# IFT3395 Devoirs 1

Professur: Aaron Courville

Boumediene Boukharouba, Farzin Faridfar

October 12, 2017

## 1 Petit exercice de probabilités

atteinte: une femme atteinte d'un cancer du sein
positif: résultat de test positif

Bayes:

$$P(atteinte|positif) = \frac{P(positif|atteinte)P(atteinte)}{P(positif)}$$

résultat de test positif = positif parmi ce qui sont atteintes + positif et parmi ce qui ne sont pas atteintes

$$\begin{aligned} P(positif|atteinte) &= 0.8 \\ P(atteinte) &= 0.01 \\ P(positif) &= 0.8 \times 0.01 + 0.096 \times (1 - 0.01) \\ &= 0.103 \end{aligned}$$

$$\begin{aligned} P(atteinte|positif) &= \frac{P(positif|atteinte)P(atteinte)}{P(positif)} \\ &= \frac{0.8 \times 0.01}{0.103} \\ &= 0.0777 = 7.77\% \end{aligned}$$

Comme le résultat montre il faut au moins un cours de probabilité pour les médecins pour qu'ils ne fassent pas peur aux pauvres patientes. Leur choix de réponse est à cause qu'ils calculent la probabilité d'un test positif sashant que la patiente est atteinte d'un cancer.

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

## 2.1 Gaussien isotropique

(a) La moyenne: $\mu \in \mathbb{R}^{\mathrm{d}}$ et la variance: $\sigma^2 \in \mathbb{R}$
(Dans le cas de Gaussien isotropique, la matrice de covariance ($\Sigma$) est diagonale avec la même valeur sur sa diagonale, car on considère que les variances sont égales dans toute les dimensions. Donc, le seul paramètre est la variance elle-même)

(b) On considère que $|D| = n$ et que la dimension des données est égal à $d$. On a:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\sigma^2 = \frac{1}{nd}\sum_{i=1}^{n}(x_i - \mu)^T(x_i - \mu) \tag{2}$$

Les valeurs optimum de chacun des paramètres (le moyen et la variance) peut être calculé en dérivant les équation (1) et (2) par rapport à $x_i$.

(c) $O(n.d)$ (Il faut calculer la moyenne et la variance pour chaque exemple de données (On en a $n$ fois) qui sont en $d$ dimensions)

(d)

$$\hat{P}_{gaus-isotrop}(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{(x_i-\mu)^T(x_i-\mu)}{2\sigma^2}}$$

(e) Il faut calculer la moyenne, la variance et le log de vraisemblance pour chacun des exemples de données. Donc la complexité est $O(m.d)$ où $m$ et la taille d'ensemble du test.

## 2.2 Fenêtre de Parzen avec un noyau Gaussien Isotropique

(a) Comme le $\sigma$ est fixé on a que calculer la moyenne et ensuite le log de vraisemblance pour trouver la classe appropriée. Ici, l'ensemble d'entraînements joue le rôle de paramètre qui doit être mémorisé pour la phase *"entraînement-apprentissage"* et il n'existe aucune phase d'optimisation des paramètre ou la sélection d'hyper-paramètre.

(b) Pour un estimateur Parzen avec le noyau Gaussien isotropique la largeur

de fenêtre de Parzen $h = \sigma$.

$$\hat{P}_{mathrmParzen}(x) = \frac{1}{n}\sum_{i=1}^{n} N_{X_i,\sigma}(x_i)$$

$$N_{X_i,\sigma}(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{(x_i-x)^T(x_i-x)}{\sigma^2}}$$

(c) Dans ce cass aussi, il faut calculer la moyenne et le log de vraisemblance pour chacun des exemples de données. Donc, la complexité est $O(m.d)$ où $m$ et la taille d'ensemble du test.

## 2.3 Capacité

(a) L'approche Parzen a la plus forte capacité. Dans cette approche la variance est une hyper-paramètre qui définie la rayon de fenêtre de Parzen et qui peut prendre une infinité de valeur. Donc il faut l'optimiser pour avoir la meilleur valeur pour cet hyper-paramètre. Alors que, dans l'approche Gaussienne, la variance est calculé en utilisant les données d'entraînements un degré de liberté moins que l'approche précédente.

(b) Approche Parzen. Comme dans cette approche on peut choisir la variance (qui a le rôle de la rayon de fenêtre de Parzen), il est toujours possible d'avoir une surface d'apprentissage plus précise. Cependant, le sigma choisi peut retourner une très bon résultat pour l'ensemble d'entraînements et il est très probable que cette valeur donne un mauvais résultat pour l'ensemble de validation, autrement dit le cas de sur-apprentissage. En choisissant les sigma trop petit, on a le risque d'être en sur-apprentissage.

(c) Parce que dans les fenêtre de Parzen, $\sigma$ est choisi comme une constante par l'utilisateur. Alors que, dans l'approche paramétrique Gaussienne, cette paramètre est choisi selon l'optimisation qui se fait dans le modèle en utilisant les données d'entraînements.

