

Devoir 1

Jean Archambault

Abou Nassif Mahmoud

Section 1 Petit exercice de probabilités

La probabilité d'un cancer du sein chez les femmes dans la quarantaine participant à un test de routine (mammographie) est de 1%. Si une femme est atteinte d'un cancer du sein, il y a 80% 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 ?

Soit X le fait d'un test de dépistage avec $X = 1$ que le test soit positif et $X = 0$ que le test soit négatif.

Soit le fait d'avoir le cancer du sein pour les femmes dans la quarantaine participant à un test de routine est Y avec $Y = 1$ d'avoir le cancer et $Y = 0$ ne pas avoir le cancer.

Selon les données du problème, la probabilité à priori que $Y = 1$ ("La probabilité d'un cancer du sein chez les femmes dans la quarantaine participant à un test de routine (mammographie) est de 1%") est :

$$p(Y = 1) = 0.01$$

Par ailleurs, "si une femme est atteinte d'un cancer du sein ($Y = 1$), il y a 80% de chances que le test soit positif ($X = 1$)" s'exprime donc comme :

$$p(X = 1 | Y = 1) = 0.80$$

De même, "chez une femme qui n'est pas atteinte de cancer ($Y = 0$), il y a une probabilité de 9,6% que le test soit positif ($X = 1$)" se formalise comme suit :

$$P(X = 1 | Y = 0) = 0.096$$

Alors, la question "Une femme dans la quarantaine ayant passé ce test de routine reçoit un résultat de test positif ($X = 1$). Quelle est la probabilité qu'elle soit effectivement atteinte d'un cancer du sein ($Y = 1$) ?" est formulé comme :

$$p(Y = 1 | X = 1) = ?$$

Par Bayes :

$$p(Y = 1 / X = 1) = \frac{p(X = 1 / Y = 1) * p(Y = 1)}{p(X = 1 / Y = 1) * p(Y = 1) + p(X = 1 / Y = 0) * p(Y = 0)}$$

Donc selon les données :

$$p(Y = 1 / X = 1) = \frac{0.8 * 0.01}{0.8 * 0.01 + 0.096 * 0.99}$$

Avec $p(Y=0) = 1.0 - p(Y=1) = 1.0 - 0.01 = 0.99$

$$p(Y = 1 / X = 1) = 0.0776$$

Donc la probabilité qu'une femme dans la quarantaine ayant passé ce test de routine avec un résultat positif soit effectivement atteinte d'un cancer du sein est de 7.76% donc moins de 10% (choix F).

Donc les médecins ont mal estimé cette probabilité puisqu'ils ont oublié, dans leurs calculs, de tenir compte de l'a priori que "La probabilité d'un cancer du sein chez les femmes dans la quarantaine participant à un test de routine (mammographie) est de 1%."

Section 2 Estimation de densité paramétrique Gaussienne vs estimation de densité par fenêtres de Parzen

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

P1. Entraînement des 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.

1. Les moyennes de chaque $x^{(i)}$ dans un vecteur $\mu = \{\mu^{(1)}, \mu^{(2)}, \dots, \mu^{(d)}\}$ de dimension d, donc d paramètres $\mu^{(i)}$.
2. Puisqu'on considère une Gaussienne isotropique, soit que la variance pour tous les traits/dimensions est la même, donc cette variance scalaire de dimension 1 σ^2 est l'autre paramètre à estimer.

(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)

Pour chaque μ_i du vecteur μ des moyennes des traits :

$$\mu_i = \frac{1}{n} \sum_{j=1}^n x_i^{(j)}$$

et sous forme vectorielle :

$$\mu = (\mu_1, \dots, \mu_d)$$

et pour la variance σ^2 :

$$\sigma^2 = \frac{\sum_i (x^{(i)} - \mu)^T (x^{(i)} - \mu)}{n * d}$$

(c) Quelle est la complexité algorithmique de cet apprentissage (entraînement) c.a.d. du calcul de ces paramètres ?

$$O(nd)$$

(d) Pour un point de test x , écrivez la fonction qui donnera la densité de probabilité prédite au point x :

$$\hat{p}_{\text{gauss-isotrop}}(x) = \frac{1}{(2\pi)^{d/2} * \sigma^d} * e^{\frac{-|x-\mu|^2}{2\sigma^2}}$$

(e) Quelle est la complexité algorithmique pour le calcul de cette prédiction à chaque nouveau point x ?

$$O(d)$$

P2. Estimation de la densité de la probabilité des fenêtres de Parzen avec un noyau Gaussien isotropique de largeur (écart-type) σ , avec ces fenêtres de Parzen entraînées sur D.

(a) Soit que σ est fixé. En quoi consiste la phase « entraînement/apprentissage » pour ces fenêtres de Parzen ?

Elle consiste :

1. À estimer la moyenne μ_i de chaque dimension/trait ($i = 1$ à d) à partir des exemples/données $D = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ où $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)}\}$ soit d μ_i calculées par l'équation :

$$\mu_i = \frac{1}{n} \sum_{j=1}^n x_i^{(j)}$$

Ce qui donne le vecteur de moyenne

$$\mu = (\mu_1, \mu_2, \mu_3, \dots, \mu_d)$$

2. À stocker tous les points de l'ensemble d'entraînement pour calculer, de façon pondérée selon les $p(x)$, la classe majoritaire qui sera associée au point test x selon ses plus proches voisins de l'ensemble d'entraînement.

