2.2 Séance de TP sur la discrimination linéaire

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 premier TP est la discrimination linéaire dans un problème à deux classes et la précision de l'estimation des performances.

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 à l'enseignant pendant la séance. Ce TP donne lieu à un compte rendu d'une page A4 (recto uniquement, taille de police 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. Le compte rendu doit aussi poser des questions destinées à alimenter une discussion qui aura lieu lors du prochain cours. Chaque groupe travaille à son rythme et rédige son compte rendu en autonomie.

Contexte de l'étude

On considère un problème synthétique de discrimination à 2 classes pour lequel l'espace des mesures caractéristiques est $\mathbb{R}^N$. A tout vecteur $\boldsymbol{x}$ de cet espace, on associe le scalaire $d = 1$ si la classe de $\boldsymbol{x}$ est 1 et $d = -1$ si la classe de $\boldsymbol{x}$ est 2. Le problème est dit "linéairement séparable", s'il existe un vecteur $\boldsymbol{w}_o$ telle que $\forall \boldsymbol{x}$

$$d = \text{sgn}(\boldsymbol{w}_o^T \boldsymbol{x}) \tag{37}$$

où $\text{sgn}(x) = 1$ si $x \geq 0$ et $\text{sgn}(x) = -1$ sinon, et ^{T} est l'opérateur transpose. La frontière linéaire $\boldsymbol{w}_o$, qu'elle existe ou non, est estimée par un discriminateur linéaire à partir d'une base d'apprentissage

$$\mathcal{B}_{app} = \left\{ \left(\boldsymbol{x}^{(p)}, d^{(p)} \right) ; p \in I_{app} \right\} \quad \text{avec } I_{app} \text{ un sous-ensemble de } \mathbb{N} \text{ de taille } P_{app} \tag{38}$$

Tout discriminateur linéaire est ainsi caractérisé par un vecteur $\boldsymbol{w} = f(\mathcal{B}_{app})$ et l'évaluation des performances de ce discriminateur s'effectue avec une base de généralisation

$$\mathcal{B}_{gen} = \left\{ \left(\boldsymbol{x}^{(p)}, d^{(p)} \right) ; p \in I_{gen} \} \right\} \quad \text{avec } I_{gen} \text{ un sous-ensemble de } \mathbb{N} \text{ de taille } P_{gen} \text{ et } I_{app} \cap I_{gen} = \emptyset \tag{39}$$

On peut alors calculer

$$\widehat{d}^{(p)} = \text{sgn}(\boldsymbol{w}^T \boldsymbol{x}^{(p)}) \quad \forall p \in I_{gen} \tag{40}$$

et pour quantifier les performances, on calcule le taux de réussite :

$$\tau_g(\boldsymbol{w}, \mathcal{B}_{gen}) = \frac{N_{gen}}{P_{gen}} \tag{41}$$

où N_{gen} est le nombre d'éléments de la base de généralisation pour lesquels $\widehat{d}^{(p)} = d^{(p)}$. Dans la suite, on appellera τ_g le taux de généralisation.

2.2.1 Représentation des bases d'apprentissage en 2 dimensions ($N = 2$)

Dans cette partie, on considère 2 types de bases de données synthétiques avec $N = 2$. Pour chacune, on génère aléatoirement des bases d'apprentissage et on analyse les frontières linéaires obtenues avec les algorithmes de Hebb et Pseudo-inverse.

1.a/ On considère une base linéairement séparable. Exécutez (plusieurs fois) le script `main.m`. A l'aide des figures obtenues, analysez les performances des deux discriminateurs Hebb et PI en précisant votre critère.

1.b/ Modifier la ligne 9 du script `main.m` de manière à choisir $P_{app} = 2000$, relancez (plusieurs fois) le script, et commentez les nouveaux résultats obtenus.