## 2.4 Densité Gussienne diagonale

(a) Le moyen: $\mu \in \mathbb{R}^{\mathrm{d}}$ et la matrice de covariance: $\Sigma \in \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$

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

(b)

(c)

$$\hat{R} = \frac{1}{n}\sum_{i=1}^{n} L(f(x), y)$$

$$L(f(x), y) = -\log p(x) \implies \hat{R} = -\sum_{i=1}^{n} \log p(x_i)$$

(d)

$$\begin{aligned}
\hat{R} &= -\sum_{i=1}^{n} \log p(x_i) \\
&= -\sum_{i=1}^{n} log\Big(\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} e^{(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}\Big) \\
&= -\sum_{i=1}^{n} \left( log\Big(\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}}\Big) + log\Big(e^{(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}\Big)\right) \\
&= -\sum_{i=1}^{n} \left( \log 1 - \log\left((2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}\right) + (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\
&= \sum_{i=1}^{n} \left( \frac{d}{2} log(2\pi) + \log(\sqrt{|\Sigma|}) - (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\
&= \frac{nd}{2}\log(2\pi) + n\log(\sqrt{|\Sigma|}) - \sum_{i=1}^{n}\Big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big)
\end{aligned}$$

Pour obtenir les paramètres optimaux,il faut résoudre les équations suivants:

$$\frac{\partial \hat{R}}{\partial \mu} = 0 \text{ et } \frac{\partial \hat{R}}{\partial \sigma} = 0$$

Pour la moyenne:

$$\begin{aligned}
\frac{\partial \hat{R}}{\partial \mu} &= 0 \\
&= \frac{\partial}{\partial \mu}\left( \frac{nd}{2} log(2\pi) + n\log(\sqrt{|\Sigma|}) - \sum_{i=1}^{n}\Big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big)\right) \\
&= 0 + 0 + \sum_{i=1}^{n} \frac{\partial}{\partial \mu}\Big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big)
\end{aligned}$$

On a $(f.g)' = f'.g + f.g'$. Donc:

$$\frac{\partial}{\partial \mu}\Big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big) = \Sigma^T\big(-(x_i-\mu)-(x_i-\mu)\big)$$
$$= 2\Sigma^T(x_i-\mu)$$

et ainsi:

$$\frac{\partial \hat{R}}{\partial \mu} = 0$$
$$\sum_{i=1}^{n} 2\Sigma^T(x_i-\mu) = 0$$
$$\sum_{i=1}^{n}(x_i-\mu) = 0$$
$$\sum_{i=1}^{n} x_i + n\mu = 0 \implies$$
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Pour la variance ou l'écart-type:

$$\frac{\partial \hat{R}}{\partial \sigma} = 0$$
$$= \frac{\partial}{\partial \sigma}\left(\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log(|\Sigma|) - \sum_{i=1}^{n}\Big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big)\right)$$
$$= 0 + \frac{\partial}{\partial \sigma}\big(\frac{n}{2}\log(|\Sigma|)\big) + \sum_{i=1}^{n}\frac{\partial}{\partial \sigma}\Big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big)$$

Pour $|\Sigma|$ on a:

$$|\Sigma| = \Pi_{j=1}^{d}\sigma^2 \implies \log|\Sigma| = \log\Pi_{j=1}^{d}\sigma_j^2 \quad (3)$$
$$= \sum_{j=1}^{d}(\log\sigma_j^2) \quad (4)$$

$$\frac{\partial}{\partial \sigma}\sum_{j=1}^{d}(\log\sigma_j^2) = \sum_{j=1}^{d}\frac{2}{\sigma_j} \text{ et pour un } \sigma_k \implies \frac{2}{\sigma_k} \quad (5)$$

$$\text{Aussi pour un } \sigma_k : \frac{\partial}{\partial \sigma}\Sigma^{-1} = \frac{2}{\sigma_k^3} \quad (6)$$

En replaçant les équations (5) et (6) on aura:

$$= \frac{2n}{\sigma} + \frac{2\sum_{i=1}^{n}\left((x_i-\mu)^T(x_i-\mu)\right)}{\sigma^3}$$

$$n\sigma^2 = \sum_{i=1}^{n}\left((x_i-\mu)^T(x_i-\mu)\right)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left((x_i-\mu)^T(x_i-\mu)\right)$$

### 2.5 Bayes

(a) Pour un classifieur de Bayes il faut d'abords calculer la probabilité de chaque classe dans notre ensemble d'entraînements ce qui nous donne la probabilité à priori de chaque classe.
Pour le faire, on divise les données d'entraînements selon leurs cible/classe et on crée un sous-ensemble pour chaque classe.
Ensuite, utilisant un noyau (e.g. Gaussien, Parzen, ...), on obtient la probabilité conditionnelle de chaque sous-ensemble.
Ayant dans la main ces deux valeurs (la probabilité à priori et la probabilité conditionnelle), on applique l'équation de Bayes pour obtenir la probabilité vraisemblance et finalement classifier les données d'entraînements.

(b)

$$P(c|x) = \frac{\hat{p}_c(x)\hat{P}_c}{\sum_{c'=1}^{m}\hat{p_{c'}}(x)\hat{P}_{c'}}$$

## 3 Partie pratique : estimation de densité

Pour exécuter le code vous pourriez taper, s'il vous plaît, sur un terminal:
`python3 main.py`

## 4 Partie pratique : classifieur de Bayes

Pour exécuter le code vous pourriez taper, s'il vous plaît, sur un terminal:
`python3 main.py`