(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 :

$$\hat{p}_{Parzen}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{(2\pi)^{d/2} \sigma^d} e^{\frac{-|x-\mu|}{2\sigma^2}}$$

(c) Quelle est la complexité algorithmique pour le calcul de cette prédiction à chaque nouveau point x ?

$$O(d)$$

P3. Capacité

(a) Laquelle de ces 2 approches (paramétrique Gaussienne v.s. Parzen à noyau Gaussien) a la plus forte capacité ? Expliquez.

L'approche Parzen à noyau Gaussien est basée sur une moyenne de Gaussienne comportant " d " moyennes μ_i estimées des données, un hyper paramètre, la variance σ^2 , fixe et l'ensemble des données d'entraînement qui sont aussi des paramètres du modèle.

L'approche paramétrique Gaussienne implique une seule Gaussienne comportant " d " moyennes μ_i et une variance σ^2 toutes estimées des données.

L'approche Parzen à noyau Gaussien a donc une plus forte capacité puisqu'elle comporte beaucoup plus de paramètres, soit les d moyennes de chaque dimension μ_i et toutes les données d'entraînement. Elle est donc plus riche, plus flexible et comporte un degré de liberté plus élevé que l'approche Gaussienne.

(b) Avec laquelle de ces approches, et dans quel cas précis, a-t-on toutes les chances d'être en sur-apprentissage ?

L'approche Parzen à noyau Gaussien lorsque l'hyper paramètre variance est fixé trop élevé par rapport aux distances entre les points d'entraînement. À la limite, si celle-ci est trop grande, elle englobera tous les points d'entraînement et seule la classe majoritaire de l'ensemble d'entraînement sera choisie pour tous les points tests expérimentés.

(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 ?

Parce que le σ des fenêtres de Parzen est fixé avant l'apprentissage et il ne peut être optimisé qu'au moyen d'un ou plusieurs ensemble(s) de validation.

D'un autre côté, le σ de l'approche densité paramétrique Gaussienne est considéré comme un paramètre puisqu'il est appris et optimisé à l'apprentissage sur l'ensemble d'entraînement.

P4. 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 dimensions.

$$p(x) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Où d = dimension du vecteur d'entrée x

Et les paramètres sont :

μ : le vecteur colonne de dimension d contenant les moyennes μ_i ($i = 1$ à d) des données d'entraînement de chaque trait/dimension i ;

et Σ est la matrice de covariance de dimension $d \times d$ des données d'entraînement, et dans le cas d'une Gaussienne **diagonale** :

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \\ & & & \sigma_d^2 \end{pmatrix}$$

Où seules les variances σ_i^2 sur la diagonale ne sont pas nulles et toutes les autres $\sigma_{i,j}^2 = 0$; il y a donc d σ_i^2 dans la matrice des covariances Σ .

Et

$|\Sigma|$ est le déterminant de la matrice de covariances.

(b) Démontrez que les composantes d'un vecteur aléatoire qui suit une distribution Gaussienne diagonale sont des variables aléatoires indépendantes.

Dans le cas d'une distribution Gaussienne **diagonale**, toutes les covariances $\sigma_{i,j}^2$ entre chacune des variables aléatoires x_i et x_j composant le vecteur aléatoire x sont nulles ce qui implique que toutes ces variables aléatoires x_i et x_j sont mutuellement décorrélées entre elles.

La démonstration qui suit tirée de <http://cs229.stanford.edu/section/gaussians.pdf> pour un exemple à deux variables distribuées selon une Gaussienne diagonale multivariée:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

En introduisant ces variables et ces paramètres dans l'équation de la Gaussienne, on obtient:

$$\begin{aligned} p(x; \mu, \Sigma) &= \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{1/2}} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right) \\ &= \frac{1}{2\pi(\sigma_1^2 \cdot \sigma_2^2 - 0 \cdot 0)^{1/2}} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right), \end{aligned}$$

En simplifiant cette dernière équation, on obtient :

$$\begin{aligned} p(x; \mu, \Sigma) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2}(x_1 - \mu_1) \\ \frac{1}{\sigma_2^2}(x_2 - \mu_2) \end{bmatrix} \right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 \right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp \left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right). \end{aligned}$$

Or cette dernière équation est en fait la multiplication des deux densités Gaussiennes séparées des variables aléatoires x_1 et x_2 qui composent le vecteur aléatoire X . Donc ceci devient :

$$p(x; \mu, \Sigma) = p(x_1, x_2; \mu, \Sigma) = p(x_1; \mu_1, \sigma_1) * p(x_2; \mu_2, \sigma_2)$$

Or on sait que 2 variables aléatoires x_1 et x_2 sont indépendantes si et seulement si

$$P(X_1, X_2) = P(X_1) * P(X_2)$$

Ce qui correspond à l'équation qui précède.

Ceci peut être généralisé à toutes les variables aléatoires x_i d'un vecteur aléatoire de dimension d distribué selon une Gaussienne diagonale.

Donc les variables aléatoires d'un vecteur aléatoire distribué selon une Gaussienne diagonale sont 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)

$$\theta^* = \arg \min_{\theta} \hat{R}(p_{\theta}, D_n) = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n (-\log(p(x^{(i)}|\theta)))$$

$$\theta^* = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \left(-\log \left(\frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-u)^T \Sigma^{-1} (x-u)} \right) \right)$$

(d) Résolvez cette équation de manière analytique pour obtenir les paramètres optimaux.

