HE VD HAUTE ÉCOLE D'INGÉNIERIE ET DE GESTION DU CANTON DE VAUD
IG

ISCD : **Apprentissage supervisé (APV)**	**TP 5**
S. Robert, F. Hêche	Novembre 2023

Le but de ce travail pratique est de construire la méthode de classification bayésienne naïve (Naïve Bayes Classifier).

Plus précsiément l'exercise 1 permet de comprendre la motivation théorique de cette méthode. L'exercice 2 consiste à appliquer cette approche dans le cas discret. Ensuite à la fin de la partie théorique, nous allons montrer comment nous pouvons adapter cette méthode dans le cas continu. Finalement, dans le jupyter notebook nous allons implémenter un Naïve Bayes Classifier dans le cas continu.

Exercice 1

Soit $\mathcal{X}$, un espace discret (par exemple $\mathcal{X} = \{0,1\}^k$) et $\mathcal{Y} := \{C_1, \dots C_m\}$ l'ensemble des classes possibles. Etant donné $x \in \mathcal{X}$, on aimerait estimer à quelle classe appartient x.

La méthode bayésienne revient à prédire la classe qui minimise le risque. Plus précisément, pour chaque classe C_j, on construit une fonction $L_j : \mathcal{Y} \to \mathbb{R}$ où $L_j(C_i)$ détermine le coût de prédire la classe C_j alors que la vrai classe est C_i. Ensuite, on définit le risque de prédire la classe C_j comme étant $r_j(x) := \mathbb{E}_{C_i \sim p(\cdot|x)}[L_j] = \sum_{i=1}^{m} L_j(C_i)p(C_i|x)$. Finalement, notre modèle h_θ devient

$$h_\theta(x) := \text{argmin}_{C_j \in \mathcal{Y}} \quad r_j(x)$$

Plus explicitement, on obtient

$$h_\theta(x) := \text{argmin}_{C_j \in \mathcal{Y}} \quad \mathbb{E}_{p(\cdot|x)}[L_j] = \text{argmin}_{C_j \in \mathcal{Y}} \quad \sum_{i=1}^{m} L_j(C_i)p(C_i|x) \tag{1}$$

En vous souvenant de la formule de Bayes, montrer que si on définit L_j comme $L_j(C_j) = 0$ et $L_j(C_i) = 1$ pour $i \neq j$ alors le modèle définit ci-dessus peut s'écrire sous la forme suivante

$$h_\theta(x) = \text{argmax}_{C_j \in \mathcal{Y}} \quad p(x|C_j)p(C_j)$$

Exercice 2

Reprenons le modèle construit à l'exercice 1

$$h_\theta(x) = \text{argmax}_{C_j \in \mathcal{Y}} p(x|C_j)p(C_j)$$

Le problème avec cette expression est qu'il est difficile d'estimer $p(x|C_i)$ en utilisant des données d'entraînement. En effet, il est très peu probable d'observer au moins une fois $x = (x_1, \dots x_k) \in \mathcal{X}$ dans notre jeux de données. Pour éviter ce problème, nous allons encore supposer que les composantes de x sont indépendantes entre elles. Ce qui nous permet d'écrire

$$p(x|C_j) = p((x_1, \dots x_k) \mid C_j) = \Pi_{i=1}^{k} p(x_i|C_j)$$

Ensuite,nous pouvons estimer les $p(x_i|C_j)$ en utilisant un dataset. Le modèle ainsi construit s'appelle 'Naïve Bayes Classifier'.

Utiliser cette approche ainsi que les données d'entraînement présentées dans le tableau ci-dessous, afin d'estimer si une fleur qui possède des sépales vertes et arrondies et de petites pétales jaunes est une iris versicolores ou une iris setosa.

