A2DI : Théorie de la décision

John Klein

Université de Lille - CRIStAL UMR CNRS 9189

Où en est-on dans notre problème d'apprentissage supervisé ?

- En théorie on veut minimiser Err_{gen}.
- En pratique on minisera Err_{train} tout en s'assurant que Err_{train} ne dévie pas de Err_{gen}.

Le calcul de ces 2 erreurs dépend de la perte L. Dans ce chapitre, nous allons étudier l'influence de L en détail.

Comment prendre de bonnes décisions ?

- Imaginons un problème d'apprentissage supervisé où une valeur y est associée à un exemple $\mathbf{x}$.
- Supposons que notre problème consiste à choisir une action $a \in \mathcal{A}$ quand nous observons $\mathbf{x}$.
- Enfin, nous avons connaissance d'une fonction de perte $L : \mathbb{Y} \times \mathcal{A} \to \mathbb{R}$.

Exemple

Problème de régression : agir = choisir un $\hat{y}$, $\mathcal{A} = \mathbb{Y}$ et

$$L(y, a) = (y - a)^2 .$$

Ce formalisme reprend donc celui vu au chapitre précédent où on s'intéressait à l'erreur de généralisation = bonne prédiction en moyenne. Dans le chapitre précédent, Err_{gen} nous a servi à contrôler la qualité de l'apprentissage. Ici, on va s'en servir pour orienter notre apprentissage !

Comment prendre de bonnes décisions ? minimiser la perte attendue.

- La démarche optimale consiste à sélectionner la fonction prédictrice f^* qui en moyenne causera le moins de perte :

$$f^*(\mathbf{x}) = \arg\min_{a \in \mathcal{A}} \mathbb{E}[L]. \tag{1}$$

- ... où est $\mathbf{x}$ dans cette expression ?
- dans la distribution sous laquelle le calcul d'espérance est fait :

$$f^*(\mathbf{x}) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{Y|X=\mathbf{x}}[L(y, a)].$$

Comment prendre de bonnes décisions ? minimiser la perte attendue.

- Dans cette section du chapitre, on se concentre sur la solution bayésienne.
- Ce cadre de travail se nomme théorie Bayésienne de la décision.
- La perte $\mathbb{E}_{y|\mathbf{x}}[L]$ est appelée perte attendue a posteriori.
- On note souvent :

$$\rho(a|\mathbf{x}) = \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}}[L(y,a)] = \int_{\mathbb{Y}} L(y,a)\, p_{Y|\mathbf{X}=\mathbf{x}}(y)\, dy. \qquad (2)$$

Théorie Bayésienne de la décision

avec la **0-1 loss** pour un problème de classification ($\mathbb{Y} = \mathcal{A} = \mathcal{C}$).

$$\begin{aligned} L : \mathcal{C} \times \mathbb{X} &\rightarrow \{0; 1\}, && (3) \\ (c_i, \hat{c}(\mathbf{x_i})) &\rightarrow \begin{cases} 0 & \text{si } c_i = \hat{c}(\mathbf{x_i}) \\ 1 & \text{sinon} \end{cases}. && (4) \end{aligned}$$

La perte attendue a posteriori s'écrit :

$$\begin{aligned} \rho(a|\mathbf{x}) &= \sum_{c \in \mathcal{C}} (1 - \mathbb{I}_c(a)) \, p_{Y|\mathbf{X}=\mathbf{x}}(c), && (5) \\ &= 1 - p_{Y|\mathbf{X}=\mathbf{x}}(a). && (6) \end{aligned}$$

La fonction prédictrice qui minimise ρ est donc :

$$f^*(\mathbf{x}) = \arg\max_{a \in \mathcal{C}} p_{Y|\mathbf{X}=\mathbf{x}}(a). \quad (7)$$

Théorie Bayésienne de la décision

avec une perte à option de rejet pour un problème de classification ($\mathbb{Y} = \mathcal{C}$ et $\mathcal{A} = \mathcal{C} \cup \{c_r\}$) :

- Dans certains domaines, laisser le système prendre une décision risquée n'est pas acceptable.
- On ajoute une nouvelle classe fictive c_r appelée classe de rejet.
- La perte correspondante s'exprime alors comme suit :

$$
\begin{aligned}
L : \mathcal{C} \times \mathbb{X} \quad &\rightarrow \quad \{0;1\}, && (8)\\
(c_i, \hat{c}(\mathbf{x_i})) \quad &\rightarrow \quad \begin{cases} 0 & \text{si } c_i = \hat{c}(\mathbf{x_i}) \\ \lambda_r & \text{si } c_r = \hat{c}(\mathbf{x_i}) \\ \lambda & \text{sinon} \end{cases}, && (9)
\end{aligned}
$$

avec $0 < \lambda_r < \lambda$.