Il s'agit donc de dériver (dérivées partielles) la fonction de coût $\mathcal{J}(\theta)$ par rapport aux paramètres μ et Σ de la distribution Gaussienne diagonale des données et de mettre celle-ci à 0 :

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta} = 0$$

En premier, on dérive par rapport au vecteur des moyennes : $\mu = (\mu_1, \dots, \mu_d)$

$$\frac{\partial \mathcal{J}(\theta)}{\partial (\mu_1, \dots, \mu_d)} = 0$$

Puisqu'il s'agit, pour les moyennes μ_k ($k = 1$ à d), de la même situation que le cas de la Gaussienne isotropique démontré au cours après une longue dérivation, on obtient les paramètres optimaux pour μ :

$$\mu_k = \frac{1}{n} \sum_{i=1}^n x_k^{(i)}$$

Et la moyenne empirique

$$\mu = \frac{1}{n} \sum_{i=1}^n x^{(i)}$$

Par la suite, on dérive par rapport à Σ :

$$\frac{\partial l(\theta)}{\partial \Sigma} = 0$$

Or nous considérons une distribution Gaussienne diagonale pour laquelle, comme démontré au point P4 (b) plus haut, les variables aléatoires x_i du vecteur aléatoire x sont indépendantes et les $\sigma_{i,j} = 0$. Donc cette équation peut être appliquée pour déterminer indépendamment les σ_i de chaque variable x_i indépendante du vecteur x , ces σ_i étant sur la diagonale de la matrice de covariance Σ .

Donc ceci se réduit à trouver, pour chaque σ_i de Σ :

$$\frac{\partial l(\theta)}{\partial \sigma_i} = 0$$

Ceci donne l'équivalent de l'équation présentée aux notes de cours pour le cas isotropique mais considérant σ_i pour les données $x_i^{(j)}$ pour chaque dimension i de 1 à d . On obtient alors :

$$\sigma_{i \text{ max vrais}}^2 = \frac{\sum_{j=1}^n (x_i^{(j)} - \mu_i)^T (x_i^{(j)} - \mu_i)}{n}$$

P5. Problème de classification avec $D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$. Construire un classifieur de Bayes en se basant sur l'une ou l'autre de ces 3 estimations de densité : soit Gaussienne isotropique (variance σ^2 la même pour toutes les dimensions), Parzen à noyau Gaussien isotropique (variance σ^2 fixée avant l'apprentissage) et Gaussienne diagonale (variances $\sigma_{i,j}^2 = 0$ lorsque $i \neq j$ et $\sigma_i^2 \neq 0$).

(a) Écrivez dans vos propres mots comment vous apprendriez un classifieur de Bayes (phase d'entraînement)

L'apprentissage du classifieur de Bayes se fait par l'apprentissage/l'estimé des paramètres de la distribution Gaussienne qui modélise la distribution des données du problème au moyen des données d'entraînement, soit :

1. La moyennes μ_i de chaque dimension/trait pour le vecteur μ de dimension d , soit celle de l'entrée du système, pour les 3 estimations,
2. La variance σ^2 pour l'estimation par la Gaussienne isotropique pour l'ensemble des données d'entraînement et
3. Les variances σ_i^2 de la covariance diagonale Σ pour l'estimation par la Gaussienne diagonale pour chaque dimension i à partir des données d'entraînement de chaque dimension,

Le tout par l'application du principe de maximum de vraisemblance. Pour chaque classe, on calcule les valeurs de ces paramètres qui maximisent la log-vraisemblance des données d'entraînement de cette classe.

On notera que la variance pour le Parzen à noyau Gaussien n'est pas calculée puisqu'il s'agit d'un hyper-paramètre fixé avant l'apprentissage.

Le calcul de la moyenne empirique de chaque dimension/trait k maximisant la log-vraisemblance des données d'entraînement de cette dimension/trait se fait, pour les 3 estimations, au moyen de l'équation suivante :

$$\mu_k = \frac{1}{n} \sum_{i=1}^n x_k^{(i)}$$

La moyenne empirique de chaque dimension/trait k est une composante du vecteur μ :

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n x^{(i)}$$

$$\hat{\mu}_{ML} = (\mu_1, \dots, \mu_d)$$

Le calcul de la variance σ^2 pour la Gaussienne isotropique pour l'ensemble des données d'entraînement se fait par l'équation suivante sur l'ensemble des données d'entraînement:

$$\hat{\sigma}_{\max \text{ vrais}}^2 = \frac{\sum_i (x_i - \mu)^T (x_i - \mu)}{n * d}$$

Le calcul des variances σ_i^2 de chaque dimension/trait de la covariance diagonale Σ de la Gaussienne diagonale se fait par l'équation suivante appliquée pour chaque trait/dimension d'entraînement:

$$\hat{\sigma}_{i \max \text{ vrais}}^2 = \frac{\sum_{j=1}^n (x_i^{(j)} - \mu_i)^T (x_i^{(j)} - \mu_i)}{n}$$

(b) Pour un point de test x , écrivez, la fonction qui donnera le vecteur de probabilité prédite pour chaque classe au point x :

Cas Gaussienne isotropique :

$$g(x) = \frac{1}{(2\pi)^{d/2} \sigma^d} e^{-\frac{|x-\mu|^2}{2\sigma^2}} * p_{\text{à priori}}(x)$$

Cas Gaussienne diagonale :

$$g(x) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} e^{-\frac{(x-\mu)^T (x-\mu)}{2}} * p_{\text{à priori}}(x)$$

Cas Parzen à noyau Gaussien isotropique :

$$g(x) = \frac{1}{n} \sum \left(\frac{1}{(2\pi)^{d/2} \sigma^d} e^{-\frac{|x-\mu|^2}{2\sigma^2}} \right) * p_{\text{à priori}}(x)$$

Section 3 Partie pratique : estimation de densité

P1. 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 `compute_predictions` qui calcule les log de densité.

