Apprentissage Statistique

1. Ce document contient une liste des exercices et des problèmes qui avec les exercices du TD vous permettra de bien préparer votre examen. Pour la résolution de ces exercices vous aurez besoin de vous rappeller certaines notions de probabilités et statistiques: distribution de densités de probabilités usuelles, Règle de Bayes, Maximum de Vraisemblance, Maximum A Posteriori, etc. Au moment de l'examen vous aurez accès aux documents du cours, mais pas aux notes corrigées de cette liste des exercices !
2. Prévoir aussi des questions de compréhension du cours !

Exercice 1. *Nous jetons deux fois un dé idéal, dont les faces sont numérotées de 1 à 6. Quelle est la probabilité qu'il y aura au moins 5 ?*

Exercice 2. *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 aété testé positif?*

Exercice 3. *Distribution Uniforme.*
Soit X un ensemble de variables aléatoires d'une distribution Uniforme avec paramètres a et b, X $\sim$ *Uniforme* $(x|a,b)$*, où* $a = 0$ *et* $b = \frac{1}{2}$.

1. *Calculez la fonction de densité de probabilité (pdf) de X.*
2. *Calculez* $p(X = 0.00027|a,b)$.
3. *Calculez* $Pr(X = 0.00027|a,b)$ *(la probabilité que X = 0.00027) .*

Exercice 4. *Vous devez être testé pour une maladie qui a la fréquence dans la population de 1 à 1000. Le test de laboratoire utilisé n'est pas toujours parfait : Il a un taux faux-positif de* 1%. *[Un résultat faux-positif est quand le test est positif, bien que la maladie ne soit pas présente.] le taux négatif faux du test est zéro. [Un négatif faux est quand le résultat de test est négatif quand en fait la maladie est présente.]*

1. *Si vous êtes testé et vous obtenez un résultat positif, quelle est la probabilité que vous ayez en réalité la maladie ?*
2. *Dans ces conditions, est-ce qu'il est plus probable que vous ayez la maladie ou non ?*
3. *Les réponses à a) et / ou b) seraient diffèrentes si vous utilisez la vrasemblance maximale plutôt que méthode d'évaluation basée sur le maximum a posteriori ? Commentez votre réponse.*

Exercice 5. *On suppose que nous avons un ensemble de données a décrit les trois variables suivantes: Cheveux = B, D, où B=blond, D=dark. Hauteur = T, S, où T=tall, S=short. Pays =*

G, P, où G=Greenland, P=Poland. On vous donne l'ensemble de données d'entraînement suivant (Cheveux, Hauteur, Pays):

$$\begin{pmatrix} (B,T,G) & (B,T,G) & (B,T,P) \\ (D,T,G) & (D,T,G) & (B,T,P) \\ (D,T,G) & (D,T,G) & (B,T,P) \\ (D,T,G) & (D,T,G) & (D,T,P) \\ (B,T,G) & (B,T,G) & (D,T,P) \\ (B,S,G) & (B,S,G) & (D,S,P) \\ (B,S,G) & (B,S,G) & (B,S,P) \\ (D,S,G) & (D,S,G) & (D,S,P) \end{pmatrix}$$

Nous voulons répondre à la question suivante : si vous observez un nouvel individu grand ayant cheveux blonds, quel est son pays d'origine le plus probable?

1. *Donner la solution maximum a posteriori (MAP) à la question, en utilisant le cadre naïf de Bayes. Détaillez votre travail.*

2. *Donner la solution maximum a posteriori (MAP) à la question, sans utiliser le cadre naïf de Bayes. Détaillez votre travail.*

3. *Donner la solution de maximum de vraisemblance (MLE) à la question, en utilisant le cadre naïf de Bayes. Détaillez votre travail.*

4. *Donner la solution de maximum de vraisemblance (MLE) à la question, sans utiliser le cadre naïf de Bayes. Détaillez votre travail.*

5. *Parmi les méthodes ci-dessus laquelle a gagné votre confiance pour traiter les problèmes suivants et pourquoi ?*

   (a) *Un grand nombre d'exemples décrit par un petit nombre d'attributs.*

   (b) *Un petit nombre d'exemples décrit par un grand nombre d'attributs.*

   Indication : plus de paramètres nous devonsévaluer plus d'exemples nous avons besoin afin d'obtenir des estimation fiables.