Théorie Bayésienne de la décision

avec une perte à option de rejet pour un problème de classification ($\mathbb{Y} = \mathcal{C}$ et $\mathcal{A} = \mathcal{C} \cup \{c_r\}$) :

Pour une classifcation binaire, on peut résumer les pertes comme suit :

	$\hat{c}(\mathbf{x}) = a = c_0$	$\hat{c}(\mathbf{x}) = a = c_1$	$\hat{c}(\mathbf{x}) = a = c_r$
$y = c_0$	0	λ	λ_r
$y = c_1$	λ	0	λ_r

La perte attendue a posteriori s'écrit :

$$\rho(a|\mathbf{x}) = \left\{ \right. \tag{10}$$

Théorie Bayésienne de la décision

avec une perte à option de rejet pour un problème de classification
($\mathbb{Y} = \mathcal{C}$ et $\mathcal{A} = \mathcal{C} \cup \{c_r\}$) :
Quant à la fonction prédictrice qui minimise ρ, on a :

$$f^* (\mathbf{x}) = \tag{11}$$

Théorie Bayésienne de la décision

avec une perte asymétrique pour un problème de classification binaire $(\mathbb{Y} = \mathcal{A} = \mathcal{C})$:

	$\hat{c}(\mathbf{x}) = a = c_0$	$\hat{c}(\mathbf{x}) = a = c_1$
$y = c_0$	0	λ_{FP}
$y = c_1$	λ_{FN}	0

La perte attendue a posteriori s'écrit :

$$\rho(a|\mathbf{x}) = \left\{ \right. \tag{12}$$

Théorie Bayésienne de la décision
avec une perte asymétrique pour un problème de classification binaire $(\mathbb{Y} = \mathcal{A} = \mathcal{C})$: Exercice

Si on a $\frac{\lambda_{FP}}{\lambda_{FN}} = \alpha > 1$, c'est à dire qu'un faux positif coûte α plus cher qu'un faux négatif, à partir de quel seuil τ de probabilité a posteriori vais je choisir la classe c_1 ?

Théorie Bayésienne de la décision

avec une perte asymétrique pour un problème de classification binaire ($\mathbb{Y} = \mathcal{A} = \mathcal{C}$) : Matrice de confusion.

- Ayant déterminé τ, on peut utiliser notre règle de décision $\hat{c}$ sur un ensemble de test $\mathcal{D}_{\text{test}}$.
- La taille de cet ensemble est notée $\sharp\mathcal{D}_{\text{test}} = n_{\text{test}}$.
- On peut alors détailler les probabilités de se tromper ou non d'une manière plus fine que l'erreur de généralisation :

	$y = c_0$	$y = c_1$
$\hat{c}(\mathbf{x}) = a = c_0$	$\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_0}(c_i)\,\mathbb{I}_{c_0}(\hat{c}(\mathbf{x}_i))$	$\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}(c_i)\,\mathbb{I}_{c_0}(\hat{c}(\mathbf{x}_i))$
$\hat{c}(\mathbf{x}) = a = c_1$	$\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_0}(c_i)\,\mathbb{I}_{c_1}(\hat{c}(\mathbf{x}_i))$	$\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}(c_i)\,\mathbb{I}_{c_1}(\hat{c}(\mathbf{x}_i))$

- On peut calculer une telle matrice pour du multi-classe.

Théorie Bayésienne de la décision

avec une perte asymétrique pour un problème de classification binaire ($\mathbb{Y} = \mathcal{A} = \mathcal{C}$) : Matrice de confusion.

- En normalisant cette matrice par colonne par colonne, on obtient une estimation empirique de $p\left(\hat{c}|c\right)$.
- Avant normalisation :

	$y = c_0$	$y = c_1$
$\hat{c}\left(\mathbf{x}\right) = a = c_0$	$\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_0}\left(c_i\right)\mathbb{I}_{c_0}\left(\hat{c}\left(\mathbf{x}_i\right)\right)$	$\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}\left(c_i\right)\mathbb{I}_{c_0}\left(\hat{c}\left(\mathbf{x}_i\right)\right)$
$\hat{c}\left(\mathbf{x}\right) = a = c_1$	$\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_0}\left(c_i\right)\mathbb{I}_{c_1}\left(\hat{c}\left(\mathbf{x}_i\right)\right)$	$\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}\left(c_i\right)\mathbb{I}_{c_1}\left(\hat{c}\left(\mathbf{x}_i\right)\right)$
Somme	$n_0 = \sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_0}\left(c_i\right)$	$n_1 = \sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}\left(c_i\right)$

Théorie Bayésienne de la décision

avec une perte asymétrique pour un problème de classification binaire $(\mathbb{Y} = \mathcal{A} = \mathcal{C})$: Matrice de confusion.

