Loss and risk

Fabrice Rossi

Exercise 1

We consider a tiny data set with 10 observations $(\mathbf{x}_i,y_i)_{1\leq i\leq 10}$, with $\mathbf{x}_i \in \mathcal{X}$ et $y_i \in \{-1,1\}$. Using different machine learning algorithms, two models are built from this data set: g_1 and g_2. The outputs of the models on the learning data set are given by the following table:

$\mathbf{x}_i$	$g_1(\mathbf{x}_i)$	$g_2(\mathbf{x}_i)$	y_i
$\mathbf{x}_1$	1	1	1
$\mathbf{x}_2$	1	-1	1
$\mathbf{x}_3$	-1	1	1
$\mathbf{x}_4$	1	-1	1
$\mathbf{x}_5$	1	-1	1
$\mathbf{x}_6$	1	-1	-1
$\mathbf{x}_7$	-1	-1	-1
$\mathbf{x}_8$	-1	-1	-1
$\mathbf{x}_9$	-1	1	-1
$\mathbf{x}_{10}$	1	-1	-1

Question 1 Compute the confusion matrices of both models on the learning set.

Question 2 We use the loss function l_1 given by:

$l_1(p,t)$	$t=-1$	$t=1$
$p=-1$	0	1
$p=1$	3	0

where p is the predicted value and t the true value. Compute the empirical risk of both models on the learning set for l_1.

Question 3 Determine the best model based on the available information using the loss function $l_0(p,t) = \mathbf{1}_{p\neq t}$.

Exercise 2

In this exercise, we study a classification problem in which the target variable $\mathbf{Y}$ can take three different values in $\mathcal{Y} = \{A, B, C\}$. From a learning set $\mathcal{D}$, two models have been constructed g_1 and g_2. Their predictions on a new set $\mathcal{D}'$ are summarized by the following confusion matrices (we use the convention that the predicted values are in rows while the true values are in columns):

g_1

	A	B	C
A	44	0	0
B	5	62	1
C	1	8	54

g_2

	A	B	C
A	44	4	5
B	2	64	3
C	4	2	47

Question 1 Using the confusion matrices, compute an estimation of the distribution of $\mathbf{Y}$, i.e. of the probabilities $\mathbb{P}(\mathbf{Y} = \mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}$.

Question 2 What minimal consistency checks between $\mathcal{D}$ and $\mathcal{D}'$ should be done?

Question 3 Compute the accuracy of each model on $\mathcal{D}'$ (the accuracy is the percentage of correct classification).

Question 4 Determine the best model between g_1 and g_2 according to the loss function $l_0(p,t) = \mathbf{1}_{p \neq t}$ using the empirical risk on $\mathcal{D}'$.

Question 5 Is the selected model the best one according to the risk associated to l_0?

Question 6 We define a new loss function l_2 as follows:

$l_2(p,t)$		t		
		A	B	C
p	A	0	2	1
	B	1	0	1
	C	2	1	0

We use the convention that p is the predicted value and t the true value. Compute the empirical risk of each model according to this loss function on $\mathcal{D}'$.