

IFT3395 Fondements de l'apprentissage machine

Professeur: Guillaume Rabusseau

Devoir 1

- Ce devoir est à faire en équipe de 2 personnes. Assurez-vous d'avoir noté le nom de tous les coéquipiers en tête du rapport et en tête en commentaires de chacun des fichiers que vous remettrez.
- On demande, la remise d'un rapport en format électronique (.pdf). Tous les fichiers de code source que vous aurez créé ou adapté devront également être remis. La partie pratique est à faire en python (en utilisant les librairies numpy et matplotlib), et vous pouvez bien entendu fortement vous inspirer de ce qui a été fait pendant les labos.
- Vous pouvez remettre votre code python sous la forme d'un notebook ipython .ipynb. Pour produire un rapport avec des formules mathématiques vous pouvez utiliser le logiciel de votre choix: L^AT_EX; L^AT_EX; Word; voire même écrire directement les parties théoriques dans le notebook (en entrant les équations en format MathJaX pour qu'elles s'affichent). Dans tous les cas on vous demande d'exporter votre rapport en .pdf que vous remettrez.
- La remise doit se faire via StudiUM. Une seule remise par équipe (un seul des coéquipiers effectue la remise). Assurez-vous d'avoir noté le nom de tous les coéquipiers en tête du rapport. Si vous avez beaucoup de fichiers à remettre vous pouvez aussi (c'est peut-être plus pratique) en faire une archive (.zip ou .tar.gz) et téléverser le fichier d'archive.

Important:

- Vous pouvez faire au choix l'exercice 1 ou l'exercice 2 (vous pouvez également faire ces deux exercices, auquel cas l'exercice ayant obtenu le maximum de points sera utilisé pour la note finale).
- La question 3.4 est optionnelle, vous pouvez y répondre pour obtenir des points en *bonus*.

1 Petit exercice de probabilités [20 points]

Une étude réalisée aux États-Unis il y a quelques années auprès de médecins, pour mesurer leur “intuition probabiliste” comportait la question suivante:

La probabilité d’un cancer du sein chez les femmes dans la quarantaine participant à un test de routine (mammographie) est de 1.5%. Si une femme est atteinte d’un cancer du sein, il y a 87% de chances que le test soit positif. Chez une femme qui n’est pas atteinte de cancer, il y a une probabilité de 9,6% que le test soit positif.

Une femme dans la quarantaine ayant passé ce test de routine reçoit un résultat de test positif. Quelle est la probabilité qu’elle soit effectivement atteinte d’un cancer du sein?

- A) plus de 90%
- B) entre 70% et 90%
- C) entre 50% et 70%
- D) entre 30% et 50%
- E) entre 10% et 30%
- F) moins de 10%

95% des médecins interrogés ont répondu B). Qu’en pensez-vous?

Formalisez la question et calculez la probabilité exacte.

Indication: utilisez la règle de Bayes...

2 Fléau de la dimensionalité et intuition géométrique en haute dimension [20 points]

1. On considère un hyper-cube en dimension d (c’est une généralisation du carré 2D et du cube 3D) de côté c (qui peut être exprimé en cm par ex.). Quel est son volume V ?
2. On définit une variable aléatoire vectorielle X en dimension d ($x \in \mathbb{R}^d$) distribuée uniformément au sein de l’hypercube (la densité de probabilité est $p(x) = 0$ pour x en dehors du cube). Quelle est la densité de

probabilité $p(x)$ à l'intérieur du cube? Indiquez quelle(s) propriété(s) des densités de probabilité vous permet d'arriver à ce résultat.

3. On considère une zone de bordure étroite d'une largeur de 3% de a bordant l'hypercube vers l'intérieur. Par exemple si $a = 100cm$, la bordure sera de $3cm$ (à gauche, à droite, en haut, etc...) et délimitera donc un second hypercube de côté $100 - 3 - 3 = 94cm$. Si on génère un point x de la distribution précédemment définie (on effectue un tirage), quelle est la probabilité qu'il tombe dans la zone de bordure? Quelle est la probabilité qu'il tombe dans l'hypercube plus petit?
4. Calculez numériquement la probabilité que x tombe dans l'étroite bordure pour les valeurs suivantes de d : 1, 2, 3, 5, 10, 100, 1000.
5. Qu'en concluez-vous sur la répartition des points en haute dimension, qui est contraire à notre intuition en faible dimension?

3 Estimation de densité paramétrique Gaussienne, v.s. estimation de densité par fenêtres de Parzen [30 points]

Dans cette question on considère un ensemble de données $D = \{x^{(1)}, \dots, x^{(n)}\}$ avec $x \in \mathbb{R}^d$.

1. Supposons que l'on ait entraîné les paramètres d'une densité Gaussienne **isotropique** sur D (pour maximiser la vraisemblance) pour en estimer la densité de probabilité.
 - (a) Nommez ces paramètres et indiquez-en les dimensions.
 - (b) Si on apprend les paramètres en utilisant le principe de maximum de vraisemblance, exprimez en fonction des points de D la formule qui nous donnera la valeur des paramètres optimaux (indiquez seulement la formule qui calcule le résultat, on ne vous demande pas de la redériver, les formules de maximum de vraisemblance se trouvent dans les slides 5 sur la distribution de Gauss vers la fin.)
 - (c) Quelle est la complexité algorithmique de cet apprentissage (entraînement) c.a.d. du calcul de ces paramètres?