- En normalisant cette matrice par colonne par colonne, on obtient une estimation empirique de $p(\hat{c}|c)$.
- Après normalisation :

	$y = c_0$	$y = c_1$
$\hat{c}(\mathbf{x}) = a = c_0$	$\frac{1}{n_0}\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_0}(c_i)\,\mathbb{I}_{c_0}(\hat{c}(\mathbf{x}_i))$	$\frac{1}{n_1}\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}(c_i)\,\mathbb{I}_{c_0}(\hat{c}(\mathbf{x}_i))$
$\hat{c}(\mathbf{x}) = a = c_1$	$\frac{1}{n_0}\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_0}(c_i)\,\mathbb{I}_{c_1}(\hat{c}(\mathbf{x}_i))$	$\frac{1}{n_1}\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}(c_i)\,\mathbb{I}_{c_1}(\hat{c}(\mathbf{x}_i))$

Théorie Bayésienne de la décision

avec une perte asymétrique pour un problème de classification binaire $(\mathbb{Y} = \mathcal{A} = \mathcal{C})$: Matrice de confusion.

- En normalisant cette matrice par colonne par colonne, on obtient une estimation empirique de $p\left(\hat{c}|c\right)$.
- Après normalisation :

	$\hat{c}\left(\mathbf{x}\right) = a = c_0$	$\hat{c}\left(\mathbf{x}\right) = a = c_1$
$y = c_0$	$\hat{p}\left(\hat{c} = c_0 \| c = c_0\right)$	$\hat{p}\left(\hat{c} = c_1 \| c = c_0\right)$
$y = c_1$	$\hat{p}\left(\hat{c} = c_0 \| c = c_1\right)$	$\hat{p}\left(\hat{c} = c_1 \| c = c_1\right)$

Théorie Bayésienne de la décision

avec une perte asymétrique pour un problème de classification binaire ($\mathbb{Y} = \mathcal{A} = \mathcal{C}$) : Matrice de confusion.

- Quand la matrice est normalisée, chaque entrée a un nom spécifique

	$\hat{c}(\mathbf{x}) = a = c_0$	$\hat{c}(\mathbf{x}) = a = c_1$
$y = c_0$	taux de vrais négatifs spécificité	taux de faux négatifs taux de cibles manquées erreur de type II
$y = c_1$	taux de faux positifs erreur de type I	taux de vrais positifs rappel sensibilité

Théorie Bayésienne de la décision

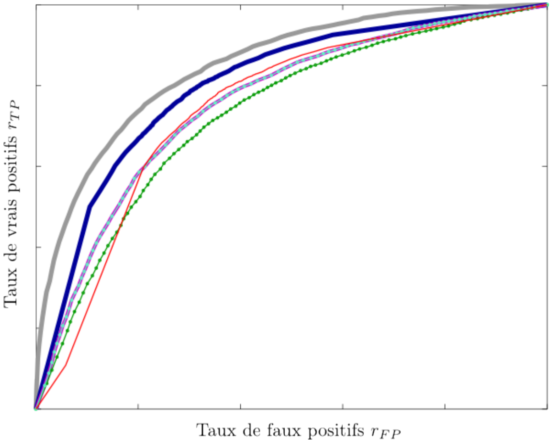
avec une perte asymétrique pour un problème de classification binaire ($\mathbb{Y} = \mathcal{A} = \mathcal{C}$) : courbe ROC.

- Choisir un seuil τ fixe permet de calculer tous ces critères d'évaluation d'un classifieur binaire, mais pour comparer 2 classifieurs ne faudrait-il pas le faire pour tout τ ?
- La courbe ROC se propose d'atteindre cet objectif en calculant les deux types d'erreur pour différentes valeurs de τ.

Théorie Bayésienne de la décision

avec une perte asymétrique pour un problème de classification binaire ($\mathbb{Y} = \mathcal{A} = \mathcal{C}$) : courbe ROC.

- Chaque paire (r_{FP}, r_{FN}) obtenue pour une valeur de τ forme un point de la courbe.



Théorie Bayésienne de la décision

avec une perte asymétrique pour un problème de classification binaire $(\mathbb{Y} = \mathcal{A} = \mathcal{C})$: autres critères.