P2. 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 `compute_predictions` qui calcule les log de densité.

P3. Densités 1D : Parmi l'ensemble de données Iris, choisissez le sous-ensemble correspondant à une des classes (de votre choix), et un des trait 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.

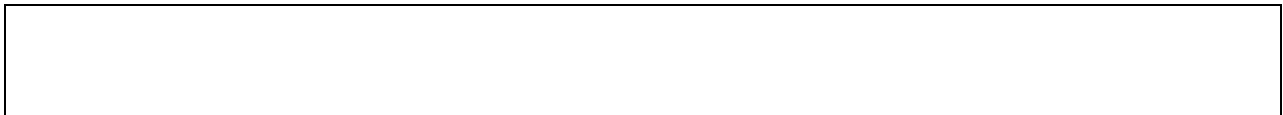
P4. Densités 2D : Ajoutez maintenant un second trait caractéristique d'iris, afin d'avoir des entrées en dimension $d = 2$ et produisez 4 graphiques, chacun affichant les points du sous-ensemble de données (avec la fonction `plot`), et les lignes de contours de la densité estimée (à l'aide de la fonction `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.



Section 4 Partie pratique : classifieur de Bayes

P1. Mélangez les exemples de Iris (utilisez `numpy.random.shuffle` après avoir initialisé le générateur aléatoire comme suit `numpy.random.seed(123)`). Puis divisez l'ensemble de tous les exemples en 2 : un ensemble d'entraînement, et un ensemble de validation.

Préparez deux versions de chacun de ces ensembles : une version complète comportant les $d = 4$ traits caractéristiques. Et une version avec seulement les $d = 2$ premiers traits caractéristiques qu'on utilisera pour fins de visualisation.

P2. Classifieur de Bayes basé sur des densités paramétriques Gaussiennes diagonales.

(a) Implémentez l'algorithme de classifieur de Bayes basé sur des densités paramétriques Gaussiennes diagonales.

(b) Visualisation en dimension $d = 2$. Considérant seulement les deux premiers traits caractéristiques d'iris, entraînez votre classifieur de Bayes sur votre ensemble d'entraînement ; affichez un graphique avec les régions de décision obtenues (ainsi que les points des ensembles d'entraînement et de validation).

(c) Calcul des erreurs en dimension $d = 2$: calculez et affichez le taux d'erreur de votre classifieur (entraîné sur les 2 premiers traits caractéristiques), à la fois sur l'ensemble d'entraînement et de validation.

(d) Calcul des erreurs en dimension $d = 4$: Entraînez votre classifieur en utilisant tous les traits caractéristiques. Puis calculez et affichez le taux d'erreur de votre classifieur (entraîné sur tous les traits caractéristiques), à la fois sur l'ensemble d'entraînement et de validation.

P3. Classifieur de Bayes basé sur des densités de Parzen avec noyau Gaussien isotropique

(a) Implémentez l'algorithme de classifieur de Bayes basé sur des densités de Parzen avec noyau Gaussien isotropique.

(b) Visualisation en dimension $d = 2$. Considérant seulement les deux premiers traits caractéristiques d'iris, entraînez votre classifieur de Bayes sur votre ensemble d'entraînement ; affichez un graphique avec les régions (surface) de décision obtenues (ainsi que les points des ensembles d'entraînement et de validation). Produisez 3 tels graphiques de régions de décision : avec un σ trop petit, trop grand, et approprié.

(c) Courbes d'apprentissage avec $d = 2$. Calculez le taux d'erreur de classification, à la fois sur l'ensemble d'entraînement, et sur l'ensemble de validation, en fonction des valeurs de l'hyper-paramètre σ (calculez-les pour une centaine de valeurs différentes de cet hyper-paramètre afin de pouvoir afficher la courbe). Indiquez la meilleure valeur de l'hyper-paramètre σ que vous avez trouvé.

(d) Courbes d'apprentissage avec $d = 4$. On va maintenant utiliser tous les traits caractéristiques. Calculez le taux d'erreur de classification, à la fois sur l'ensemble d'entraînement, et sur l'ensemble de validation, en fonction des valeurs de l'hyper-paramètre σ (calculez-les pour une centaine de valeurs différentes de cet hyper-paramètre afin de pouvoir afficher la courbe). Indiquez la meilleure valeur de l'hyper-paramètre σ que vous avez trouvé.

4. D'après ces expériences, pour le problème de classification d'Iris (et pour cette division entraînement/validation particulière) indiquez quel est le meilleur choix d'algorithme entre classifieur de Bayes avec Gaussiennes diagonales, et fenêtres de Parzen, et des autres hyper-paramètres : dimension de l'entrée (2 ou 4), et σ (s'il y a lieu). Précisez les taux d'erreurs de classification qu'ils permettent d'atteindre.

