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.

Plan du chapitre

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 apprendre !

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.

- 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.

- 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 ($\mathbb{Y} = \mathcal{A} = \mathcal{C}$).

$$L : \mathcal{C} \times \mathbb{X} \rightarrow \{0; 1\}, \tag{3}$$

$$(c_i, \hat{c}(\mathbf{x_i})) \rightarrow \begin{cases} 0 & \text{si } c_i = \hat{c}(\mathbf{x_i}) \\ 1 & \text{sinon} \end{cases}. \tag{4}$$

La perte attendue a posteriori s'écrit :

$$\rho(a|\mathbf{x}) = \sum_{c \in \mathcal{C}} (1 - \mathbb{I}_c(a)) \, p_{Y|\mathbf{X}=\mathbf{x}}(c), \tag{5}$$

$$= 1 - p_{Y|\mathbf{X}=\mathbf{x}}(a). \tag{6}$$

La fonction prédictrice qui minimise ρ est donc :

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

Théorie Bayésienne de la décision

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

- 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 :

$$L : \mathcal{C} \times \mathbb{X} \rightarrow \{0; 1\}, \tag{8}$$

$$(c_i, \hat{c}(\mathbf{x_i})) \rightarrow \begin{cases} 0 & \text{si } c_i = \hat{c}(\mathbf{x_i}) \\ \lambda_r & \text{si } c_r = \hat{c}(\mathbf{x_i}), \\ \lambda & \text{sinon} \end{cases} \tag{9}$$

avec $0 < \lambda_r < \lambda$.

Théorie Bayésienne de la décision

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

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 :

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

Théorie Bayésienne de la décision

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

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 $\frac{\lambda_{FP}}{\lambda_{FP}} = \alpha > 1$, 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 :

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

- 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 $p(\hat{c}|c)$.
- Avant normalisation :

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

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 $p(\hat{c}|c)$.
- Après normalisation :

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

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 $p(\hat{c}|c)$.
- Après normalisation :

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

Théorie Bayésienne de la décision

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

- 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

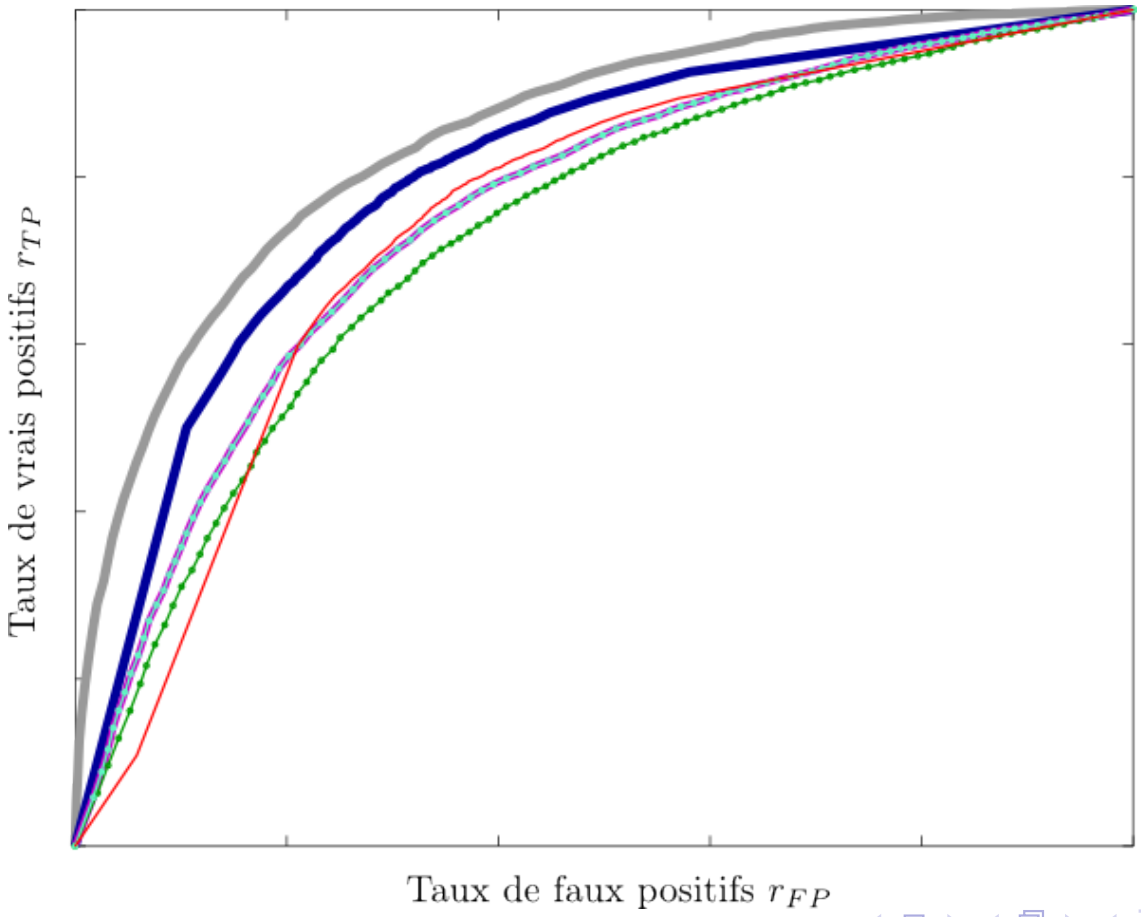
avec une perte asymétrique pour un problème de classification binaire ($\mathbb{Y} = \mathcal{A} = \mathcal{C}$) : courbe ROC.

