i, y_i)\}_{i \in \llbracket 1, n \rrbracket}$  be a learning sample with  $x_i \in \mathbb{R}^d$  and  $y_i \in \{-1, 1\}$ .

Let  $I^+ = \{i \in \llbracket 1, n \rrbracket, y_i = 1\}$  and  $I^- = \{i \in \llbracket 1, n \rrbracket, y_i = -1\}$ . We assume that neither  $I^+$  nor  $I^-$  are empty and we consider the problem :

$$(P) : \begin{cases} \arg \inf_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

where  $C$  is a constant which is strictly positive.

1. [2pt] Show that the inf of  $(P)$  needs to be searched only in a bounded region.

**Correction :**

For  $w, b$  equal to zero and all the components of  $\xi$  equal to 1 all the conditions are satisfied and the value of the function is  $nC$  so there is no need to search the inf in a domain where the function exceeds  $nC$ . Therefore we can limit the search to the region where  $\|w\| \leq \sqrt{2nC}$ .

For  $\xi$  the search can be limited to the domain for which all the  $\xi_i$  satisfy  $|\xi_i| \leq n$  otherwise the function would exceed  $nC$ . If now we consider an index  $i$  for which  $y_i = 1$  the condition  $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$  implies that  $b \geq 1 - \xi_i - \langle w, x_i \rangle$  and therefore gives the lower bound  $1 - n - \|x_i\| \sqrt{2nC}$  for  $b$ . Considering now an index  $i$  for which  $y_i = -1$  we get in the same way an upper bound for  $b$ . So at the end the inf needs to be searched only within a bounded region for the parameters  $w, b$  and  $\xi$  Q.E.D.

2. [0.5pt] Explain why the inf is reached (and so in fact is a min).

**Correction :**

The function to minimize is continuous and the search of the inf is on a bounded and thus compact region, therefore the inf is reached.

3. [2pt] Show that (P) is equivalent to

$$(Q) : \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \max[0, 1 - y_i(\langle w, x_i \rangle + b)]$$

**Correction :**

Let  $w^*, b^*, \xi^*$  be a solution of (P). We remark that :

$\xi_i^* > 0 \implies 1 - y_i(\langle w, x_i \rangle + b) > 0$  because otherwise  $\xi_i^{**} = 0$  would satisfy

$\xi_i^{**} \geq 1 - y_i(\langle w, x_i \rangle + b)$  and would enable to reach a better solution. Also,

if  $1 - y_i(\langle w, x_i \rangle + b) > 0$  then not only  $\xi_i^* \geq 1 - y_i(\langle w, x_i \rangle + b)$  but in fact

$\xi_i^* = 1 - y_i(\langle w, x_i \rangle + b)$  because otherwise  $\xi_i^{**} = 1 - y_i(\langle w, x_i \rangle + b)$  would satisfy all the conditions and would enable to build a better solution (as  $\xi_i^{**} < \xi_i^*$ ). So the  $\xi_i^*$  solutions of the problem verify :

either  $\xi_i^* = 0$  or  $\xi_i^* = 1 - y_i(\langle w, x_i \rangle + b)$  if  $1 - y_i(\langle w, x_i \rangle + b) > 0$

so,  $\xi_i^* = \max[0, 1 - y_i(\langle w, x_i \rangle + b)]$ . As a consequence for the optimisation problem the variable  $\xi$  can be replaced by its value at the optimum, and thus we get the formulation (Q) of the problem.

Let  $l_h : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$  be the hinge loss function defined by :

$$l_h(y, z) = \max[0, 1 - yz].$$

4. [0.5pt] Show that :

$$(Q) \iff \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n l_h(y_i, \langle w, x_i \rangle + b)$$

and that :

$$\forall (y, z) \in \{-1, 1\} \times \mathbb{R}, 1 - yz = 0 \iff y = z.$$

**Correction : Obvious.**

Let  $\sigma : \mathbb{R} \rightarrow \mathbb{R}^+$  be the sigmoid function defined by

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

In a logistic model we assume that :  $P(Y_i = 1 | X_i = x_i) = \sigma(\langle w, x_i \rangle + a)$ .

5. [0.5pt] Show that in a logistic model :
- $$P(Y_i = -1 | X_i = x_i) = \sigma(-\langle w, x_i \rangle - a)$$

**Correction :**

$$\begin{aligned} P(Y_i = -1 | X_i = x_i) &= 1 - P(Y_i = 1 | X_i = x_i) \\ &= \frac{\exp(-\langle w, x_i \rangle - a)}{1 + \exp(-\langle w, x_i \rangle - a)} = \frac{1}{1 + \exp(\langle w, x_i \rangle + a)} = \sigma(-\langle w, x_i \rangle - a) \end{aligned}$$

6. [1pt] Show that in a logistic model the log likelihood to maximise is :

$$L(w, a) = - \sum_{i=1}^{i=n} \ln(1 + \exp(-y_i[\langle w, x_i \rangle + a]))$$

**Correction :**

According to the previous questions

$$P(Y_i = y_i | X_i = x_i) = \sigma(y_i[\langle w, x_i \rangle + a])$$

therefore the log likelihood for the sample is :

$$L(w, a) = \sum_{i=1}^{i=n} \ln(\sigma(y_i[\langle w, x_i \rangle + a])) = \sum_{i=1}^{i=n} -\ln(1 + \exp(-y_i[\langle w, x_i \rangle + a]))$$

7. [2pt] Show that if  $(x_i)_{i \in I^+}$  and  $(x_i)_{i \in I^-}$  are strictly separable by an hyperplane then  $\sup_{w \in \mathbb{R}^d, a \in \mathbb{R}} L(w, a) = 0$ .