Devoir 1

Jean Archambault

Abou Nassif Mahmoud

Section 1 Petit exercice de probabilités

La probabilité d'un cancer du sein chez les femmes dans la quarantaine participant à un test de routine (mammographie) est de 1%. Si une femme est atteinte d'un cancer du sein, il y a 80% 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 ?

Soit X le fait d'un test de dépistage avec $X = 1$ que le test soit positif et $X = 0$ que le test soit négatif.

Soit le fait d'avoir le cancer du sein pour les femmes dans la quarantaine participant à un test de routine est Y avec $Y = 1$ d'avoir le cancer et $Y = 0$ ne pas avoir le cancer.

Selon les données du problème, la probabilité à priori que $Y = 1$ ("La probabilité d'un cancer du sein chez les femmes dans la quarantaine participant à un test de routine (mammographie) est de 1%") est :

$$p(Y = 1) = 0.01$$

Par ailleurs, "si une femme est atteinte d'un cancer du sein ($Y = 1$), il y a 80% de chances que le test soit positif ($X = 1$)" s'exprime donc comme :

$$p(X = 1 | Y = 1) = 0.80$$

De même, "chez une femme qui n'est pas atteinte de cancer ($Y = 0$), il y a une probabilité de 9,6% que le test soit positif ($X = 1$)" se formalise comme suit :

$$P(X = 1 | Y = 0) = 0.096$$

Alors, la question "Une femme dans la quarantaine ayant passé ce test de routine reçoit un résultat de test positif ($X = 1$). Quelle est la probabilité qu'elle soit effectivement atteinte d'un cancer du sein ($Y = 1$) ?" est formulé comme :

$$p(Y = 1 | X = 1) = ?$$

Par Bayes :

$$p(Y = 1 / X = 1) = \frac{p(X = 1 / Y = 1) * p(Y = 1)}{p(X = 1 / Y = 1) * p(Y = 1) + p(X = 1 / Y = 0) * p(Y = 0)}$$

Donc selon les données :

$$p(Y = 1 / X = 1) = \frac{0.8 * 0.01}{0.8 * 0.01 + 0.096 * 0.99}$$

Avec $p(Y=0) = 1.0 - p(Y=1) = 1.0 - 0.01 = 0.99$

$$p(Y = 1 / X = 1) = 0.0776$$

Donc la probabilité qu'une femme dans la quarantaine ayant passé ce test de routine avec un résultat positif soit effectivement atteinte d'un cancer du sein est de 7.76% donc moins de 10% (choix F).

Donc les médecins ont mal estimé cette probabilité puisqu'ils ont oublié, dans leurs calculs, de tenir compte de l'a priori que "La probabilité d'un cancer du sein chez les femmes dans la quarantaine participant à un test de routine (mammographie) est de 1%."

Section 2 Estimation de densité paramétrique Gaussienne vs estimation de densité par fenêtres de Parzen

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

P1. Entraînement des 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.

1. Les moyennes de chaque $x^{(i)}$ dans un vecteur $\mu = \{\mu^{(1)}, \mu^{(2)}, \dots, \mu^{(d)}\}$ de dimension d, donc d paramètres $\mu^{(i)}$.
2. Puisqu'on considère une Gaussienne isotropique, soit que la variance pour tous les traits/dimensions est la même, donc cette variance scalaire de dimension 1 σ^2 est l'autre paramètre à estimer.

(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)

Pour chaque μ_i du vecteur μ des moyennes des traits :

$$\mu_i = \frac{1}{n} \sum_{j=1}^n x_i^{(j)}$$

et sous forme vectorielle :

$$\mu = (\mu_1, \dots, \mu_d)$$

et pour la variance σ^2 :

$$\sigma^2 = \frac{\sum_i (x^{(i)} - \mu)^T (x^{(i)} - \mu)}{n * d}$$

(c) Quelle est la complexité algorithmique de cet apprentissage (entraînement) c.a.d. du calcul de ces paramètres ?

$$O(nd)$$

(d) Pour un point de test x , écrivez la fonction qui donnera la densité de probabilité prédite au point x :

$$\hat{p}_{\text{gauss-isotrop}}(x) = \frac{1}{(2\pi)^{d/2} * \sigma^d} * e^{\frac{-|x-\mu|^2}{2\sigma^2}}$$

(e) Quelle est la complexité algorithmique pour le calcul de cette prédiction à chaque nouveau point x ?

$$O(d)$$

P2. Estimation de la densité de la probabilité des fenêtres de Parzen avec un noyau Gaussien isotropique de largeur (écart-type) σ , avec ces fenêtres de Parzen entraînées sur D.

(a) Soit que σ est fixé. En quoi consiste la phase « entraînement/apprentissage » pour ces fenêtres de Parzen ?

Elle consiste :

1. À estimer la moyenne μ_i de chaque dimension/trait ($i = 1$ à d) à partir des exemples/données $D = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ où $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)}\}$ soit d μ_i calculées par l'équation :

$$\mu_i = \frac{1}{n} \sum_{j=1}^n x_i^{(j)}$$

Ce qui donne le vecteur de moyenne

$$\mu = (\mu_1, \mu_2, \mu_3, \dots, \mu_d)$$

2. À stocker tous les points de l'ensemble d'entraînement pour calculer, de façon pondérée selon les $p(x)$, la classe majoritaire qui sera associée au point test x selon ses plus proches voisins de l'ensemble d'entraînement.