- 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 ($\mathbb{Y} = \mathcal{A} = \mathbb{R}$).

$$L : \mathbb{R} \times \mathbb{X} \rightarrow \mathbb{R}, \tag{15}$$

$$(y_i, \hat{y}(\mathbf{x_i})) \rightarrow (y_i - \hat{y}(\mathbf{x_i}))^2 . \tag{16}$$

La perte attendue a posteriori s'écrit :

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

La fonction prédictrice qui minimise ρ est alors :

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

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 ($\mathbb{Y} = \mathcal{A} = \mathbb{R}$).

$$L : \mathbb{R} \times \mathbb{X} \rightarrow \mathbb{R}, \tag{20}$$

$$(y_i, \hat{y}(\mathbf{x_i})) \rightarrow |y_i - \hat{y}(\mathbf{x_i})|. \tag{21}$$

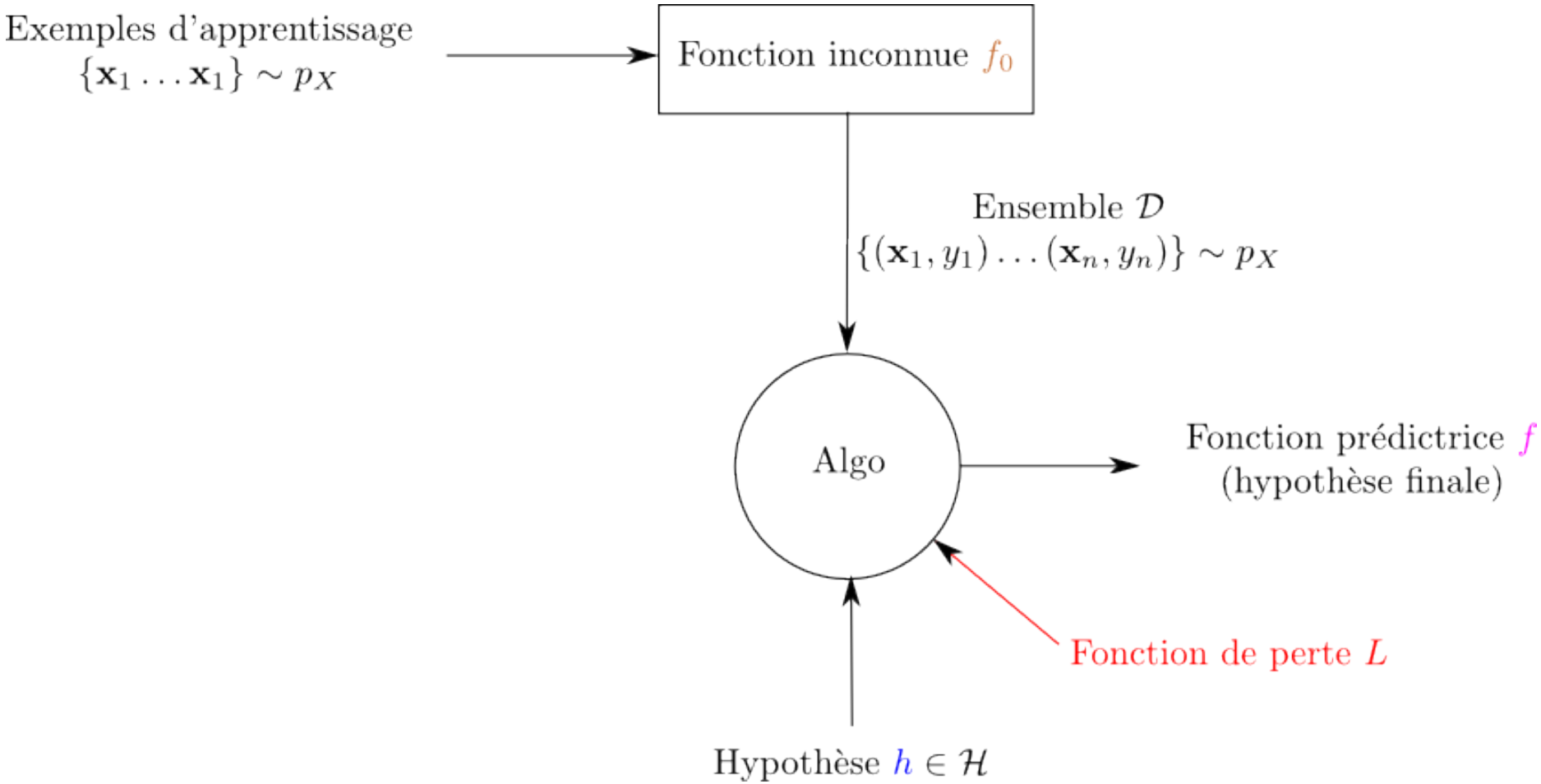
La perte attendue a posteriori s'écrit :

