2.8 Séance de TP sur le risque de Bayes

Remarques générales
Cette séance de travaux pratiques (TP) est l'occasion de mettre en œuvre les théories étudiées en cours. Il est donc indispensable d'avoir une bonne connaissance du cours pour analyser les résultats des expériences menées pendant la séance. Le sujet abordé pour ce quatrième TP porte sur l'application du risque de Bayes. Les étudiants doivent construire leur raisonnement en s'appuyant sur une démarche expérimentale. Pour les aider dans cette démarche, ils sont encouragés à poser des questions aux enseignants pendant la séance. Ce TP donne lieu à un compte rendu écrit d'une page A4 (recto uniquement, taille des polices 11) qui est remis avant le jour suivant la séance de TP au format pdf. Le rôle de ce compte rendu est de décrire les quelques points qui vous paraissent essentiels et non de faire une liste exhaustive des points abordés en TP. Chaque groupe travaille à son rythme et rédige son compte rendu en autonomie.

Risque de Bayes
On considère un problème à C classes $\{\omega_1, ..., \omega_C\}$ pour lequel les vecteurs de caractéristiques $\boldsymbol{x} \in \mathbb{R}^N$ sont distribués suivant des lois normales :

$$P(\boldsymbol{x}|\omega_c) = f(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Gamma}_c) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Gamma}_c|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^T \boldsymbol{\Gamma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right] \quad \forall c \in \{1, .., C\} \tag{59}$$

Au précédent TP, en supposant les classes équiréparties, on a analysé le discriminateur

$$\widehat{d}_{\mu,\Gamma \text{ connus}}(\boldsymbol{x}) = \arg \max_{c\in\{1,...,C\}} (P(\boldsymbol{x}|\omega_c)) \tag{60}$$

Pour ce TP, on utilise comme critère le risque de Bayes. On introduit le prior $P(\omega_c)$, la vraisemblance $P(\boldsymbol{x}|\omega_c)$, la densité a posteriori $P(\omega_c|\boldsymbol{x})$ et la loi du couple $P(\omega_c, \boldsymbol{x})$. On peut noter que $P(\boldsymbol{x}|\omega_c) = f(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Gamma}_c)$. Pour une partition de l'espace $(R_i)_{i=1,...,C}$ donnée, le risque de Bayes est défini par

$$\mathcal{R} = \sum_{i=1}^{C}\sum_{j=1}^{C} \int_{R_i} \alpha_{ij} P(\omega_j, \boldsymbol{x})\, d\boldsymbol{x} \tag{61}$$

où α_{ij} est le coût associé à la décision la classe de $\boldsymbol{x}$ est ω_i alors que la vraie classe de $\boldsymbol{x}$ est ω_j.
Avec la relation $P(\omega_j, \boldsymbol{x}) = P(\omega_j|\boldsymbol{x})P(\boldsymbol{x})$, on obtient

$$\mathcal{R} = \sum_{i=1}^{C} \int_{R_i} \rho_i(\boldsymbol{x})P(\boldsymbol{x})\, d\boldsymbol{x} \quad \text{où} \quad \rho_i(\boldsymbol{x}) = \sum_{j=1}^{C} \alpha_{ij} P(\omega_j|\boldsymbol{x}) \tag{62}$$

Le discriminateur qui minimise le risque de Bayes est donc défini par

$$\widehat{d}_{Bayes}(\boldsymbol{x}) = \arg \min_{i=1,...,C} \rho_i(\boldsymbol{x}) \tag{63}$$

Comme $P(\omega_j|\boldsymbol{x}) = \frac{P(\omega_j)P(\boldsymbol{x}|\omega_j)}{P(\boldsymbol{x})}$, on peut aussi écrire

$$\widehat{d}_{Bayes}(\boldsymbol{x}) = \arg\min_i \rho'_i(\boldsymbol{x}) \quad \text{où} \quad \rho'_i(\boldsymbol{x}) = \sum_{j=1}^{C} \alpha_{ij} P(\omega_j) P(\boldsymbol{x}|\omega_j) \tag{64}$$

