

Challenge MIMSE 2014

Adrien Todeschini

Table des matières

1	Fonction de coût	1
2	Matrice de confusion	1
3	Risque de prédiction	2
3.1	Exemple	2
3.1.1	La fonction de coût 0-1	2
3.1.2	La fonction de coût pondérée	3
4	Règle de décision	3
4.1	Exemple	3

1 Fonction de coût

Soit $\hat{Y} \in \mathcal{Y}$ la sortie prédite d'un individu et $Y \in \mathcal{Y}$ sa vraie valeur. Soit $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ une fonction de coût telle que $L(Y, \hat{Y})$ est le coût associé à la décision : "prédire \hat{Y} alors que la vraie valeur est Y ".

En classification, on a $\mathcal{Y} = \{1, \dots, K\}$ et la fonction de coût peut être résumée par une matrice de coûts C de taille $K \times K$ telle que $C_{kl} = L(k, l)$ (voir table 1).

		classe prédite			
		$l = 1$	$l = 2$...	$l = K$
vraie classe	$k = 1$	$C_{11} \leq 0$	$C_{12} \geq 0$...	$C_{1K} \geq 0$
	$k = 2$	$C_{21} \geq 0$	$C_{22} \leq 0$...	$C_{2K} \geq 0$
	...	\vdots	\vdots	\ddots	\vdots
	$k = K$	$C_{K1} \geq 0$	$C_{K2} \geq 0$...	$C_{KK} \leq 0$

TABLE 1 – Matrice de coûts

La fonction de coût s'écrit :

$$L(Y, \hat{Y}) = C_{Y\hat{Y}} = \sum_{k=1}^K \sum_{l=1}^K C_{kl} \mathbb{1}(Y = k \text{ et } \hat{Y} = l)$$

2 Matrice de confusion

Soit un échantillon (Y_i, \hat{Y}_i) , $i = 1, \dots, n$ de couples classe réelle, classe prédite. La matrice de confusion comptabilise les occurrences des prédictions en fonction des vraies valeurs dans un tableau de contingence (voir table 2).

		classe prédite			
		$l = 1$	$l = 2$...	$l = K$
vraie classe	$k = 1$	$p_{11} \times n$	$p_{12} \times n$...	$p_{1K} \times n$
	$k = 2$	$p_{21} \times n$	$p_{22} \times n$...	$p_{2K} \times n$
	...	\vdots	\vdots	\ddots	\vdots
	$k = K$	$p_{K1} \times n$	$p_{K2} \times n$...	$p_{KK} \times n$

TABLE 2 – Matrice de confusion

où p_{kl} est la proportion d'individus de classe k auxquels on a prédit la classe l

$$p_{kl} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(Y_i = k \text{ et } \hat{Y}_i = l)$$

3 Risque de prédiction

Le risque théorique (aussi appelé taux d'erreur théorique) est l'espérance de la fonction de coût

$$E = \mathbb{E} \left[L(Y, \hat{Y}) \right]$$

Le risque théorique peut être approché par le risque empirique (aussi appelé taux d'erreur empirique) donné par

$$\begin{aligned} \hat{E}_n &= \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{Y}_i) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^K C_{kl} \mathbb{1}(Y_i = k \text{ et } \hat{Y}_i = l) \\ &= \sum_{k=1}^K \sum_{l=1}^K C_{kl} p_{kl} \end{aligned}$$

Il s'agit du coût moyen sur l'échantillon.

3.1 Exemple

Dans le challenge, on a $K = 2$ classes avec 1=Good et 2=Bad. On note $FP = p_{21} \times n$ le nombre de faux positifs et $FN = p_{12} \times n$ le nombre de faux négatifs. Les deux fonctions de coût suivantes sont considérées.

3.1.1 La fonction de coût 0-1

Elle ne distingue pas les deux types d'erreurs et elle est associée à la matrice de coûts $C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

L'expression du coût moyen devient

$$\begin{aligned} \hat{E}_n &= \sum_{k=1}^2 \sum_{l=1}^2 C_{kl} p_{kl} \\ &= C_{11} p_{11} + C_{21} p_{21} + C_{12} p_{12} + C_{22} p_{22} \\ &= p_{21} + p_{12} \\ &= \frac{FP + FN}{n} \end{aligned}$$

ce qui correspond au taux d'erreur standard (appelé "taux d'erreur" dans le challenge).