6. *Expliquer comment vous résoudriez la même question si au lieu de blond/dark nous aurions une mesure estimée de la couleur des cheveux continue et au lieu de tall/short nous aurions la hauteur réelle en centimètres ?*

Exercice 6. *On considére le problème de classification à 2 classes. On suppose que $p(x|Class = k)$ suit une loi gaussienne multivariée $(k = 1, 2)$, avec le vecteur moyen m_k et la matrice de covariance S. On suppose que la matrice covariance est la même pour les deux classes. Dans ce cas, montrez que la frontière de décision du classificateur gaussien est l'hyper-plan linéaire (c'est-à-dire de la forme $w'x + b = 0$ avec un certain vecteur de poids w et un certain scalaire b). (Ici la notation w' signifie le transposé de w.)*

Exercice 7. *Dans la figure 1 ci-dessous, les chiffres représentent les valeurs prises par une fonction cible réelle. Calculez la valeur estimée de la fonction cible au point de test x_q par l'algorithme d'apprentissage 5-ppv.*

Exercice 8. *Dans la figure 2 ci-dessous, le $+$ et $-$ représentent les exemples positifs et négatifs d'une fonction cible booléenne dans un espace bidimensionnel des cas. Déterminez comment les algorithmes d'apprentissage $1, 3, 5$-ppv classifierait le nouveau cas x_q.*

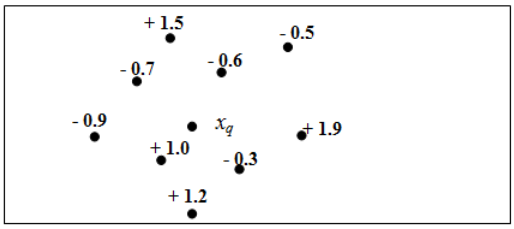

Figure 1: **Valeurs de la fonction cible**

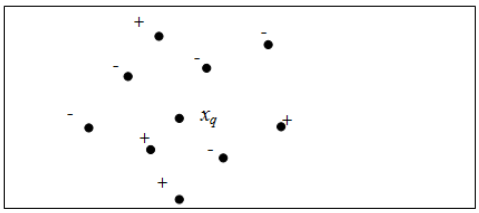

Figure 2: **Valeurs de la fonction cible**

Exercice 9. *Soit $x \in \mathbb{R}^p$ un vecteur. La loi normale p-dimensionnelle a la densité*

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)).$$

Onécrit $x = (x_1, x_2)'$ et $\mu = (\mu_1, \mu_2)'$ de dimension $(q \times 1, (p-q) \times 1)'$. On pose

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

de dimension

$$\begin{pmatrix} q \times q & q \times (p-q) \\ (p-q) \times q & (p-q) \times (p-q) \end{pmatrix}$$

1. *Caculez les distributions marginales et jointe. Montrez:*

$$x_1 \sim N(\mu_1, \Sigma_{11}), x_2 \sim N(\mu_2, \Sigma_{22}), x \sim N(\mu, \Sigma).$$

2. *Caculez la distribution conditionnelle. Montrez:*

$$x_1|x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

(Ici la notation w' signifie le transposé de w.)

Exercice 10. *On suppose que dans un problème la distance n'a pas ete explicitement spécifiée. A la place , on vous donne une boîte noire oùvous saisissez un ensemble de données $x_1, x_2, ..., x_n$ et un nouveau point de test x Les sorties de la boîte noires en utilisant l'algorithme $1-$ plus proche voisin de x, donnent x_i et sonétiquette de classe correspondante Y_i. Est cela possible de construire un algorithme de classification $k-ppv$, où $k > 1$, basée sur cette boîte noire seule ? Si la réponse est positive , quelle est la procédure ? Si la réponse est négative, pourquoi ?*

Exercice 11. *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 12. *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 ?*

Exercice 13. *On considère un problème de classification mixte où chaque observation est décrite par une variable discrète D à valeurs dans $\{0,1\}$ et une variable continue C à valeurs dans $\mathbb{R}$. La classe de chaque observation est donnée par la variable Y à valeurs dans $\{0,1\}$.*

1. *Ecrire la vraisemblance d'une observation (d,c,y) en notant $p(C|D,Y,\theta)$ la densité conditionnelle de C sachant D et Y, où θ désigne un vecteur de paramètres pour la densité conditionnelle.*

2. *Donner la forme simplifiée de la vraisemblance quand on fait l'hypothèse du classifieur bayésien naïf (que l'on maintiendra à partir de cette question).*

3. *On se donne N observations, $(d_i, c_i, y_i)_{1 \leq i \leq N}$ supposées i.i.d. Déterminer l'estimation de $P(D = 1|Y = y)$ (pour $y \in \{0,1\}$) par maximum de vraisemblance des N observations.*

4. *On suppose maintenant que la distribution de C sachant Y est gaussienne. Déterminer l'estimation des paramètres des gaussiennes par maximisation de la vraisemblance des N observations.*

Exercice 14. *On suppose que X est une variable de Bernoulli de paramètre θ (soit $P(X = 1) = \theta$). On se donne N répliques i.i.d. de X, $X_1, \ldots, X_N$.*