- (d) Pour un point de test x , écrivez la fonction qui donnera la densité de probabilité prédite au point x :
 $\hat{p}_{\text{gauss-isotrop}}(x) = ?$
 - (e) Quelle est la complexité algorithmique pour le calcul de cette prédiction à chaque nouveau point x ?
2. Considérons à présent qu'on utilise plutôt pour estimer la densité de probabilité des fenêtres de Parzen avec un noyau Gaussien isotropique de largeur (écart-type) σ , et qu'on ait entraîné ces fenêtres de Parzen sur D .
- (a) Supposons que l'utilisateur ait fixé σ . En quoi consiste la phase "entraînement/apprentissage" pour ces fenêtres de Parzen?
 - (b) Pour un point de test x , écrivez en *une seule formule détaillée* (c.a.d. avec des exponentielles), la fonction qui donnera la densité de probabilité prédite au point x :
 $\hat{p}_{\text{Parzen}}(x) = ?$
 - (c) Quelle est la complexité algorithmique pour le calcul de cette prédiction à chaque nouveau point x ?
3. Capacité
- (a) Laquelle de ces 2 approches (paramétrique Gaussienne v.s. Parzen à noyau Gaussien) a la plus forte *capacité*? Expliquez.
 - (b) Avec laquelle de ces approches, et dans quel *cas précis*, a-t-on toutes les chances d'être en sur-apprentissage?
 - (c) Le σ dans les fenêtres de Parzen est généralement traité comme un hyper-paramètre, alors que pour une densité paramétrique Gaussienne il est généralement traité comme un paramètre. Pourquoi?
4. **[Bonus]** Considérons maintenant une estimation de densité paramétrique avec une densité Gaussienne **diagonale**.
- (a) Exprimez l'équation d'une densité Gaussienne diagonale dans \mathbb{R}^d . Et précisez ce que sont ses paramètres et leurs dimension.
 - (b) Démontrez que les *composantes* d'un vecteur aléatoire qui suit une distribution Gaussienne diagonale sont des variables aléatoires **indépendantes**.

- (c) En utilisant comme coût $-\log p(x)$ écrivez l'équation qui correspondrait à la minimisation du risque empirique sur l'ensemble d'entraînement D (pour apprendre les paramètres)
- (d) Résolvez cette équation de manière analytique pour obtenir les paramètres optimaux.

4 Partie pratique: estimation de densité [50 points]

1. Implémentez un estimateur de densité paramétrique Gaussien diagonal. Il devra pouvoir fonctionner pour des données de dimension d arbitraire. Comme vu dans les labos, il devrait avoir une méthode **train** pour apprendre les paramètres et une méthode **predict()** qui calcule les *log* de densité.
2. Implémentez un estimateur de densité de Parzen à noyau Gaussien isotropique. Il devra pouvoir fonctionner pour des données de dimension d arbitraire. De même il devrait avoir une méthode **train** pour apprendre les paramètres et une méthode **predict()** qui calcule les *log* de densité.
3. Densités 1D: Parmi l'ensemble de données Iris, choisissez le sous-ensemble correspondant à une des classes (de votre choix), et un des traits caractéristique, de sorte qu'on sera en dimension $d = 1$ et produisez un unique graphique (à l'aide de la fonction `plot`) comportant:
 - (a) les points du sous-ensemble de données (affichés sur l'axe des x)
 - (b) une courbe de la densité estimée par votre estimateur paramétrique Gaussien
 - (c) une courbe de la densité estimée par l'estimateur de Parzen avec un hyper-paramètre σ (écart type) trop petit
 - (d) une courbe de la densité estimée par l'estimateur de Parzen avec un hyper-paramètre σ un peu trop grand
 - (e) une courbe de la densité estimée par l'estimateur de Parzen avec un hyper-paramètre σ que vous jugerez plus approprié.
Utilisez une couleur différente pour chaque courbe, et munissez votre graphique d'une légende claire.
 - (f) Expliquez comment vous avez choisi votre hyper-paramètre σ

4. Densités 2D: Ajoutez maintenant un second trait caractéristique d'iris, afin d'avoir des entrées en dimension $d = 2$ et produisez 4 graphiques, chacun affichant les points du sous-ensemble de données (avec la fonction `plot`), et les lignes de contours de la densité estimée (à l'aide de la fonction `contour`) suivante:
- (a) par votre estimateur paramétrique Gaussien diagonal
 - (b) par l'estimateur de Parzen avec un hyper-paramètre σ (écart type) trop petit
 - (c) par l'estimateur de Parzen avec un hyper-paramètre σ un peu trop grand
 - (d) par l'estimateur de Parzen avec un hyper-paramètre σ que vous jugerez plus approprié.
 - (e) Expliquez comment vous avez choisi votre hyper-paramètre σ