Si on ne connaît pas les paramètres de $P(\boldsymbol{x}|\omega_j)$, mais qu'on dispose d'une base d'apprentissage, on peut alors définir le discriminateur de Bayes "quadratique"

$$\widehat{d}_{Bayes,quad}(\boldsymbol{x}) = \arg\min_i \widehat{\rho}'_{i,quad}(\boldsymbol{x}) \quad \text{où} \quad \widehat{\rho}'_{i,quad}(\boldsymbol{x}) = \sum_{j=1}^{C} \alpha_{ij} P(\omega_j) \widehat{P}_{quad}(\boldsymbol{x}|\omega_j) \tag{65}$$

où $\widehat{P}_{quad}(\boldsymbol{x}|\omega_j) = f(\boldsymbol{x}, (\widehat{\boldsymbol{\mu}}_c, \widehat{\boldsymbol{\Gamma}}_c))$, c'est à dire qu'on remplace les paramètres par leur estimation au sens du maximum de vraisemblance (sans supposer que $\boldsymbol{\Gamma}_1 = ... = \boldsymbol{\Gamma}_C$). De même, on considère aussi $\widehat{d}_{Bayes,lin}$ qui est similaire, à la différence qu'il utilise la connaissance a priori que $\boldsymbol{\Gamma}_1 = ... = \boldsymbol{\Gamma}_C$.

2.8.1 Exemple didactique dans $\mathbb{R}^2$

1.1/ *Fonction coût standard et prior uniforme*

On considère un problème à $C = 3$ classes en dimension $N = 2$. On suppose un prior uniforme ($P(\omega_i) = 1/C$) et on considère la fonction coût $\alpha_{ij} = 0$ pour $i = j$ et $\alpha_{ij} = 1$ sinon. En lançant le programme, on génère une base de généralisation de taille P_{gen}. On compare $\widehat{d}_{\mu,\Gamma \text{ connus}}$ et $\widehat{d}_{Bayes}$ dont les résultats sont représentés respectivement Figure 1 et 3 (les ronds représentent les classes estimées et en noir sont représentées les surfaces discriminantes). De plus, Figure 13 représente les vraisemblances pour chacune des classes ainsi que les densités a posteriori pour chacune des classes. Effectuez le lien entre ces différentes figures. Validez avec un encadrant.

1.2/ *Estimation du risque*

Les Figures 2 et 4 représentent les matrices de confusion correspondantes. Chaque élément (i, j) de cette matrice représente le nombre de fois que le discriminateur a attribué à un vecteur de la base de généralisation la classe i alors que celui-ci appartient à la classe j. Pour estimer le risque à partir de la réalisation de la base de généralisation, on part de l'expression $\mathcal{R} = \sum_{i=1}^{C} \sum_{j=1}^{C} \int_{R_i} \alpha_{ij} P(\omega_j, \boldsymbol{x})\, d\boldsymbol{x}$ et on remarque que

$$\int_{R_i} P(\omega_j, \boldsymbol{x}) d\boldsymbol{x} = P(\omega_j) \int_{R_i} P(\boldsymbol{x}|\omega_j) d\boldsymbol{x} = P(\omega_j) P(\omega_i|\omega_j) \tag{66}$$

où $P(\omega_i|\omega_j)$ est la probabilité de choisir la classe ω_i quand la classe est en réalité ω_j. Ainsi

$$\mathcal{R} = \sum_{i=1}^{C} \sum_{j=1}^{C} \alpha_{ij} P(\omega_j) P(\omega_i|\omega_j) \tag{67}$$

Comment estimer $P(\omega_i|\omega_j)$ à partir de la matrice de confusion ? Commentez les taux de généralisation ainsi que les risques estimés. Validez avec un encadrant.

1.3/ *Influence du choix de la fonction coût*

En modifiant la ligne 13, on modifie le choix de la fonction coût. Plus précisément, on considère

$$\alpha_{ij} = \begin{cases} \alpha_{12} = 100 \\ \alpha_{ij} = 0 \quad \text{pour } i = j \\ \alpha_{ij} = 1 \quad \text{sinon} \end{cases} \tag{68}$$

Observez et commentez les différences avec les précédentes questions. Validez avec un encadrant.

1.4/ *Influence du prior*

En modifiant la ligne 16, on modifie le prior de sorte à avoir $P(\omega_1) = P(\omega_3)$ et $P(\omega_2) = 10P(\omega_1)$. Observez et commentez les différences avec les précédentes questions. Validez avec un encadrant.

