IFT3395 - Fondements de l'apprentissage machine

Travail Pratique 1

Nom: Paul CHAFFANET
Matricule: 1009543

Nom: Émile Labbé
Matricule: 20019813

1. Petit exercice de probabilités

Définissons deux variables aléatoires X et Y tels que:

$$X = \begin{cases} 0 & \text{si une femme n'a pas le cancer du sein} \\ 1 & \text{sinon} \end{cases}$$

$$Y = \begin{cases} 0 & \text{si le test est négatif} \\ 1 & \text{sinon} \end{cases}$$

On tire du texte que:

- la probabilité d'un cancer du sein chez les femmes dans la quarantaine participant à un test de routine (mammographie) est de 1% : $P(X=1)=0.01$ => $P(X=0)=0.99$
- si une femme est atteinte d'un cancer du sein, il y a 80% de chances que le test soit positif: $P(Y=1 \mid X=1)=0.8$
- chez une femme qui n'est pas atteinte de cancer, il y a une probabilité de 9,6% que le test soit positif : $P(Y=1 \mid X=0)=0.096$

Une femme dans la quarantaine ayant passé ce test de routine reçoit un résultat de test positif. On recherche la probabilité qu'elle soit effectivement atteinte d'un cancer du sein. Ainsi, on recherche la valeur de $P(X=1 \mid Y=1)$.

D'après la règles de Bayes:

$$\begin{aligned} P(X=1 \mid Y=1) &= \frac{P(Y=1 \mid X=1)P(X=1)}{P(Y=1)} \\ &= \frac{P(Y=1 \mid X=1)P(X=1)}{P(Y=1 \mid X=1)P(X=1)+P(Y=1 \mid X=0)P(X=0)} \\ &= \frac{0.8*0.01}{0.8*0.01+0.096*0.99} \\ &= 0.07764 \end{aligned}$$

Ainsi il y a $7,764\%$ de chances qu'une femme recevant un test positif pour le cancer du sein soit effectivement atteinte du cancer du sein. La probabilité est ainsi inférieure à 10%.

2. Estimation de densité : paramétrique Gaussienne, v.s. fenêtres de Parzen

1.(a) Les paramètres d'une densité gaussienne isotropique sur D sont la moyenne $\mu \in \mathbb{R}^d$ (de dimension d) et $\sigma^2 \in \mathbb{R}$ (de dimension 1).

1.(b) $\hat{\mu}_{MV} = \frac{1}{n}\sum_{i=1}^{n} x(i)$ et $\widehat{\sigma^2}_{MV} = \frac{1}{nd}\sum_{i=1}^{n} (x(i)-\hat{\mu})^T (x(i)-\hat{\mu})$ sont les formules donnant la valeur des paramètres optimaux par le principe du maximum de vraisemblance.

1.(c) Pour la calcul de la moyenne empirique, nous effectuons d sommes pour l'addition du vecteur $x(i)$. Étant donné que nous disposons de n vecteurs, on peut dire que le calcul s'effectue en $\mathrm{O}(nd)$.

Pour le calcul de la variance, celui-ci s'effectue en $\mathrm{O}(nd)$ car nous effectuons $2d$ soustractions, d multiplications et d sommes à chaque passage de boucle (afin de réduire notre variance à un scalaire).

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

1.(e) Le complexité est de $\mathrm{O}(d)$.

2.(a) La phase d'apprentissage consiste à mémoriser l'ensemble des données d'entraînements afin de pouvoir les utiliser lors de la prédiction d'un nouveau point.

2.(b) $\hat{p}_{mathrmParzen}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{1}{2}\frac{\left\|x-x(i)\right\|^2}{\sigma^2}}$

2.(c) À chaque nouveaux point à prédire, on doit effectuer d soustractions et ceci n fois, donc la complexité du calcul de la prédiction est de $\mathrm{O}(nd)$.

3.(a) Il n'y a que deux paramètres non libres dans l'approche paramétrique Gaussienne (entièrement déterminé par l'ensemble de données après l'entraînement) alors que les fenêtres de Parzen dispose de n variables déterminées par les données et de l'hyperparamètre σ qui permet de faire varier la largeur des noyaux. Ainsi, les fenêtres de Parzen dispose d'un ensemble de fonctions plus riches et donc l'approche par la paramétrique Gaussienne présente une plus faible capacité que les fenêtres de Parzen.

3.(b) Puisque que les fenêtres de Parzen présentent une plus grande capacité, cette approche sera plus sujette au surapprentissage que l'approche gaussienne paramétrique. Avec un hyperparamètre σ très petit, la largeur des noyaux se trouvent réduite pour les fenêtres de Parzen et donc on risque d'être en situation de surapprentissage. Cette modélisation est mauvaise car elle est sensible au bruit de l'ensemble de données.

