Théorie bayésienne de la décision

Introduction à l'apprentissage automatique – GIF-4101 / GIF-7005

Professeur: Christian Gagné

Semaine 2

UNIVERSITÉ LAVAL

2.1 Formule de Bayes

Rappel sur les statistiques

- Expérience aléatoire ($\mathcal{E}$) : expérience dont l'issue n'est pas prévisible avec certitude à l'avance
- Univers (U) : ensemble des issues possibles d'une expérience
    - Univers discret : ensemble fini d'issues possibles
    - Univers continu : issues possibles non énumérables
- Événement aléatoire (A) : résultat d'une expérience aléatoire, sous-ensemble de l'univers ($A \subset U$)
- Probabilité ($P(A)$) : associe un nombre réel représentant une application d'un événement quelconque (A) lié à une expérience aléatoire ($A \subset U$), satisfaisant les axiomes des probabilités
    1. $0 \leq P(A) \leq 1, \forall A$
    2. $P(U) = 1$
    3. Supposons que les événements A_i, $i = 1, \ldots, n$ sont mutuellement exclusifs ($A_i \cap A_j = \emptyset, \forall j \neq i$), alors $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$

Probabilité et inférence

- Lancement d'une pièce de monnaie : $U = \{\text{pile,face}\}$
- Variable aléatoire $X = \{0, 1\}$ (0=face, 1=pile)
  - Distribution de Bernoulli : $P(x \in X) = (1 - p_1)^{1-x}\, p_1^x$
- Ensemble d'échantillons $\mathbf{X}$ tirés selon une distribution de probabilité paramétrée par p_1 (probabilité de pile)
  - Ensemble de N échantillons : $\mathbf{X} = \{x^t\}_{t=1}^{N}$ avec $x^t \in X$
  - Estimation de p_1 par échantillonnage : $\hat{p}_1 = \frac{\#pile}{\#tir} = \frac{\sum_{i=1}^{N} x^t}{N}$
- Prédiction du prochain tir x^{N+1} : si $\hat{p}_1 > 0{,}5$ alors pile, sinon face
- Exemple de tirage : $\mathbf{X} = \{1, 1, 1, 0, 1, 0, 0, 1, 1\}$
  - Estimation de la probabilité : $\hat{p}_1 = \frac{\sum_{t=1}^{N} x^t}{N} = \frac{6}{9}$

Classement

- Exemple de l'évaluation du risque au crédit
  - Données d'entrée : $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, avec x_1 comme le revenu et x_2 le niveau d'épargne
  - Classes possibles : $C \in \{0, 1\}$ où $C = 1$ dénote un individu à haut risque de défaut de paiement et $C = 0$ un individu à faible risque
- Si on connaît $P(C|x_1,x_2)$ alors :
  - Sélectionner : $\begin{cases} C = 1 & \text{si } P(C = 1|x_1,x_2) > 0{,}5 \\ C = 0 & \text{autrement} \end{cases}$
- Formulation équivalente :
  - Sélectionner : $\begin{cases} C = 1 & \text{si } P(C = 1|x_1,x_2) > P(C = 0|x_1,x_2) \\ C = 0 & \text{autrement} \end{cases}$

Probabilité conditionnelle

- Probabilité conditionnelle $P(E|F)$: probabilité que l'événement E se produit si l'événement F est survenu :

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

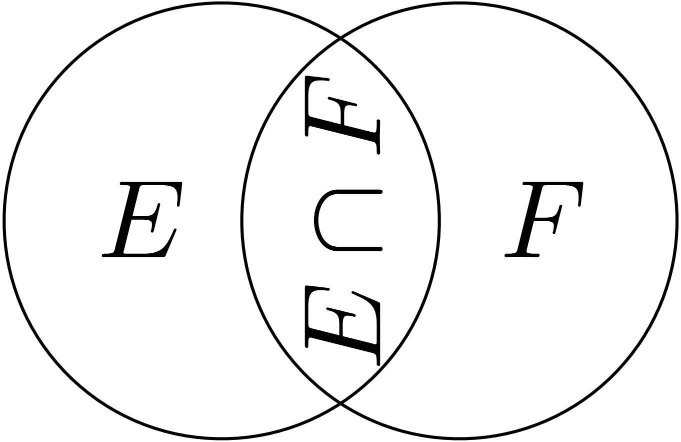
- Comme $\cap$ est commutatif :

$$P(E \cap F) = P(E|F)\,P(F) = P(F|E)\,P(E)$$

- Formule de Bayes :

$$P(F|E) = \frac{P(E|F)\,P(F)}{P(E)}$$

Diagramme de Venn et formule de Bayes



$$P(E \cap F) = P(E|F)\,P(F) = P(F|E)\,P(E) = P(F \cap E)$$

Formule de Bayes

$$\underbrace{P(C|\mathbf{x})}_{\text{a posteriori}} = \frac{\overbrace{P(C)}^{\text{a priori}}\ \overbrace{p(\mathbf{x}|C)}^{\text{vraisemblance}}}{\underbrace{p(\mathbf{x})}_{\text{évidence}}}$$