(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 :

$$\hat{p}_{Parzen}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{(2\pi)^{d/2} \sigma^d} e^{\frac{-|x-\mu|}{2\sigma^2}}$$

(c) Quelle est la complexité algorithmique pour le calcul de cette prédiction à chaque nouveau point x ?

$$O(d)$$

P3. Capacité

(a) Laquelle de ces 2 approches (paramétrique Gaussienne v.s. Parzen à noyau Gaussien) a la plus forte capacité ? Expliquez.

L'approche Parzen à noyau Gaussien est basée sur une moyenne de Gaussienne comportant " d " moyennes μ_i estimées des données, un hyper paramètre, la variance σ^2 , fixe et l'ensemble des données d'entraînement qui sont aussi des paramètres du modèle.

L'approche paramétrique Gaussienne implique une seule Gaussienne comportant " d " moyennes μ_i et une variance σ^2 toutes estimées des données.

L'approche Parzen à noyau Gaussien a donc une plus forte capacité puisqu'elle comporte beaucoup plus de paramètres, soit les d moyennes de chaque dimension μ_i et toutes les données d'entraînement. Elle est donc plus riche, plus flexible et comporte un degré de liberté plus élevé que l'approche Gaussienne.

(b) Avec laquelle de ces approches, et dans quel cas précis, a-t-on toutes les chances d'être en sur-apprentissage ?

L'approche Parzen à noyau Gaussien lorsque l'hyper paramètre variance est fixé trop élevé par rapport aux distances entre les points d'entraînement. À la limite, si celle-ci est trop grande, elle englobera tous les points d'entraînement et seule la classe majoritaire de l'ensemble d'entraînement sera choisie pour tous les points tests expérimentés.

(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 ?

Parce que le σ des fenêtres de Parzen est fixé avant l'apprentissage et il ne peut être optimisé qu'au moyen d'un ou plusieurs ensemble(s) de validation.

D'un autre côté, le σ de l'approche densité paramétrique Gaussienne est considéré comme un paramètre puisqu'il est appris et optimisé à l'apprentissage sur l'ensemble d'entraînement.

P4. 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 dimensions.

$$p(x) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Où d = dimension du vecteur d'entrée x

Et les paramètres sont :

μ : le vecteur colonne de dimension d contenant les moyennes μ_i ($i = 1$ à d) des données d'entraînement de chaque trait/dimension i ;

et Σ est la matrice de covariance de dimension $d \times d$ des données d'entraînement, et dans le cas d'une Gaussienne **diagonale** :

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \\ & & & \sigma_d^2 \end{pmatrix}$$

Où seules les variances σ_i^2 sur la diagonale ne sont pas nulles et toutes les autres $\sigma_{i,j}^2 = 0$; il y a donc d σ_i^2 dans la matrice des covariances Σ .

Et

$|\Sigma|$ est le déterminant de la matrice de covariances.

(b) Démontrez que les composantes d'un vecteur aléatoire qui suit une distribution Gaussienne diagonale sont des variables aléatoires indépendantes.

Dans le cas d'une distribution Gaussienne **diagonale**, toutes les covariances $\sigma_{i,j}^2$ entre chacune des variables aléatoires x_i et x_j composant le vecteur aléatoire x sont nulles ce qui implique que toutes ces variables aléatoires x_i et x_j sont mutuellement décorréées entre elles.

La démonstration qui suit tirée de <http://cs229.stanford.edu/section/gaussians.pdf> pour un exemple à deux variables distribuées selon une Gaussienne diagonale multivariée:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

En introduisant ces variables et ces paramètres dans l'équation de la Gaussienne, on obtient:

$$\begin{aligned} p(x; \mu, \Sigma) &= \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{1/2}} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right) \\ &= \frac{1}{2\pi(\sigma_1^2 \cdot \sigma_2^2 - 0 \cdot 0)^{1/2}} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right), \end{aligned}$$

En simplifiant cette dernière équation, on obtient :

$$\begin{aligned} p(x; \mu, \Sigma) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2}(x_1 - \mu_1) \\ \frac{1}{\sigma_2^2}(x_2 - \mu_2) \end{bmatrix} \right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 \right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp \left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right). \end{aligned}$$

Or cette dernière équation est en fait la multiplication des deux densités Gaussiennes séparées des variables aléatoires x_1 et x_2 qui composent le vecteur aléatoire X . Donc ceci devient :

$$p(x; \mu, \Sigma) = p(x_1, x_2; \mu, \Sigma) = p(x_1; \mu_1, \sigma_1) * p(x_2; \mu_2, \sigma_2)$$

Or on sait que 2 variables aléatoires x_1 et x_2 sont indépendantes si et seulement si

$$P(X_1, X_2) = P(X_1) * P(X_2)$$

Ce qui correspond à l'équation qui précède.

Ceci peut être généralisé à toutes les variables aléatoires x_i d'un vecteur aléatoire de dimension d distribué selon une Gaussienne diagonale.

Donc les variables aléatoires d'un vecteur aléatoire distribué selon une Gaussienne diagonale sont 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)

$$\theta^* = \arg \min_{\theta} \hat{R}(p_{\theta}, D_n) = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n (-\log(p(x^{(i)}|\theta)))$$

$$\theta^* = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \left(-\log \left(\frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-u)^T \Sigma^{-1} (x-u)} \right) \right)$$

(d) Résolvez cette équation de manière analytique pour obtenir les paramètres optimaux.