Id	Couleur des sépales	Forme des sépales	Couleur des pétales	Taille des pétales	Iris
1	Bleue	Ovale	Violet	Grande	Versicolores
2	Bleue	Arrondie	Violet	Grande	Versicolores
3	Rouge	Ovale	Violet	Grande	Versicolores
4	Verte	Ovale	Jaune	Petite	Versicolores
5	Bleue	Ovale	Jaune	Grande	Versicolores
6	Verte	Ovale	Violet	Grande	Versicolores
7	Rouge	Arrondie	Jaune	Petite	Setosa
8	Rouge	Ovale	Violet	Petite	Setosa
9	Rouge	Arrondie	Jaune	Grande	Setosa
10	Bleue	Ovale	Jaune	Petite	Setosa
11	Verte	Arrondie	Violet	Petite	Setosa
12	Verte	Arrondie	Jaune	Grande	Setosa

Naïve Bayes dans le cas continu

Pour le moment, nous avons supposé que $\mathcal{X}$ était un espace discret. Nous allons maintenant adapter notre modèle dans le cas où $\mathcal{X}$ est un espace continu. L'idée de base est donc de construire un classificateur de la forme

$$h_\theta(x) = \textbf{argmax}_{C_j \in \mathcal{Y}} \quad p(C_j)\Pi_{i=1}^{k} p(x_i|C_j) \tag{2}$$

Cependant dans le cas continu, on a $p(x_i|C_j) = 0$.

Pour éviter ce problème, nous allons supposer que la variable aléatoire $X_i|C_j$ admet comme fonction de densité $f_{\theta_{i,j}}$. Et nous allons considérer le classificateur suivant

$$h_\theta(x) = \textbf{argmax}_{C_j \in \mathcal{Y}} \quad p(C_j)\Pi_{i=1}^{k} f_{\theta_{i,j}}(x_i) \tag{3}$$

Ce modèle peut être justifié de la manière suivante. Commençons par prendre un $\delta > 0$

suffisament petit et considérons les intervalles de la forme $[x_i - \frac{\delta}{2}, x_i + \frac{\delta}{2}]$. Ensuite, remarquons que

$$\begin{aligned} p\left(X_i \in [x_i - \frac{\delta}{2}; x_i + \frac{\delta}{2}] \mid C_j\right) &= \int_{x_i - \frac{\delta}{2}}^{x_i + \frac{\delta}{2}} f_{\theta_{i,j}}(x)dx \\ &\simeq \int_{x_i - \frac{\delta}{2}}^{x_i + \frac{\delta}{2}} f_{\theta_{i,j}}(x_i)dx \\ &= f_{\theta_{i,j}}(x_i)\left(x_i + \frac{\delta}{2} - x_i + \frac{\delta}{2}\right) \\ &= f_{\theta_{i,j}}(x_i)\delta \end{aligned}$$

Par conséquent, en appliquant la version discrète de Naïve Bayes aux intervalles $[x_i - \frac{\delta}{2}, x_i + \frac{\delta}{2}]$, nous trouvons que

$$\begin{aligned} h_\theta(x) &= \textbf{argmax}_{C_j \in \mathcal{Y}} \quad p(C_j)\Pi_{i=1}^k p\left(X_i \in [x_i - \frac{\delta}{2}; x_i + \frac{\delta}{2}] \mid C_j\right) \\ &\simeq \textbf{argmax}_{C_j \in \mathcal{Y}} \quad p(C_j)\Pi_{i=1}^k f_{\theta_{i,j}}(x_i)\delta \\ &= \textbf{argmax}_{C_j \in \mathcal{Y}} \quad \delta^k p(C_j)\Pi_{i=1}^k f_{\theta_{i,j}}(x_i) \\ &= \textbf{argmax}_{C_j \in \mathcal{Y}} \quad p(C_j)\Pi_{i=1}^k f_{\theta_{i,j}}(x_i) \end{aligned}$$

Cependant le produit de ces fonctions de densités générer des très petits nombres et donc de potentiels problèmes numériques. Pour gérer ce problème, en pratique nous prenons le log de l'expression trouvée ci-dessus. Notre modèle final dans le cas continu devient donc

$$h_\theta(x) = \textbf{argmax}_{C_j} \log(p(C_j)) + \sum_{i=1}^{n} \log(f_{\theta_{i,j}}(x_i))$$