3.(c) Dans le cas de l'approche paramétrique gaussienne, le σ est calculé lors de la phase d'entraînement. Dans le cas des fenêtres de Parzen, le σ est fixé par l'utilisateur. Ainsi, on parle de paramètre pour σ et d'hyperparamètre pour σ de Parzen.

4.(a) L'équation d'une densité gaussienne diagonale dans $\mathbb{R}^d$ est:

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

où les paramètres sont la moyenne $\mu \in \mathbb{R}^d$ et la matrice de variance-covariance diagonale $\Sigma \in \mathbb{R}^{d \times d}$.

4.(b) Soit le vecteur aléatoire $x = \begin{pmatrix} x_1 & x_2 & \cdots & x_d \end{pmatrix}$. On sait que notre matrice de variance-covariance diagonale est:

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

Comme c'est une matrice diagonale, ses propriétés nous permettent d'exprimer $|\Sigma|$ et Σ^{-1} facilement:

$$|\Sigma| = \sigma_1^2 \sigma_2^2 ... \sigma_d^2$$

$$\Sigma^{-1} = \begin{pmatrix} 1/\sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_d^2 \end{pmatrix}$$

Nous allons séparer l'expression $p(x)$ en deux parties. Les première partie est:

$$\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} = \frac{1}{\left(\prod_{i=1}^{d}(2\pi)^{1/2}\right)\sqrt{\sigma_1^2 \sigma_2^2 ... \sigma_d^2}} = \frac{1}{\left(\prod_{i=1}^{d}(2\pi)^{1/2}\right)\sqrt{(\sigma_1 \sigma_2 ... \sigma_d)^2}}$$

$$= \frac{1}{\left(\prod_{i=1}^{d}(2\pi)^{1/2}\right)\sigma_1 \sigma_2 ... \sigma_d} = \frac{1}{\prod_{i=1}^{d}(2\pi)^{1/2}\sigma_i}$$

La deuxième partie est:

$$(x-\mu)^T \Sigma^{-1}(x-\mu) = \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_d-\mu_d \end{pmatrix} \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & 0 \\ 0 & 1/\sigma_1^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\sigma_1^2 \end{pmatrix} \begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 & \cdots & x_d-\mu_d \end{pmatrix}$$

$$= \begin{pmatrix} \frac{(x_1-\mu_1)}{\sigma_1^2} \\ \frac{(x_2-\mu_2)}{\sigma_2^2} \\ \vdots \\ \frac{(x_d-\mu_d)}{\sigma_d^2} \end{pmatrix} \begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 & \cdots & x_d-\mu_d \end{pmatrix} = \sum_{i=1}^{d} \frac{(x_i-\mu_i)^2}{\sigma_i^2}$$

En fusionnant ces deux parties, on obtient que:

$$p(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} = \frac{1}{\prod_{i=1}^{d}(2\pi)^{1/2}\sigma_i} e^{-\frac{1}{2}\sum_{i=1}^{d}\frac{(x_i-\mu_i)^2}{\sigma_i^2}}$$

$$p(x) = \frac{1}{\prod_{i=1}^{d}(2\pi)^{1/2}\sigma_i} \prod_{i=1}^{d} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}}$$

$$p(x) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}}$$

Ainsi, les composantes d'un vecteur aléatoire x constituent bien des variables aléatoires indépendantes dans le cas d'une densité gausienne diagonale.

4.(c) L'équation qui correspondrait à la minimisation du risque empirique sur l'ensemble d'entrainement D (pour apprendre les paramètres) est :

$$J(\theta) = \hat{R}(f_\theta, D) = \frac{1}{n}\sum_{i=1}^{n} -\log\big(p(x(i))\big)$$

Et l'on cherche donc le vecteur de paramètres optimal $\theta^* = \arg\min_\theta J(\theta)$, $\theta = (\mu, \Sigma)$

4.(d) $$J(\theta) = \hat{R}(f_\theta, D) = \frac{1}{n}\sum_{i=1}^{n} -\log\big(p(x(i))\big) = \frac{1}{n}\sum_{i=1}^{n} -\log\left(\prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\frac{\left(x_j(i)-\mu_j\right)^2}{\sigma_j^{\,2}}}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} -\sum_{j=1}^{d} -\frac{1}{2}\log(2\pi) - \log\left(\sigma_j\right) - \frac{1}{2}\frac{\left(x_j(i)-\mu_j\right)^2}{\sigma_j^{\,2}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \frac{1}{2}\log(2\pi) + \log\left(\sigma_j\right) + \frac{1}{2}\frac{\left(x_j(i)-\mu_j\right)^2}{\sigma_j^{\,2}}$$