Il s'agit donc de dériver (dérivées partielles) la fonction de coût $l(\theta)$ par rapport aux paramètres μ et Σ de la distribution Gaussienne diagonale des données et de mettre celle-ci à 0 :

$$\frac{\partial l(\theta)}{\partial \theta} = 0$$

En premier, on dérive par rapport au vecteur des moyennes : $\mu = (\mu_1, \dots, \mu_d)$

$$\frac{\partial l(\theta)}{\partial (\mu_1, \dots, \mu_d)} = 0$$

Puisqu'il s'agit, pour les moyennes μ_k ($k = 1$ à d), de la même situation que le cas de la Gaussienne isotropique démontré au cours après une longue dérivation, on obtient les paramètres optimaux pour μ :

$$\mu_k = \frac{1}{n} \sum_{i=1}^n x_k^{(i)}$$

Et la moyenne empirique

$$\mu = \frac{1}{n} \sum_{i=1}^n x^{(i)}$$

Par la suite, on dérive par rapport à Σ :

$$\frac{\partial l(\theta)}{\partial \Sigma} = 0$$

Or nous considérons une distribution Gaussienne diagonale pour laquelle, comme démontré au point P4 (b) plus haut, les variables aléatoires x_i du vecteur aléatoire x sont indépendantes et les $\sigma_{i,j} = 0$. Donc cette équation peut être appliquée pour déterminer indépendamment les σ_i de chaque variable x_i indépendante du vecteur x , ces σ_i étant sur la diagonale de la matrice de covariance Σ .

Donc ceci se réduit à trouver, pour chaque σ_i de Σ :

$$\frac{\partial l(\theta)}{\partial \sigma_i} = 0$$

Ceci donne l'équivalent de l'équation présentée aux notes de cours pour le cas isotropique mais considérant σ_i pour les données $x_i^{(j)}$ pour chaque dimension i de 1 à d . On obtient alors :

$$\sigma_{i \text{ max vrais}}^2 = \frac{\sum_{j=1}^n (x_i^{(j)} - \mu_i)^T (x_i^{(j)} - \mu_i)}{n}$$

P5. Problème de classification avec $D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$. Construire un classifieur de Bayes en se basant sur l'une ou l'autre de ces 3 estimations de densité : soit Gaussienne isotropique (variance σ^2 la même pour toutes les dimensions), Parzen à noyau Gaussien isotropique (variance σ^2 fixée avant l'apprentissage) et Gaussienne diagonale (variances $\sigma_{i,j}^2 = 0$ lorsque $i \neq j$ et $\sigma_i^2 \neq 0$).

(a) Écrivez dans vos propres mots comment vous apprendriez un classifieur de Bayes (phase d'entraînement)

L'apprentissage du classifieur de Bayes se fait par l'apprentissage/l'estimé des paramètres de la distribution Gaussienne qui modélise la distribution des données du problème au moyen des données d'entraînement, soit :

1. La moyennes μ_i de chaque dimension/trait pour le vecteur μ de dimension d , soit celle de l'entrée du système, pour les 3 estimations,
2. La variance σ^2 pour l'estimation par la Gaussienne isotropique pour l'ensemble des données d'entraînement et
3. Les variances σ_i^2 de la covariance diagonale Σ pour l'estimation par la Gaussienne diagonale pour chaque dimension i à partir des données d'entraînement de chaque dimension,

Le tout par l'application du principe de maximum de vraisemblance. Pour chaque classe, on calcule les valeurs de ces paramètres qui maximisent la log-vraisemblance des données d'entraînement de cette classe.

On notera que la variance pour le Parzen à noyau Gaussien n'est pas calculée puisqu'il s'agit d'un hyper-paramètre fixé avant l'apprentissage.

Le calcul de la moyenne empirique de chaque dimension/trait k maximisant la log-vraisemblance des données d'entraînement de cette dimension/trait se fait, pour les 3 estimations, au moyen de l'équation suivante :

$$\mu_k = \frac{1}{n} \sum_{i=1}^n x_k^{(i)}$$

La moyenne empirique de chaque dimension/trait k est une composante du vecteur μ :

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n x^{(i)}$$

$$\hat{\mu}_{ML} = (\mu_1, \dots, \mu_d)$$

Le calcul de la variance σ^2 pour la Gaussienne isotropique pour l'ensemble des données d'entraînement se fait par l'équation suivante sur l'ensemble des données d'entraînement:

$$\hat{\sigma}_{\max \text{ vrais}}^2 = \frac{\sum_i (x_i - \mu)^T (x_i - \mu)}{n * d}$$

Le calcul des variances σ_i^2 de chaque dimension/trait de la covariance diagonale Σ de la Gaussienne diagonale se fait par l'équation suivante appliquée pour chaque trait/dimension d'entraînement:

$$\hat{\sigma}_{i \max \text{ vrais}}^2 = \frac{\sum_{j=1}^n (x_i^{(j)} - \mu_i)^T (x_i^{(j)} - \mu_i)}{n}$$

(b) Pour un point de test x , écrivez, la fonction qui donnera le vecteur de probabilité prédite pour chaque classe au point x :

Cas Gaussienne isotropique :

$$g(x) = \frac{1}{(2\pi)^{d/2} \sigma^d} e^{-\frac{|x-\mu|^2}{2\sigma^2}} * p_{\text{à priori}}(x)$$

Cas Gaussienne diagonale :

$$g(x) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} e^{-\frac{(x-\mu)^T (x-\mu)}{2}} * p_{\text{à priori}}(x)$$

Cas Parzen à noyau Gaussien isotropique :

$$g(x) = \frac{1}{n} \sum \left(\frac{1}{(2\pi)^{d/2} \sigma^d} e^{-\frac{|x-\mu|^2}{2\sigma^2}} \right) * p_{\text{à priori}}(x)$$

Section 3 Partie pratique : estimation de densité

P1. 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 `compute_predictions` qui calcule les log de densité.