- On préfère parfois la courbe rappel/précision à la courbe ROC.
- La précision est la proportion de vrais positifs parmi ceux détectés comme tel :

$$\hat{p}\left(c = c_1 | \hat{c} = c_1\right) = \frac{\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}\left(c_i\right) \mathbb{I}_{c_1}\left(\hat{c}\left(\mathbf{x}_i\right)\right)}{\sum_{i=1}^{n_{\text{test}}} \mathbb{I}_{c_1}\left(\hat{c}\left(\mathbf{x}_i\right)\right)} \tag{13}$$

- Pour essayer de combiner précision et rappel en une seule valeur, on peut utilise le F-score :

$$\text{F-score} = \frac{2 \times \text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}} \tag{14}$$

Théorie Bayésienne de la décision

avec la **quadratic loss** pour un problème de régression ($\mathbb{Y} = \mathcal{A} = \mathbb{R}$).

$$\begin{aligned} L : \mathbb{R} \times \mathbb{X} &\rightarrow \mathbb{R}, & (15)\\ (y, \hat{y}(\mathbf{x})) &\rightarrow (y - \hat{y}(\mathbf{x}))^2 . & (16) \end{aligned}$$

La perte attendue a posteriori s'écrit :

$$\rho(a|\mathbf{x}) = \int_{\mathbb{R}} (y-a)^2 \, p_{Y|\mathbf{X}=\mathbf{x}}(y) \, dy. \quad (17)$$

La fonction prédictrice qui minimise ρ est alors :

$$f^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X}=\mathbf{x}] . \quad (18)$$

Théorie Bayésienne de la décision

Pour preuve, résolvons :

$$\frac{d\rho}{da}(a|\mathbf{x}) = 0, \tag{19}$$

Théorie Bayésienne de la décision

avec la **absolute loss** pour un problème de régression ($\mathbb{Y} = \mathcal{A} = \mathbb{R}$).

$$\begin{aligned} L : \mathbb{R} \times \mathbb{X} &\rightarrow \mathbb{R}, & (20) \\ (y, \hat{y}(\mathbf{x})) &\rightarrow |y - \hat{y}(\mathbf{x})| . & (21) \end{aligned}$$

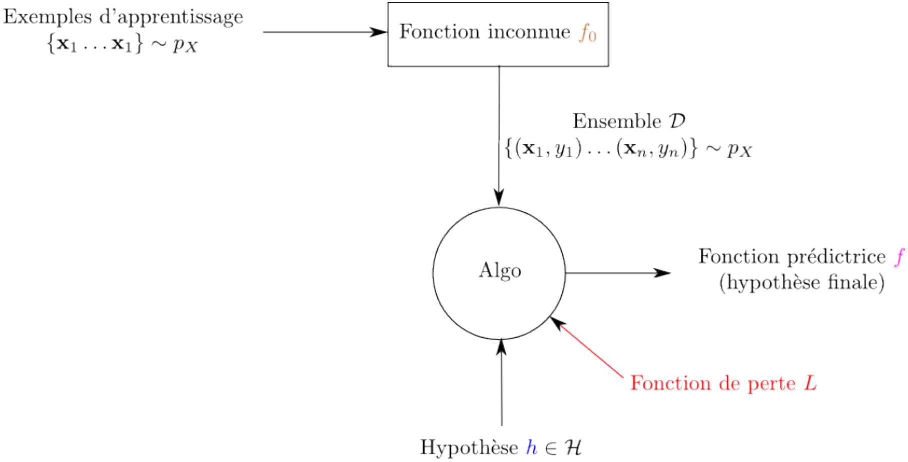
La perte attendue a posteriori s'écrit :

$$\rho(a|\mathbf{x}) = \int_{\mathbb{R}} |y - a| \, p_{Y|\mathbf{X}=\mathbf{x}}(y) \, dy. \quad (22)$$

La fonction prédictrice qui minimise ρ est alors la médiane de la distribution conditionnelle $p_{Y|\mathbf{X}=\mathbf{x}}(y)$.

Théorie Bayésienne de la décision

La fonction de perte influe directement sur la solution retenue.



$\boxed{\text{Messages importants du chapitre}}$:

- Les pertes classiques mènent à des classifieurs/régresseurs optimaux mais qui nécessitent de connaître $p_{Y|\mathbf{X}}$.

- Les pertes peuvent être *customisées* selon le contexte applicatifs.

- Elles influent directement sur la solution retenue.

Notion de bruit : $B = Y - f_0(\mathbf{X})$. Si le bruit en centré alors

$$\begin{aligned} \mathbb{E}[B|X = \mathbf{x}_i] &= 0, \\ \Leftrightarrow \mathbb{E}[Y|X = \mathbf{x}_i] &= f_0(\mathbf{x}_i). \end{aligned} \tag{23}$$