Afin de minimiser $J(\theta)$, on calcule ses dérivées partielles:

$$\frac{dJ}{d\mu} = 0$$

$$\equiv \frac{d}{d\mu}\left(\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{x_j^{\,2}(i) + \mu_j^{\,2} - 2x_j(i)\mu_j}{\sigma_j^{\,2}}\right) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{2\mu_j}{\sigma_j^{\,2}} - \frac{2x_j(i)}{\sigma_j^{\,2}}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{\mu_j}{\sigma_j^{\,2}} - \frac{x_j(i)}{\sigma_j^{\,2}} = 0$$

$$\equiv \sum_{i=1}^{n}\sum_{j=1}^{d}\frac{\mu_j}{\sigma_j^{\,2}} = \sum_{i=1}^{n}\sum_{j=1}^{d}\frac{x_j(i)}{\sigma_j^{\,2}}$$

$$\equiv \sum_{i=1}^{n}\sum_{j=1}^{d}\mu_j = \sum_{i=1}^{n}\sum_{j=1}^{d}x_j(i)$$

$$\equiv n\sum_{j=1}^{d}\mu_j = \sum_{i=1}^{n}\sum_{j=1}^{d}x_j(i)$$

$$\equiv \sum_{j=1}^{d}\mu_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}x_j(i)$$

$$\equiv \mu_j^{\,*} = \frac{1}{n}\sum_{i=1}^{n}x_j(i)$$

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

$$\frac{dJ}{d\sum} = 0$$

$$\equiv \frac{d}{d\sum}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\log\left(\sigma_j\right) + \frac{1}{2}\frac{x_j(i)^2 + \mu_j{}^2 - 2x_j(i)\mu_j}{\sigma_j{}^2} \right) = 0$$

$$\equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{1}{\sigma_j} - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{x_j(i)^2 + \mu_j{}^2 - 2x_j(i)\mu_j}{\sigma_j{}^3} = 0$$

$$\equiv \sum_{i=1}^{n}\sum_{j=1}^{d}\frac{1}{\sigma_j} = \sum_{j=1}^{d}\sum_{i=1}^{n}\frac{x_j(i)^2 + \mu_j{}^2 - 2x_j(i)\mu_j}{\sigma_j{}^3}$$

$$\equiv n\sum_{j=1}^{d}\frac{1}{\sigma_j} = \sum_{j=1}^{d}\frac{1}{\sigma_j{}^3}\sum_{i=1}^{n}x_j(i)^2 + \mu_j{}^2 - 2x_j(i)\mu_j$$

$$\equiv \sum_{j=1}^{d}\sigma_j{}^2 = \frac{1}{n}\sum_{j=1}^{d}\sum_{i=1}^{n}x_j(i)^2 + \mu_j{}^2 - 2x_j(i)\mu_j$$

$$\equiv \sigma_j{}^{2^*} = \frac{1}{n}\sum_{i=1}^{n}(x_j(i) - \mu^*{}_j)^2$$

qui seront les valeurs situées sur la diagonale $\sum_{jj}$.

5.(a) Soit $y(i) = c \in \{1,...,m\}$ la classe d'un vecteur $x(i)$. Nous allons séparer l'ensemble D en m sous-ensembles qui contient des vecteurs $x(i)$ étiquettés par la même et unique classe. Nous allons ensuite entraîner (optimiser les paramètres) notre estimateur de densité sur chaque classe de l'ensemble c :

$$\widehat{p_c}(x) \simeq P(X = x \mid Y = c)$$

On sait comment optimiser nos paramètres pour les 3 méthodes (questions (1), (2), (4)), donc cette étape ne devrait pas poser de soucis. Ensuite, il est également nécessaire de déterminer les probabilités à priori pour chaque classe:

$$\widehat{P}_c = \frac{n_c}{n} \simeq P_Y(c) = P(Y = c)$$

5.(b) Notre vecteur de fonctions discriminantes au point x est:

$$g(x) = \left(P(Y = 1 \mid X = x), \ldots, P(Y = m \mid X = x)\right)$$

$$g(x) \simeq \left(\frac{\hat{p}_1(x)\widehat{P}_1}{\sum_{c'=1}^{m} \hat{p}_{c'}(x)\widehat{P}_{c'}}, \ldots, \frac{\hat{p}_m(x)\widehat{P}_m}{\sum_{c'=1}^{m} \hat{p}_{c'}(x)\widehat{P}_{c'}} \right) .$$

On choisira alors la composante du vecteur g(x) qui présente la plus haute probabilité.