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

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.



Plan du chapitre

Théorie fréquentiste la décision

En statistiques fréquentistes, une autre approche de decision optimale existe.

- Pour les Bayésiens, on cherche la fonction f^* qui à chaque exemple observé $\mathbf{x}$ associe une action a telle que :

$$f^*(\mathbf{x}) = \arg\min_{a \in \mathcal{A}} \rho(a|\mathbf{x}), \tag{23}$$

$$= \arg\min_{a \in \mathcal{A}} \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}}[L(y,a)]. \tag{24}$$

- Pour les fréquentistes, on cherche directement la fonction f^* qui estime les quantités observables y à partir de $\mathbf{x}$ telle que

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{X},Y}[L(y, f(\mathbf{x}))]. \tag{25}$$

Choisir un y est vu comme une action et f^* est parfois appelé *oracle*, car c'est le prédicteur idéal qui minimise Err_{gen}.

Théorie fréquentiste la décision : Généralités sur l'estimation

Pause Stat !

- En général, le résultat d'une décision n'est pas observable et est un vecteur de paramètres $\boldsymbol{\theta}_0 \in \Theta$.

Théorie fréquentiste la décision : Généralités sur l'estimation

Pause Stat !

- En général, le résultat d'une décision n'est pas observable et est un vecteur de paramètres $\boldsymbol{\theta}_0 \in \Theta$.
- Selon la philosophie fréquentiste, il existe une unique vraie valeur $\boldsymbol{\theta}_0$ de $\boldsymbol{\theta}$ et ce n'est pas une v.a. mais une constante.

Théorie fréquentiste la décision : Généralités sur l'estimation

Pause Stat !

- En général, le résultat d'une décision n'est pas observable et est un vecteur de paramètres $\boldsymbol{\theta}_0 \in \Theta$.
- Selon la philosophie fréquentiste, il existe une unique vraie valeur $\boldsymbol{\theta}_0$ de $\boldsymbol{\theta}$ et ce n'est pas une v.a. mais une constante.
- Le seul aléa est alors dans les données $\mathcal{D}$.

Théorie fréquentiste la décision : Généralités sur l'estimation

Pause Stat !

- En général, le résultat d'une décision n'est pas observable et est un vecteur de paramètres $\boldsymbol{\theta}_0 \in \Theta$.
- Selon la philosophie fréquentiste, il existe une unique vraie valeur $\boldsymbol{\theta}_0$ de $\boldsymbol{\theta}$ et ce n'est pas une v.a. mais une constante.
- Le seul aléa est alors dans les données $\mathcal{D}$.

Exemple

Ex : On mesure 100 fois la masse d'un objet lundi, cela donne $\mathcal{D}_1$. On recommence mardi cela donner $\mathcal{D}_2 \neq \mathcal{D}_1$, il y a donc bien un aléa.

Théorie fréquentiste la décision : Généralités sur l'estimation

Pause Stat !

