Apprentissage Statistique

Exercice 1. *Il y a une chance de 0.1% qu'un patient ait une certaine maladie. Le test sur cette maladie a une précision de 90% pour des résultats positifs de test (c'est-à-dire, P (test positif | a la maladie) = 0.9) et une précision de 80% pour des résultats négatifs de test (c'est-à-dire, P (test négatif | n'a pas de maladie) = 0.8). Quelle est la probabilité que le patient ait la maladie sachant qu'il a été testé positif?*

Exercice 2. *On considère deux urnes U_1 et U_2 contenant :*

- *deux billes bleues et une bille rouge pour U_1, et,*
- *deux billes rouges et une bille bleue pour U_2.*

Les billes ne diffèrent entres elles que par leur couleur. On choisit une bille au hasard de la façon suivante : on lance une pièce non truquée ; si on obtient pile on choisit une bille dans l'urne U_1 ; sinon on choisit une bille dans l'urne U_2. Le tirage dans l'urne est supposé uniforme. On considère le couple de variables aléatoires (U, C) oú U désigne l'urne choisie et C la couleur de la bille. Du point de vue apprentissage, l'objectif est de " prédire " U en connaissant la couleur de la bille obtenue.

Questions :

1. *Donner la loi (marginale) de U*
2. *Donner la loi (marginale) de C*
3. *Calculer $P(U = U_1|C = rouge)$ et $P(U = U_1|C = bleue)$*
4. *En déduire le meilleur classifieur possible au sens de l'erreur locale définie par le tableau suivant*

$$\ell(x, y) = \left\{ \begin{array}{cc|cc} & y & U_1 & U_2 \\ \hline x & U_1 & 0 & 1 \\ & U_2 & 1 & 0 \end{array} \right. \tag{1}$$

5. *Calculer le risque du classifieur optimal.*

Exercice 3. *On observe maintenant les tirages sous forme de réalisation de copies i.i.d. de (U,C), plus précisément le tableau suivant :*

U	*C*
1	rouge
1	bleue
2	bleue
1	bleue
2	rouge
2	rouge
1	bleue
2	bleue
1	rouge
1	bleue
1	bleue
2	rouge

Ce sont les donnèes d'apprentissage.

Questions :

1. *Estimer les lois marginales de U et C d'après le tableau.*
2. *Estimer $P(U = U_1|C = rouge)$ et $P(U = U_1|C = bleue)$ d'après le tableau.*
3. *En déduire le classifieur empirique optimal.*
4. *Calculer l'erreur empirique du classifieur sur les donnèes d'apprentissage avec l'erreur locale ℓ définie par (1).*

Exercice 4. *SVM linéaire.*
On considère les données d'entraînement suivants de 2 classes:

$$\textit{Classe 1 : } \{(1,1)'\} \textit{ et Classe 2 : } \{(-1,-1)', (1,0)', (0,1)'\}$$

1. *Tracer ces quatre points et la frontière de séparation linéaire pour laquelle SVM donnerait pour ces données et lister les vecteurs de support.*
2. *Sachant que l'équation d'une droite (l'hyper-plan plus généralement) a la forme $w'x + b = 0$, où x est un point de test, w est un vecteur de poids et b est un scalaire. Ecrivez l'équation de l'hyper-plan optimal que vous avez obtenu à la question a). C'est-à-dire par inspection de tracé obtenu à la question a) spécifier le vecteur de poids w et le scalaire b qui corresponde à la droite optimale séparant les classes.*

Exercice 5. *SVM non- linéaire.*
On considère les données d'entraînement suivants de 2 classes unidimensionnelles:

$$\textit{Classe 1 : } \{-5, 5\} \textit{ et Classe 2 : } \{-2, 1\}$$

1. *Tracer ces points. Sont-ils linéairement séparables ?*
2. *Soit la transformation $f : \mathbb{R} \to \mathbb{R}^2$, définie par $f(x) = (x, x^2)$. Transformez les données et tracez les points transformés. Est-ce que ceux-ci sont linéairement séparables ?*
3. *Ecrivez l'équation de l'hyper-plan optimal de séparation.*
4. *Ce l'hyper-plan optimal de séparation, correspond à une frontière de séparation non-linéaire dans l'espace d'origine ?*