FTML Exercices 3

Pour le 24 mars 2023

TABLE DES MATIÈRES

1 RISQUES DE BAYES

Retrouver les valeurs des risques de Bayes des deux parties de l'exercice 3 du premier TP. Pour chacune de ces loss, on connaît le prédicteur de Bayes comme vu en cours, et on peut ainsi calculer son risque.

Pour le premier exemple, le calcul ressemblera à ceux du 3e cours magistral. Pour le deuxième exemple (nombre de streams), il y a une espérance non triviale à calculer, mais il est possible de le faire numériquement.

2 ESTIMATEURS ET ESPÉRANCES

2.1 1

On se donne une variable aléatoire réelle X ayant un moment d'ordre 1. On note $\mathbb{E}[X]$ son espérance. On suppose qu'on dispose de n samples tirés de cette variable, $(x_1, \dots, x_n)$, qui suivent donc la loi de X.

On considère leur moyenne empirique $S_n = \sum_{i=1}^{n} x_i$. Montrer que l'espérance de S_n est $\mathbb{E}[X]$. Cela signifie que S_n est un estimateur non biaisé de $\mathbb{E}[X]$.

Simuler une variable de votre choix pour observer la convergence de S_n vers $\mathbb{E}[X]$.

2.2 2

On se donne un problème d'apprentissage supervisé usuel. On se donne un prédicteur f, fixé et indépendant du dataset. Montrer que l'espérance du risque empirique de f est le risque réel de f. autrement dit : le risque empirique de f est un estimateur non biaisé de son risque réel.

La notion d'espérance a bien un sens ici car il faut voir le dataset comme une variable aléatoire, ainsi le risque empirique de f est bien une variable aléatoire.

2.2.1 *Notations*

Si vous avez besoin de formalisation, vous pouvez prendre les notations suivantes :

— X is a random variable from an input space $\mathcal{X}$

— Y is a random variable from an output space $\mathcal{Y}$

— $(X, Y) \sim \rho$: this means that ρ is the law of the joint random variable.
https://en.wikipedia.org/wiki/Joint_probability_distribution

— The dataset D_n is a collection of n samples $\{(x_i, y_i)\}_{1 \leqslant i \leqslant n}$, that are assumed **independent and identically distributed** draws of the joint random variable (X, Y).

— Let l be a loss function.

The **risk** (or **statistical risk, generalization error, test error)** of estimator f writes

$$R(f) = E_{(X,Y)\sim\rho}[l(Y, f(X)] \tag{1}$$

The **empirical risk (ER)** of an estimator f writes

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i)) \tag{2}$$

2.2.2 *Simulation*

Simuler un problème de votre choix pour observer le résultat (c'est exactement ce qui est fait dans **simulations/lecture_3/empirical_risk_convergence**)