- La forme générale du problème d'estimation de $\boldsymbol{\theta}_0$ s'exprime alors comme suit

$$\hat{\boldsymbol{\theta}}^* = \underset{\hat{\boldsymbol{\theta}} \in \mathcal{F}}{\arg\min}\, \mathbb{E}_{\mathcal{D}|\boldsymbol{\theta}_0} \left[L\left(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}\right) \right], \tag{26}$$

où $\hat{\boldsymbol{\theta}}\left(\mathcal{D}\right)$ est un estimateur (une fonction estimant $\boldsymbol{\theta}_0$).

Théorie fréquentiste la décision : Généralités sur l'estimation

Pause Stat !

- La forme générale du problème d'estimation de $\boldsymbol{\theta}_0$ s'exprime alors comme suit

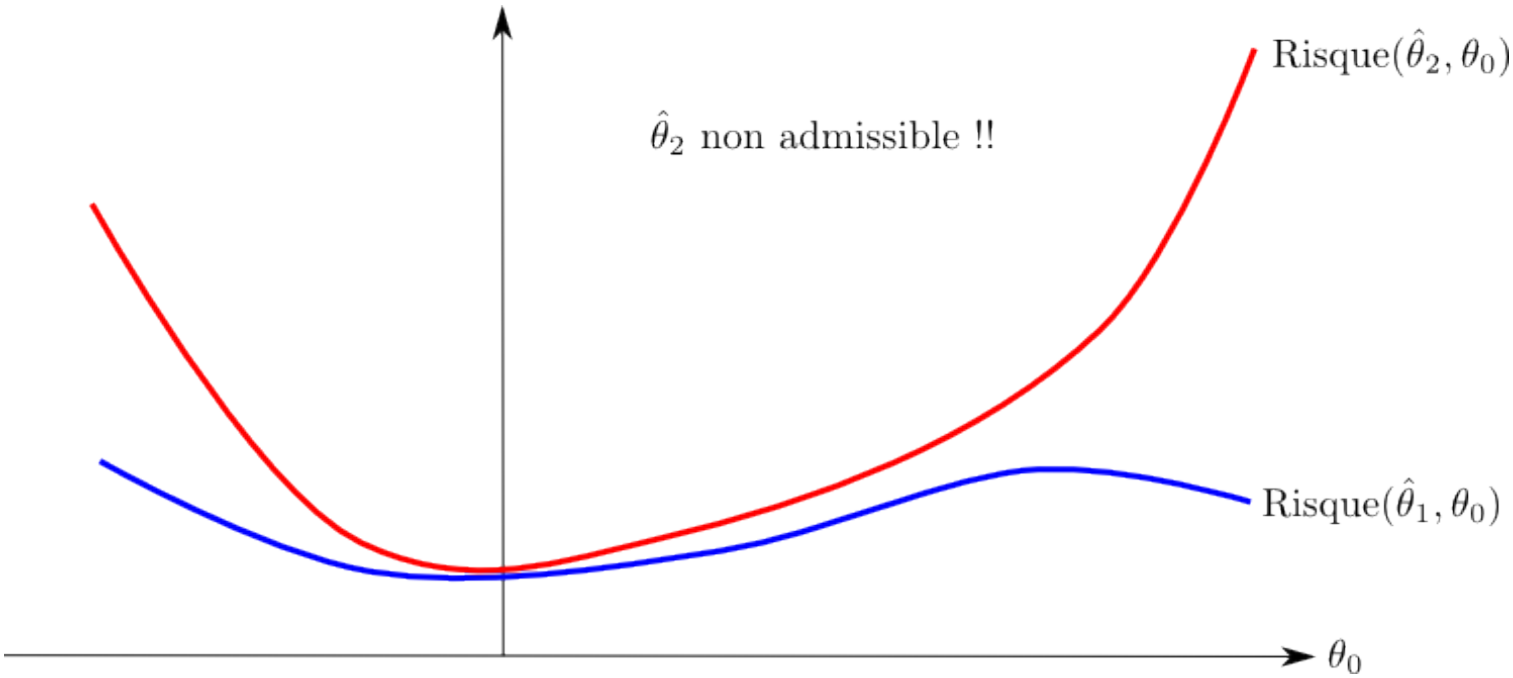
$$\hat{\boldsymbol{\theta}}^* = \underset{\hat{\boldsymbol{\theta}} \in \mathcal{F}}{\arg\min}\, \mathbb{E}_{\mathcal{D}|\boldsymbol{\theta}_0} \left[L\left(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}\right) \right], \tag{26}$$

  où $\hat{\boldsymbol{\theta}}\left(\mathcal{D}\right)$ est un estimateur (une fonction estimant $\boldsymbol{\theta}_0$).