2.8.2 Analyse du problème avec 10 chiffres

On considère le problème de discrimination avec les 10 chiffres (cf. TP n°3)

$$\boldsymbol{x} = \boldsymbol{\mu}_c + \boldsymbol{z} \quad \forall c \in \{1,\, 2,\, ...,\, 9,\, 10\} \tag{69}$$

où $\boldsymbol{\mu}_c$ est une image binaire de taille 7×6 décrivant un chiffre et $\boldsymbol{z}$ est une réalisation aléatoire d'un bruit blanc gaussien centré de variance 0.3.

2.1/ *Comparaison avec l'image des coefficients de corrélation*

Comparez les performances du discriminateur de Bayes (figure 2) avec l'image des coefficients de corrélation entre les images de chiffres (voir figure 3). Commentez. Validez avec un encadrant.

2.2/ *Comparaison entre plusieurs discriminateurs*
On analyse la méconnaissance des moyennes $\boldsymbol{\mu}_c$ et covariances $\boldsymbol{\Gamma}_c$. Pour cela, on compare $\widehat{d}_{Bayes}(\boldsymbol{x})$ (voir Eq. (64)) aux discriminateurs $\widehat{d}_{Bayes,lin}$ et $\widehat{d}_{Bayes,quad}$ (voir Eq. 65). Dans le programme `main.m`, modifiez la ligne 27. Pour une base de généralisation fixée, on trace l'évolution du risque $\mathcal{R}$ (voir figure 12345) en fonction de la taille de la base d'apprentissage (on génère une nouvelle base pour chaque valeur de P_{app}). Commentez et validez avec un encadrant.

2.3/ *Modification du choix de la fonction coût et du prior*
On souhaite voir l'influence du choix de la fonction coût et du prior sans relancer l'apprentissage car cela prend du temps. Est-il possible de recalculer le risque sur une nouvelle base de généralisation sans avoir besoin de relancer l'apprentissage ?

On utilise donc une fonction `relancer_ss_refaire_l_apprentissage.m` ce qui permettra de ne pas avoir à relancer la fonction `main.m`. Ouvrez le fichier `relancer_ss_refaire_l_apprentissage.m` et modifier `choix_cout=2` et/ou `choix_Prior=2` ligne 4 et 7. Commentez.

Proposez une synthèse en expliquant vos difficultés à utiliser la théorie du risque Bayésien. Validez avec un encadrant.

2.8.3 Réseau de neurones

Pour concevoir un discriminateur fondé sur le risque de Bayes avec un réseau de neurones, on étudie dans ce TP deux solutions qui consistent à utiliser la couche de sortie du réseau de neurones pour estimer $P(\omega_i|\boldsymbol{x})$. Soit $(\eta_i)_{i=1,\ldots,C}$ les C entrées des neurones de la couche de sortie du RN (voir Eq. (73)), la première solution consiste à estimer $P(\omega_i|\boldsymbol{x})$ avec $\frac{\beta_i}{\sum_{i=1}^{C}\beta_i}$ où $\beta_i = \frac{1}{1+e^{-2\eta_i}}$. La deuxième solution consiste à estimer $P(\omega_i|\boldsymbol{x})$ avec $\frac{\kappa_i}{\sum_{i=1}^{C}\kappa_i}$ où $\kappa_i = e^{2\eta_i}$. Dans les deux cas, on utilise ces estimations de $P(\omega_i|\boldsymbol{x})$ dans les équations (62) et Eq. (63) pour en déduire une discriminateur, que l'on note respectivement $\widehat{d}_{Bayes,RN_\beta}$ et $\widehat{d}_{Bayes,RN_\kappa}$. Commentez le choix de ces deux solutions.

3.1/ *Apprentissage du réseau de neurones et analyse des performances*
Quand vous lancez le programme `main`, par défaut, on effectue 10 apprentissages du réseau de neurones (RN) pour $P_{app} = 1000$. Commentez les performances obtenues de $\widehat{d}_{RN}$ (voir Eq. (76)), $\widehat{d}_{Bayes,RN_\beta}$ et $\widehat{d}_{Bayes,RN_\kappa}$. Si besoin, changez la taille de la base d'apprentissage (voir ligne 53). Validez avec un encadrant.

3.2/ *Modification du choix de la fonction coût et du prior*
Modifier la fonction coût (ligne 46), puis le prior (ligne 49). Commentez les résultats.