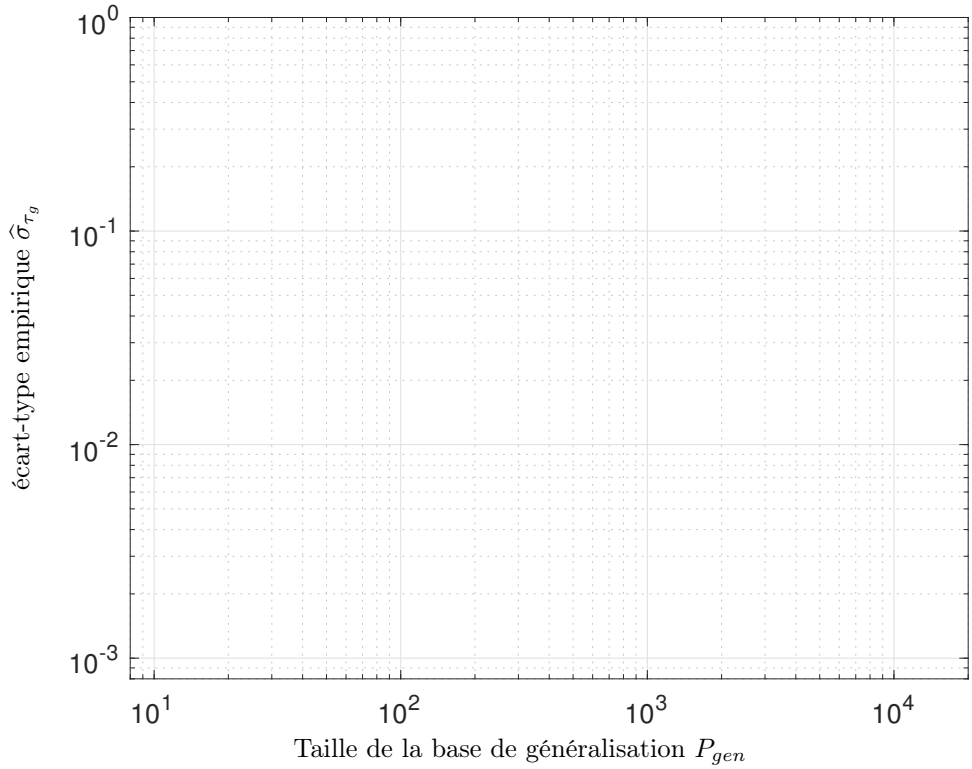
1.c/ On considère à présent une base non-linéairement séparable. Modifier la ligne 12 du script `main.m` de manière à choisir `choix_base=2;`, relancez (plusieurs fois) le script en modifiant la taille de la base et commentez les résultats obtenus.

Rédigez votre compte rendu sans essayer d'être exhaustif sur tout ce que vous avez observé. L'important est de mentionner précisément ce qui vous paraît essentiel en soulignant les difficultés rencontrées. Si besoin, formulez une (ou deux) questions pour le prochain cours.

2.2.2 Analyse de la précision d'estimation du taux de généralisation

L'objectif de cette partie est d'analyser l'écart-type du taux de généralisation $\tau_g(\boldsymbol{w}, \mathcal{B}_{gen})$ (défini Eq.(41)) en fonction de la taille de la base de généralisation P_{gen}. On considère un problème de discrimination avec $N = 42$ et $P_{app} = 42$. On analyse les performances de l'algorithme de Hebb dont on note simplement $\boldsymbol{w}$ la frontière estimée.

2.a/ L'expérience menée consiste à effectuer $M = 1000$ réalisations indépendantes d'une base de généralisation $\mathcal{B}_{gen}$ afin de calculer la moyenne empirique et l'écart-type empirique des différentes réalisations de $\tau_g(\boldsymbol{w}, \mathcal{B}_{gen})$. De plus, afin d'étudier l'influence de la taille de la base de généralisation P_{gen}, cette analyse est effectuée pour $P_{gen} = \{10, 100, 1000, 10000\}$.
En exécutant le programme avec `question=2`, vous avez à l'écran une figure qui représente, pour différentes tailles de base de généralisation (10, 100, 1000, 10000), l'histogramme des $M = 1000$ réalisations de $\tau_g(\boldsymbol{w}, \mathcal{B}_{gen})$. Dans chaque titre est indiqué la moyenne empirique $\widehat{\mu}_{\tau_g}$ et l'écart-type empirique $\widehat{\sigma}_{\tau_g}$ de $\tau_g(\boldsymbol{w}, \mathcal{B}_{gen})$. Tracez ci-dessous la valeur de l'écart-type empirique $\widehat{\sigma}_{\tau_g}$ en fonction de la taille de la base de généralisation P_{gen}.



2.b/ En vous appuyant sur le graphique ci-dessus, trouvez un lien entre l'écart-type empirique $\widehat{\sigma}_{\tau_g}$ et la taille de la base de généralisation P_{gen}.