1. *Point de vue fréquentiste*

   (a) *Donner la vraisemblance des N répliques et en déduire l'estimation de θ par maximum de vraisemblance.*

   (b) *Quelle valeur prend l'estimateur ci-dessus quand on obtient 4 fois 1 pour $N = 4$?*

2. *Point de vue de Bayes*

   Dans l'approche Bayesienne, on considère un modèle plus complexe où on choisit θ aléatoirement, puis où on observe N variables de Bernoulli du paramètre θ. On a donc $P(X_i = 1|\Theta = \theta) = \theta$. Pour simplifier les calculs, on choisit ici pour Θ une loi Beta , c'est- à-dire

$$p(\Theta = \theta|a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{Beta(a, b)}.$$

   On rappelle que la loi $Beta(a, b)$ est d'espérance $\frac{a}{a+b}$, et de mode $\frac{a-1}{a+b-2}$.

   (a) *On note $\mathcal{D} = (X_1, \ldots, X_N)$ avec*

$$P(\mathcal{D} = (x_1, \ldots, x_N)|\Theta = \theta) = \prod_{i=1}^{N} P(X_i = x_i|\Theta = \theta)$$

   Calculer $p(\Theta = \theta|\mathcal{D} = (x_1, \ldots, x_N))$.

   (b) *Déduire de l'expression précédente l'estimation de θ par maximum à postériori, c'est-à-dire le mode de $p(\Theta = \theta|\mathcal{D} = (x_1, \ldots, x_N))$.*

   (c) *On tire une nouvelle valeur X_{N+1} selon la même loi (et donc selon le même θ). Donner $P(X_{N+1} = 1|\mathcal{D} = (x_1, \ldots, x_N))$.*

Exercice 15. *On considère un problème de classification binaire (variable Y à valeurs dans $\{0, 1\}$) où chaque observation est décrite par p variables binaires, $X = (X_1, \ldots, X_p)$, supposées conditionnellement indépendantes sachant la classe. On a donc $2p$ paramètres $\theta_1^1, ..., \theta_p^1$ et $\theta_1^0, ..., \theta_p^0$, avec $P(X_i = 1|Y = y) = \theta_i^y$.*

On choisit la distribution à priori $Beta(a, b)$ pour tous les θ_i^y. On suppose $P(Y = 1) = \frac{1}{2}$ et on se donne un ensemble d'apprentissage $\mathcal{D} = ((X_1, Y_1), \ldots, (X_N, Y_N))$

1. *Donner l'estimateur du maximum à posteriori pour les $2p$ paramètres.*

2. *Donner $P(Y = 1|X, \mathcal{D})$.*

Exercice 16.

On étudie un ensemble de 10 observations, $(x_i,y_i)_{1\leq i\leq 10}$, avec $x_i \in \mathcal{X}$ et $y_i \in \{-1,1\}$. Grâce à un algorithme d'apprentissage automatique, on construit deux modèles, g_1 et g_2. Le tableau suivant donne les valeurs de $g_1(x_i)$, $g_2(x_i)$ et y_i pour tout i :

x_i	$g_1(x_i)$	$g_2(x_i)$	y_i
x_1	1	1	1
x_2	1	-1	1
x_3	-1	1	1
x_4	1	1	1
x_5	1	-1	1
x_6	1	-1	-1
x_7	-1	-1	-1
x_8	-1	-1	-1
x_9	-1	1	-1
x_{10}	1	-1	-1

Question 1 Calculer les matrices de confusion des deux modèles.

Question 2 On choisit la fonction de perte l_1 définie par :

$l_1(v,p)$	$p=-1$	$p=1$
$v=-1$	0	2
$v=1$	1	0

où p désigne la valeur prédite et v la vraie valeur. Déterminer le meilleur modèle (entre g_1 et g_2) au sens du risque empirique construit à partir de l_1.

Question 3 Quel modèle choisir au sens de la fonction de perte $l_0(v,p) = \mathbb{I}_{p\neq v}$?

Exercice 17.

On étudie des données distribuées selon le modèle $Z = (X,Y)$ suivant :

- Y est une variable aléatoire à valeurs dans $\{-1,1\}$ avec $\mathbb{P}(Y=-1) = \frac{2}{3}$;
- X est une variable aléatoire à valeurs dans $\{a,b,c\}$ dont la loi conditionnelle est donnée par :

x	a	b	c
$\mathbb{P}(X=x\|Y=-1)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
$\mathbb{P}(X=x\|Y=1)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

Question 1 Déterminer le modèle optimal de Y sous la forme d'une fonction de X au sens de la fonction de perte l_0 de l'exercice précédent.

Question 2 Calculer le risque du modèle optimal obtenu à la question précédente.

Question 3 Mêmes questions avec la fonction de perte l_1 de l'exercice précédent.