3.3/ *Recherche d'une solution plus rapide*
Proposer une solution pour ne pas avoir besoin d'effectuer un nouvel apprentissage à chaque fois que l'on modifie le prior et/ou la fonction coût. Commentez. Validez avec un encadrant.

Proposez une synthèse en expliquant vos difficultés à utiliser la théorie du risque Bayésien.

Rédigez votre compte rendu

Annexe : fonctionnement d'un réseau de neurones

On rappelle le fonctionnement d'un réseau de neurones avec N_c neurones sur la couche cachée et N_s neurones sur la couche de sortie. A l'entrée de chaque neurone $j \in \{1, ..., N_c\}$ de la couche cachée

$$\nu_j = \sum_{i=0}^{N} x_i w_{ji} = \boldsymbol{w}_j^T \boldsymbol{x} \tag{70}$$

où w_{ji} est le poids relatif à la connexion allant de l'entrée i vers le neurone de la couche cachée j. A la sortie de chaque neurone de la couche cachée on a :

$$y_j = f(\nu_j) = f\left(\boldsymbol{w}_j^T \boldsymbol{x}\right) \tag{71}$$

où la fonction d'activation f choisie pour ce TP est la sigmoïde définie par :

$$f(z) = \frac{1 - e^{-2z}}{1 + e^{-2z}} \tag{72}$$

A l'entrée de chaque neurone $k \in \{1, ..., N_s\}$ de la couche de sortie, on a :

$$\eta_k = \sum_{j=0}^{N_c} y_j w_{kj} = \boldsymbol{w}_k^T \boldsymbol{y} \tag{73}$$

où w_{kj} est le poids relatif à la connexion allant du neurone de la couche cachée j vers le neurone de sortie k. A la sortie de chaque neurone de la couche de sortie, on a :

$$z_k = f(\eta_k) = f\left(\boldsymbol{w}_k^T \boldsymbol{y}\right) = f\left(\sum_{j=1}^{N_c} w_{kj} f\left(\sum_{i=1}^{N} w_{ji} x_i + w_{j0}\right) + w_{k0}\right) \tag{74}$$

Les paramètres $(\boldsymbol{w}_j)_{j=1..,N_c} = (w_{ji})_{j\in\{1,...,N_c\},i\in\{1,...,N\}}$ et $(\boldsymbol{w}_k)_{j=1..,N_s} = (w_{kj})_{k\in\{1,...,N_s\},j\in\{1,...,N_c\}}$ sont estimés avec une base d'apprentissage $\mathcal{B}_{app} = \left\{\left(\boldsymbol{x}_{app}^{(p)}, d_{app}^{(p)}\right); p \in \{1, ..., P_{app}\}\right\}$.

Le critère utilisé pour estimer les paramètres $\boldsymbol{w}_j$ et $\boldsymbol{w}_k$ est :

$$J_T = \frac{1}{P_{app}} \sum_{\boldsymbol{x}\in\mathcal{B}_{app}} J(\boldsymbol{x}, \boldsymbol{w}_j, \boldsymbol{w}_k) \quad \text{où} \quad J(\boldsymbol{x}, \boldsymbol{w}_j, \boldsymbol{w}_k) = \frac{1}{2}\sum_{k=1}^{N_s}(t_k - z_k)^2 \tag{75}$$

où $(t_k)_{k=\{1,..,N_s\}}$ est le vecteur des valeurs cibles des neurones de la couche de sortie pour le vecteur $\boldsymbol{x}$ de la base d'apprentissage. Dans ce TP, on a choisi les conventions suivantes : N_s est égal au nombre de classe et $t_k = 1$ si $\boldsymbol{x}$ appartient à la classe k et $t_k = -1$ sinon. Une fois les matrices $\boldsymbol{w}_j$ et $\boldsymbol{w}_k$ estimées, le réseau de neurones effectue la discrimination avec

$$\widehat{d}_{RN}(\boldsymbol{x}) = \arg \max_{k=1,\,...,\,N_s} z_k \tag{76}$$

L'algorithme récursif utilisé pour estimer les paramètres $\boldsymbol{w}_j$ et $\boldsymbol{w}_k$ qui minimisent le critère J_T (voir Eq. (75)) est fondé sur un gradient stochastique. Celui-ci consiste à calculer le gradient de J pour chaque élément de la base d'apprentissage de manière itérative. Pour optimiser cet algorithme, il est souvent utile d'effectuer plusieurs "passes" à travers la base d'apprentissage.