**Correction :**

First we remark that  $\forall (w, a) \in \mathbb{R}^{d+1}$  each term of the sum is strictly negative and so  $L(w, a) < 0$ . ( $L(w, a)$  is the logarithm of a probability in  $]0, 1[$ ) Now, if the points are strictly separable then it is possible to find  $w^*, a^*$  such that :  $\forall i \in \llbracket 1, n \rrbracket, y_i[\langle w^*, x_i \rangle + a^*] > 0$  and then for each  $i \in \llbracket 1, n \rrbracket, \lim_{\lambda \rightarrow +\infty} \exp(-\lambda y_i[\langle w^*, x_i \rangle + a^*]) = 0$  and therefore

$$\lim_{\lambda \rightarrow +\infty} L(\lambda w^*, \lambda a^*) = 0.$$

8. [1pt] Show that if  $(x_i)_{i \in I^+}$  and  $(x_i)_{i \in I^-}$  are not strictly separable by an hyperplane then  $\sup_{w \in \mathbb{R}^d, a \in \mathbb{R}} L(w, a) < 0$ .

**Correction :**

In this case  $\forall (w, a) \in \mathbb{R}^{d+1} \exists i \in \llbracket 1, n \rrbracket$  such that  $y_i[\langle w, x_i \rangle + a] < 0$ .

For such  $i$  we have  $\exp(-y_i[\langle w, x_i \rangle + a]) > 1$  and as all the terms in the sum are strictly negative :  $L(w, a) < -\ln(2)$ . So if the points are not separable :  $\sup_{w \in \mathbb{R}^d, a \in \mathbb{R}} L(w, a) < -\ln(2)$ .

9. [2pt] For  $\lambda > 0$  let  $L_\lambda(w, a) = L(w, a) - \frac{\lambda}{2}\|w\|^2$ .  
Using the fact that neither  $I^+$  nor  $I^-$  are empty show that :  
 $\sup_{w \in \mathbb{R}^d, a \in \mathbb{R}} L_\lambda(w, a)$  is reached.

**Correction :**

$L_\lambda(0, 0) = -n \ln(2)$  so we search the sup in a domain such that :  
 $\frac{\lambda}{2}\|w\|^2 < n \ln(2)$  and

$$\forall i \in \llbracket 1, n \rrbracket, -\ln\left(1 + \exp(-y_i[\langle w^*, x_i \rangle + a^*])\right) > -n \ln(2).$$

This implies that :  $\forall i \in \llbracket 1, n \rrbracket, -y_i[\langle w^*, x_i \rangle + a^*] < \ln(\exp(n \ln(2)) - 1)$   
as  $\|w\|$  is bounded so are the  $\langle w^*, x_i \rangle$  and using the inequality for  $y_i$   
taking values 1 and  $-1$  show that  $a$  is bounded. So the sup of  $L_\lambda(w, a)$   
can be searched in a bounded region and as  $L_\lambda(w, a)$  is continuous the  
sup is reached.

Note that solving  $\sup_{w \in \mathbb{R}^d, b \in \mathbb{R}} L(w, b) - \frac{\lambda}{2}\|w\|^2$  is called a ridge logistic regression problem.

Let  $l_c : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be the logistic loss function defined by

$$l_c(y, z) = \ln(1 + \exp(-yz))$$

(which contrarily to some other loss functions does not vanish when  $y = z$ ).

10. [1.5pt] What is the relationship between a C-SVM classification problem and a ridge logistic regression ?

**Correction :** In a C-SVM with two classes of points we solve :

$$(Q) : \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^n \max[0, 1 - y_i(\langle w, x_i \rangle + b)]$$

$$\iff \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^n l_h(y_i, \langle w, x_i \rangle + b)$$

In a Ridge Logistic Regression problem we solve :

$$\arg \sup_{w \in \mathbb{R}^d, b \in \mathbb{R}} L(w, b) - \frac{\lambda}{2}\|w\|^2$$

$$\iff \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{\lambda}{2}\|w\|^2 - L(w, b)$$

$$\iff \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^n l_c(y_i, \langle w, x_i \rangle + b)$$

$$\iff \arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2} \|w\|^2 + \frac{1}{\lambda} \sum_{i=1}^n l_c(y_i, \langle w, x_i \rangle + b)$$

So both problems are the same written with two different loss function and with  $C = \frac{1}{\lambda}$ .