- Probabilité a priori ($P(C)$) : probabilité d'observer une instance de la classe C
- Vraisemblance de classe ($p(\mathbf{x}|C)$) : vraisemblance qu'une observation de la classe C soit $\mathbf{x}$
- Évidence ($p(\mathbf{x})$) : vraisemblance d'observer la donnée $\mathbf{x}$
- Probabilité a posteriori ($P(C|\mathbf{x})$) : probabilité qu'une observation $\mathbf{x}$ appartienne à la classe C

Formule de Bayes

$$\underbrace{P(C|\mathbf{x})}_{\text{a posteriori}} = \frac{\overbrace{P(C)}^{\text{a priori}}\ \overbrace{p(\mathbf{x}|C)}^{\text{vraisemblance}}}{\underbrace{p(\mathbf{x})}_{\text{évidence}}}$$

- Somme des probabilités a priori : $P(C=0) + P(C=1) = 1$
- Somme des probabilités a posteriori : $P(C=0|\mathbf{x}) + P(C=1|\mathbf{x}) = 1$
- Évidence : $p(\mathbf{x}) = P(C=1)\,p(\mathbf{x}|C=1) + P(C=0)\,p(\mathbf{x}|C=0)$

Exemple : formule de Bayes

- Observation de véhicules
  - Probabilité d'observer une voiture, $P(C = 1) = 0{,}7$
  - Probabilité d'observer un autre véhicule, $P(C = 0) = 0{,}3$
- Une certaine observation de véhicule $\mathbf{x}$
  - Vraisemblances de l'observation : $p(\mathbf{x}|C = 1) = 1{,}1$, $p(\mathbf{x}|C = 0) = 0{,}4$
- Évidence

$$
\begin{aligned}
p(\mathbf{x}) &= p(\mathbf{x}|C = 1)\, P(C = 1) + p(\mathbf{x}|C = 0)\, P(C = 0) \\
&= 1{,}1 \cdot 0{,}7 + 0{,}4 \cdot 0{,}3 = 0{,}77 + 0{,}12 = 0{,}89
\end{aligned}
$$

- Probabilités a posteriori

$$
\begin{aligned}
P(C = 1|\mathbf{x}) &= \frac{P(C = 1)\, p(\mathbf{x}|C = 1)}{p(\mathbf{x})} = \frac{0{,}7 \cdot 1{,}1}{0{,}89} = \frac{0{,}77}{0{,}89} = 0{,}865 \\
P(C = 0|\mathbf{x}) &= \frac{P(C = 0)\, p(\mathbf{x}|C = 0)}{p(\mathbf{x})} = \frac{0{,}3 \cdot 0{,}4}{0{,}89} = \frac{0{,}12}{0{,}89} = 0{,}134
\end{aligned}
$$

2.2 Prise de décision bayésienne

Formule de Bayes avec plusieurs classes

$$P(C_i|\mathbf{x}) = \frac{P(C_i)\,p(\mathbf{x}|C_i)}{\sum_{k=1}^{K} P(C_k)\,p(\mathbf{x}|C_k)}$$

- $P(C_i) \geq 0$ et $\sum_{i=1}^{K} P(C_i) = 1$
- Choisir classe C_i pour donnée $\mathbf{x}$ selon $C_i = \operatorname*{argmax}_{k=1}^{K} P(C_k|\mathbf{x})$

Fonction de perte

- Toutes les décisions n'ont pas le même impact
  - Prêter à un client à haut risque comparativement à ne pas prêter à un client à faible risque
  - Diagnostic médical : impacts possibles de la non-détection d'une maladie grave
  - Détection d'intrusions
- Quantifier avec une fonction de perte $\mathcal{L}(\alpha_i, C_j)$
  - Effectuer une action α_i alors que la classe véritable est C_j

Risque

- Risque espéré d'une action α :

$$R(\alpha|\mathbf{x}) = \sum_{k=1}^{K} \mathcal{L}(\alpha, C_k)\, P(C_k|\mathbf{x})$$

- Action minimisant le risque :

$$\alpha^* = \underset{\forall \alpha}{\operatorname{argmin}}\, R(\alpha|\mathbf{x})$$

- Modifier la fonction de perte modifie le risque
  - Modifier le coût associé à un faux négatif relativement au coût d'un faux positif

