

# Introduction to Statistical Learning

## Mid-term exam

*Duration : 1h30 - No documents allowed*

### Reminder/Notations

- The indicator function  $\mathbb{I}\{\Omega\}$  takes the value 1 if  $\Omega$  is true, and 0 otherwise.
- If  $A$  denotes a set, then the notation  $|A|$  denotes the cardinality of  $A$ .
- Union bound :  $\mathbb{P}\{A \cup B\} \leq \mathbb{P}\{A\} + \mathbb{P}\{B\}$  where  $A$  and  $B$  are events.
- Law of iterated expectation :  $\mathbb{E}(U) = \mathbb{E}(\mathbb{E}(U | V))$  where  $U, V$  are random variables.
- The posterior probability in the plain binary classification model with observation vectors in  $\mathbb{R}^d$  and labels in  $\{0, 1\}$  is denoted by  $\eta(x) = \mathbb{P}\{Y = 1 | X = x\}$  for any  $x \in \mathbb{R}^d$ .
- For a real-valued decision rule  $h : \mathbb{R}^d \rightarrow \mathbb{R}$ , the ROC curve of  $h$  plots the true positive rate as a function of the false positive rate. Therefore, the ROC curve of  $h$  is the parametric curve given by :  
 $t \in [-\infty, \infty] \mapsto (\mathbb{P}\{h(X) > t | Y = 0\}, \mathbb{P}\{h(X) > t | Y = 1\})$ .
- Hoeffding's inequality - consider  $Z_1, \dots, Z_n$  independent and identically distributed random variables which take values in  $[0, 1]$  almost surely, then we have : for any  $t > 0$

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}(Z_1) \right| > t \right\} \leq 2 \exp(-2nt^2) .$$

**Exercise 1** - Consider  $(X, Y)$  a random pair that models classification data with labels in  $\{0, 1\}$ .

1. For a classifier  $g : \mathbb{R}^d \rightarrow \{0, 1\}$ , define  $L(g) = \mathbb{P}\{Y \neq g(X)\}$ . What is the minimizing argument  $g^*$  (called the Bayes classifier) of  $L(g)$  over all possible classifiers  $g$ ? What is the minimal value of  $L(g)$  over all possible classifiers  $g$ ?
2. Now define  $L_c(g) = c_0 \mathbb{P}\{Y \neq g(X), Y = 1\} + c_1 \mathbb{P}\{Y \neq g(X), Y = 0\}$  where  $c_0, c_1 > 0$ . What is the minimizing argument  $g_c^*$  of  $L_c(g)$  over all possible classifiers  $g$ ? What is the minimal value of  $L_c(g)$  over all possible classifiers  $g$ ?
3. Using the same notations as in the previous question, we set  $c_0 = 1$  and  $c_1 = \lambda$  where  $\lambda \in [0; +\infty]$ . After this reparameterization of the binary classification problem with asymmetric costs, we denote  $L_c$  by  $L_\lambda$  and we consider the sequence of binary classification problems  $\{\min_g L_\lambda(g) : \lambda \in [0; +\infty]\}$  with sequence of solutions  $\{g_\lambda^* : \lambda \in [0; +\infty]\}$ . Consider the decision rule  $h^*(x) = \int_0^\infty g_\lambda^*(x) d\lambda$ .
  - (a) What is the learning problem solved by  $h^*$ ?
  - (b) Which is the criterion that  $h^*$  optimizes? Give a proof of that fact.
  - (c) Is  $h^*$  the unique optimal element for the performance criterion?

4. We now consider classifiers with reject option  $g : \mathbb{R}^d \rightarrow \{R, 0, 1\}$ , and  $L_R(g) = \mathbb{P}\{Y \neq g(X), g(X) \neq R\} + c\mathbb{P}\{g(X) = R\}$ . What is the minimizing argument  $g_R^*$  of  $L_R(g)$  over all possible classifiers  $g$  with reject option? Give an interpretation of the result.

**Exercise 2** - We consider the model for classification data where  $X$  is a random vector on  $\mathbb{R}^d$  and  $Y$  is a random variable taking values in  $\{-1, +1\}$ .

1. We consider the following problems for which the question is to compute the optimal decision rule  $g^*$ ,  $h^*$  or  $f^*$ .
  - (a)  $R(g) = \mathbb{E}((Y - g(X))^2)$  where  $g : \mathbb{R}^d \rightarrow \{-1, +1\}$
  - (b)  $R(h) = \mathbb{E}((Y - h(X))^2)$  where  $h : \mathbb{R}^d \rightarrow \mathbb{R}$
  - (c)  $A(f) = \mathbb{E}(\log_2(1 + e^{-Yf(X)}))$  where  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{-\infty, +\infty\}$  and  $\log_2(u) = \log(u)/\log(2)$  for any  $u > 0$ .

Explain why such criteria are relevant for the binary classification problem.

2. We now consider the case of 1(c).
  - (a) Determine the function  $H$  such that  $A(f^*) = \mathbb{E}(H(\eta(X)))$ .
  - (b) Plot  $H$  and state its main properties. Compare  $(u - 1/2)^2$  and  $(1 - H(u))$ .
  - (c) Consider  $L(g) = \mathbb{P}\{Y \neq g(X)\}$  and  $L^* = \inf_g L(g)$ . What upper bound can be given on the quantity  $L(\text{sgn}(f)) - L^*$  in terms of  $A(f) - A(f^*)$ ?

**Exercise 3** - Consider the plain binary classification problem with a *finite* class  $\mathcal{G}$  of candidate decision rules from supervised data  $(X_1, Y_1), \dots, (X_n, Y_n)$  where the  $(X_i, Y_i)$ 's are independent random pairs over  $\mathbb{R}^d \times \{0, 1\}$ . We set the empirical classification error of a classifier  $g$  to be  $\widehat{L}_n(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{g(X_i) \neq Y_i\}$ . We want to provide guarantees over the true classification error  $L(g)$  for any  $g \in \mathcal{G}$ .

1. Give an upper bound on the deviation probability  $\mathbb{P}\{\max_{g \in \mathcal{G}} |L(g) - \widehat{L}_n(g)| > t\}$  for any  $t > 0$ .
2. Deduce an upper bound on  $L(g)$  that holds for any  $g \in \mathcal{G}$  with a probability at least of  $1 - \delta$ . The upper bound shall depend on  $\widehat{L}_n(g), \delta, \mathcal{G}, n$ .
3. Consider the algorithm that outputs  $\widehat{g}_n = \arg \min_{g \in \mathcal{G}} \widehat{L}_n(g)$ . Propose a meaningful upper bound of the difference  $L(\widehat{g}_n) - L(g^*)$  where  $g^*$  is the Bayes classifier. Comment on this inequality.