3.1.2 La fonction de coût pondérée

On associe un coût cinq fois plus important aux faux positifs avec la matrice de coûts $C = \begin{pmatrix} 0 & 1 \\ 5 & 0 \end{pmatrix}$.

L'expression du coût moyen devient

$$\begin{aligned} \hat{E}_n &= C_{11}p_{11} + C_{21}p_{21} + C_{12}p_{12} + C_{22}p_{22} \\ &= 5 \times p_{21} + p_{12} \\ &= \frac{5 \times FP + FN}{n} \end{aligned}$$

ce qui correspond à un taux d'erreur pondéré (appelé "coût moyen" dans le challenge).

4 Règle de décision

Soit $x \in \mathcal{X}$ la variable d'entrée. La règle optimale de classification de Bayes associée à la fonction de coût L est donnée par

$$\begin{aligned} \hat{Y}(x) &= \arg \min_{l \in \{1, \dots, K\}} \mathbb{E}[L(Y, l) | X = x] \\ &= \arg \min_{l \in \{1, \dots, K\}} \sum_{k=1}^K L(k, l) \Pr(Y = k | X = x) \\ &= \arg \min_{l \in \{1, \dots, K\}} \sum_{k=1}^K C_{kl} \Pr(Y = k | X = x) \quad (\text{approche discriminante}) \\ &= \arg \min_{l \in \{1, \dots, K\}} \sum_{k=1}^K C_{kl} p(X = x | Y = k) \Pr(Y = k) \quad (\text{modèle génératif}) \end{aligned}$$

Cette règle de décision est applicable si l'on dispose d'un modèle discriminant $\Pr(Y = k | X = x)$ (ex : régression logistique) ou génératif $p(X = x | Y = k) \Pr(Y = k)$ (ex : analyse discriminante, bayésien naïf).

4.1 Exemple

On considère $K = 2$ classes et la matrice de coûts $C = \begin{pmatrix} 0 & C_{12} \\ C_{21} & 0 \end{pmatrix}$.

Alors deux cas sont à considérer :

— Si $l = 1$:

$$\begin{aligned} \mathbb{E}[L(Y, l) | X = x] &= \sum_{k=1}^2 C_{k1} \Pr(Y = k | X = x) \\ &= C_{21} \Pr(Y = 2 | X = x) \end{aligned}$$

— Si $l = 2$:

$$\begin{aligned} \mathbb{E}[L(Y, l) | X = x] &= \sum_{k=1}^2 C_{k2} \Pr(Y = k | X = x) \\ &= C_{12} \Pr(Y = 1 | X = x) \end{aligned}$$

On en déduit la règle de décision suivante :

— cas discriminant : si la condition

$$C_{21} \Pr(Y = 2 | X = x) \leq C_{12} \Pr(Y = 1 | X = x) \Leftrightarrow \Pr(Y = 1 | X = x) \geq \frac{C_{21}}{C_{12}} \Pr(Y = 2 | X = x)$$

est satisfaite alors on affectera la classe $\hat{Y}(x) = 1$, sinon on affectera la classe $\hat{Y}(x) = 2$.

— cas génératif : la condition devient

$$C_{21}p(X = x|Y = 2)\Pr(Y = 2) \leq C_{12}p(X = x|Y = 1)\Pr(Y = 1) \Leftrightarrow p(X = x|Y = 1) \geq \frac{C_{21}\pi_2}{C_{12}\pi_1}p(X = x|Y = 2)$$

où $\Pr(Y = k) = \pi_k$

Remarque Si $C_{21} = C_{12}$ (fonction de coût 0-1) on retrouve la règle d'affectation de la classe la plus probable

— cas discriminant :

$$\hat{Y}(x) = \begin{cases} 1 & \text{si } \Pr(Y = 1|X = x) \geq \Pr(Y = 2|X = x) \\ 2 & \text{sinon} \end{cases}$$

— cas génératif :

$$\hat{Y}(x) = \begin{cases} 1 & \text{si } \pi_1 p(X = x|Y = 1) \geq \pi_2 p(X = x|Y = 2) \\ 2 & \text{sinon} \end{cases}$$