Matrice de confusion (deux classes)

		Décision	
		α_0	α_1
Vérité	C_0	0	λ_{FP}
	C_1	λ_{FN}	0

- $\mathcal{L}(\alpha = 1, C = 0) = \lambda_{\mathrm{FP}}$: coût d'un faux positif
- $\mathcal{L}(\alpha = 0, C = 1) = \lambda_{\mathrm{FN}}$: coût d'un faux négatif

Matrice de confusion (K classes)

	α_0	α_1	$\cdots$	α_K
C_0	0	$\lambda_{1,0}$	$\cdots$	$\lambda_{K,0}$
C_1	$\lambda_{0,1}$	0	$\cdots$	$\lambda_{K,1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
C_K	$\lambda_{0,K}$	$\lambda_{1,K}$	$\cdots$	0

Fonction de perte zéro-un

- Fonction de perte zéro-un :

$$\mathcal{L}(\alpha_i, C_j) = \begin{cases} 0 & \text{si } i = j \\ 1 & \text{si } i \neq j \end{cases}$$

- Risque correspondant :

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{k=1}^{K} \mathcal{L}(\alpha_i, C_k)\, P(C_k|\mathbf{x}) \\ &= \sum_{k \neq i} P(C_k|\mathbf{x}) \\ &= 1 - P(C_i|\mathbf{x}) \end{aligned}$$

- Décision optimale :

$$\alpha^* = \underset{\alpha_k=\alpha_1}{\overset{\alpha_K}{\operatorname{argmax}}}\, P(C_k|\mathbf{x})$$

Option de rejet

- Pour plusieurs applications, un mauvais classement peut avoir un impact considérable
  - Ajout d'une option de rejet en cas de doute, action α_{K+1}
- Fonction de perte zéro-un avec rejet :

$$\mathcal{L}(\alpha_i, C_j) = \begin{cases} 0 & \text{si } i = j \\ \lambda & \text{si } i = K+1 \\ 1 & \text{autrement} \end{cases}$$

- Dans ce cas :

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{k \neq i} P(C_k|\mathbf{x}) = 1 - P(C_i|\mathbf{x}) \\ R(\alpha_{K+1}|\mathbf{x}) &= \sum_{k=1}^{K} \lambda P(C_k|\mathbf{x}) = \lambda \end{aligned}$$

- Décision optimale avec option de rejet :

$$\alpha^* = \operatorname*{argmin}_{\alpha_k=\alpha_1}^{\alpha_{K+1}} R(\alpha_k|\mathbf{x})$$

- Décision optimale pour fonction de perte zéro-un avec rejet :

$$\alpha^* = \begin{cases} \alpha_{K+1} & \text{si } P(C_j|\mathbf{x}) < 1-\lambda,\ \forall j = 1,\ldots,K \\ \operatorname*{argmax}_{\alpha_j=\alpha_1}^{\alpha_K} P(C_j|\mathbf{x}) & \text{autrement} \end{cases}$$

Matrice de confusion (K classes et option de rejet)

	α_0	α_1	$\cdots$	α_K	α_{K+1}
C_0	0	$\lambda_{1,0}$	$\cdots$	$\lambda_{K,0}$	$\lambda_{K+1,0}$
C_1	$\lambda_{0,1}$	0	$\cdots$	$\lambda_{K,1}$	$\lambda_{K+1,1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_K	$\lambda_{0,K}$	$\lambda_{1,K}$	$\cdots$	0	$\lambda_{K+1,K}$

Fonction discriminante

- Fonctions discriminantes pour classement : $\alpha^t = \operatorname*{argmax}_{\alpha_i=\alpha_1}^{\alpha_K} \mathrm{h}_i(\mathbf{x}^t)$
  - Dans le cas bayésien (général) : $\mathrm{h}_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$
  - Bayésien avec fonction de perte zéro-un : $\mathrm{h}_i(\mathbf{x}) = P(C_i|\mathbf{x})$
  - En ignorant normalisation selon $p(\mathbf{x})$: $\mathrm{h}_i(\mathbf{x}) = p(\mathbf{x}|C_i)\, P(C_i)$
- *Régions de décisions* : division de l'espace d'entrée selon K régions :
  - $\mathcal{R}_1, \ldots, \mathcal{R}_K$ où $\mathcal{R}_i = \{\mathbf{x}|\mathrm{h}_i(\mathbf{x}) = \max_{\forall k} \mathrm{h}_k(\mathbf{x})\}$
- Les régions de décisions sont séparées par des *frontières de décisions*
- Cas à deux classes est un *dichotomiseur*, cas à $K \geq 3$ classes est un *plurichotomiseur*

Régions et frontières de décision