2.c/ En vous fondant sur cette expérience, pensez-vous que la relation $\sigma_{\tau_g} = \sqrt{\frac{\mu_{\tau_g}(1-\mu_{\tau_g})}{P_{gen}}}$ soit vérifiée ?

2.d/ En déduire un nouvel estimateur de la précision σ_{τ_g} lorsque vous ne disposez que d'une seule réalisation d'une base de généralisation (c'est à dire $M = 1$). Validez votre analyse avec un encadrant.

Rédigez votre compte rendu pour cette partie en décrivant scientifiquement vos observations. Formulez une ou deux questions pour le prochain cours.

2.2.3 Analyse des performances en fonction de la taille de la base d'apprentissage

3.a/ On analyse les performances en discrimination pour les algorithmes de Hebb et Pseudo-inverse en fonction de la taille de la base d'apprentissage P_{app}. La dimension de l'espace est fixée à $N = 42$. Avec `question=3`, vous obtenez une figure. Pour chaque valeur de P_{app}, on génère une nouvelle base $\mathcal{B}_{app}$. On estime la frontière correspondante. On en déduit le taux d'apprentissage τ_{app} calculé avec $\mathcal{B}_{app}$ et représenté en haut. Pour le calcul du taux de généralisation τ_g, on utilise une seule et même base de généralisation $\mathcal{B}_{gen}$ de taille P_{gen}, que l'on représente en bas. Commentez les résultats obtenus en soulignant les liens avec les questions précédentes.

3.b/ La formule obtenue à la question 2.c vous permet-t-elle de décrire approximativement les variations observées ? Pour répondre, vous pourrez modifier la valeur de P_{gen} (ligne 19 du programme `main.m`). Validez avec un encadrant. Expliquez les difficultés rencontrées lors de votre analyse.

2.2.4 Analyse de l'algorithme "Ridge Approximation"

4.a/ On reste sur le même problème, mais on se focalise sur le cas où $P_{app} = 42$, on analyse les performances de l'algorithme de discrimination appelé "Ridge approximation" (voir annexe). Pour utiliser ce nouvel algorithme, il est nécessaire de régler le paramètre σ. Afin d'analyser son influence, on estime les taux d'apprentissage et de généralisation pour différentes valeurs de σ. Exécutez le programme pour `question=4`. Observez et commentez.

4.b/ Modifiez le programme pour avoir `choix_nouvelle_base_app=1` (respectivement `choix_nouvelle_base_gen=1`) afin de changer la réalisation de $\mathcal{B}_{app}$ (respectivement de $\mathcal{B}_{gen}$) à chaque nouvelle valeur de σ. Observez et commentez. Validez avec un encadrant.

4.c/ Est-ce qu'il y a un inconvénient à utiliser le taux de réussite obtenu sur la base de généralisation pour apprendre la bonne valeur de σ ?

4.d/ Proposez une méthodologie empirique qui puisse être utilisée en pratique pour choisir la valeur de σ à partir d'une unique réalisation de la base d'apprentissage. Validez avec un encadrant.

Rédigez votre compte rendu pour cette partie en décrivant scientifiquement vos observations. De plus, formulez une ou deux questions pour le prochain cours.

Annexe : Présentation des différents algorithmes de discrimination

Discriminateur 1 : Hebb

Le vecteur $\boldsymbol{w}$ est défini par

$$\boldsymbol{w} = \boldsymbol{X}\,\boldsymbol{d}$$

Discriminateur 2 : Pseudo-Inverse

Le vecteur $\boldsymbol{w}$ est défini par

$$\boldsymbol{w} = \left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\,\boldsymbol{d}$$

Si $P_{app} \geq N$, alors ci-dessus, $^{-1}$ représente l'opérateur inverse. Sinon, il représente l'opérateur pseudo-inverse.

Discriminateur 3 : Ridge approximation

Contrairement aux 2 discriminateurs précédents, celui-ci nécessite de régler un paramètre noté σ. Le vecteur $\boldsymbol{w}$ est défini par

$$\boldsymbol{w} = \left(\boldsymbol{X}\,\boldsymbol{X}^T + P_{app}\sigma^2\boldsymbol{I}\boldsymbol{d}\right)^{-1}\boldsymbol{X}\,\boldsymbol{d}$$