P2. 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 `compute_predictions` qui calcule les log de densité.

P3. Densités 1D : Parmi l'ensemble de données Iris, choisissez le sous-ensemble correspondant à une des classes (de votre choix), et un des trait 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.

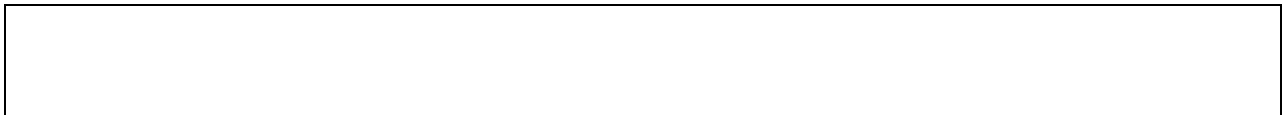
P4. Densités 2D : Ajoutez maintenant un second trait caractéristique d'iris, afin d'avoir des entrées en dimension $d = 2$ et produisez 4 graphiques, chacun affichant les points du sous-ensemble de données (avec la fonction `plot`), et les lignes de contours de la densité estimée (à l'aide de la fonction `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.



Section 4 Partie pratique : classifieur de Bayes

P1. Mélangez les exemples de Iris (utilisez `numpy.random.shuffle` après avoir initialisé le générateur aléatoire comme suit `numpy.random.seed(123)`). Puis divisez l'ensemble de tous les exemples en 2 : un ensemble d'entraînement, et un ensemble de validation.

Préparez deux versions de chacun de ces ensembles : une version complète comportant les $d = 4$ traits caractéristiques. Et une version avec seulement les $d = 2$ premiers traits caractéristiques qu'on utilisera pour fins de visualisation.

P2. Classifieur de Bayes basé sur des densités paramétriques Gaussiennes diagonales.

(a) Implémentez l'algorithme de classifieur de Bayes basé sur des densités paramétriques Gaussiennes diagonales.

(b) Visualisation en dimension $d = 2$. Considérant seulement les deux premiers traits caractéristiques d'iris, entraînez votre classifieur de Bayes sur votre ensemble d'entraînement ; affichez un graphique avec les régions de décision obtenues (ainsi que les points des ensembles d'entraînement et de validation).

(c) Calcul des erreurs en dimension $d = 2$: calculez et affichez le taux d'erreur de votre classifieur (entraîné sur les 2 premiers traits caractéristiques), à la fois sur l'ensemble d'entraînement et de validation.

(d) Calcul des erreurs en dimension $d = 4$: Entraînez votre classifieur en utilisant tous les traits caractéristiques. Puis calculez et affichez le taux d'erreur de votre classifieur (entraîné sur tous les traits caractéristiques), à la fois sur l'ensemble d'entraînement et de validation.

P3. Classifieur de Bayes basé sur des densités de Parzen avec noyau Gaussien isotropique

(a) Implémentez l'algorithme de classifieur de Bayes basé sur des densités de Parzen avec noyau Gaussien isotropique.

(b) Visualisation en dimension $d = 2$. Considérant seulement les deux premiers traits caractéristiques d'iris, entraînez votre classifieur de Bayes sur votre ensemble d'entraînement ; affichez un graphique avec les régions (surface) de décision obtenues (ainsi que les points des ensembles d'entraînement et de validation). Produisez 3 tels graphiques de régions de décision : avec un σ trop petit, trop grand, et approprié.

(c) Courbes d'apprentissage avec $d = 2$. Calculez le taux d'erreur de classification, à la fois sur l'ensemble d'entraînement, et sur l'ensemble de validation, en fonction des valeurs de l'hyper-paramètre σ (calculez-les pour une centaine de valeurs différentes de cet hyper-paramètre afin de pouvoir afficher la courbe). Indiquez la meilleure valeur de l'hyper-paramètre σ que vous avez trouvé.

(d) Courbes d'apprentissage avec $d = 4$. On va maintenant utiliser tous les traits caractéristiques. Calculez le taux d'erreur de classification, à la fois sur l'ensemble d'entraînement, et sur l'ensemble de validation, en fonction des valeurs de l'hyper-paramètre σ (calculez-les pour une centaine de valeurs différentes de cet hyper-paramètre afin de pouvoir afficher la courbe). Indiquez la meilleure valeur de l'hyper-paramètre σ que vous avez trouvé.

4. D'après ces expériences, pour le problème de classification d'Iris (et pour cette division entraînement/validation particulière) indiquez quel est le meilleur choix d'algorithme entre classifieur de Bayes avec Gaussiennes diagonales, et fenêtres de Parzen, et des autres hyper-paramètres : dimension de l'entrée (2 ou 4), et σ (s'il y a lieu). Précisez les taux d'erreurs de classification qu'ils permettent d'atteindre.