- La quantité $\mathbb{E}_{\mathcal{D}|\boldsymbol{\theta}_0} \left[L\left(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}\right) \right]$ est appelée risque.

Théorie fréquentiste la décision : Généralités sur l'estimation

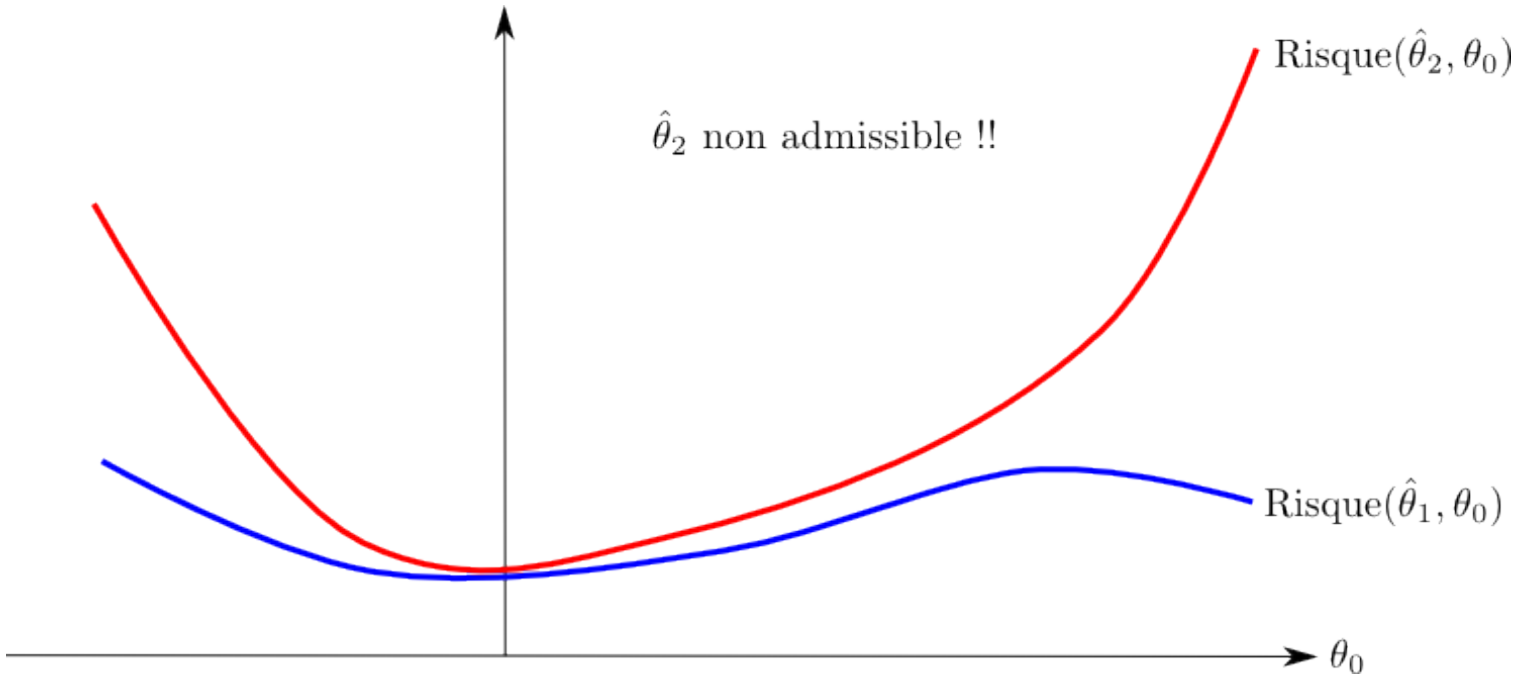
- $\rightarrow$ Gros problème : $\boldsymbol{\theta_0}$ est inconnu et le risque n'est pas calculable !
- .. mais on peut quand même savoir qu'un estimateur $\hat{\boldsymbol{\theta}}_1$ est meilleur qu'un $\hat{\boldsymbol{\theta}}_2$ si son risque est toujours plus faible (pour toute valeur de $\boldsymbol{\theta_0}$).



Théorie fréquentiste la décision : Généralités sur l'estimation

Définition

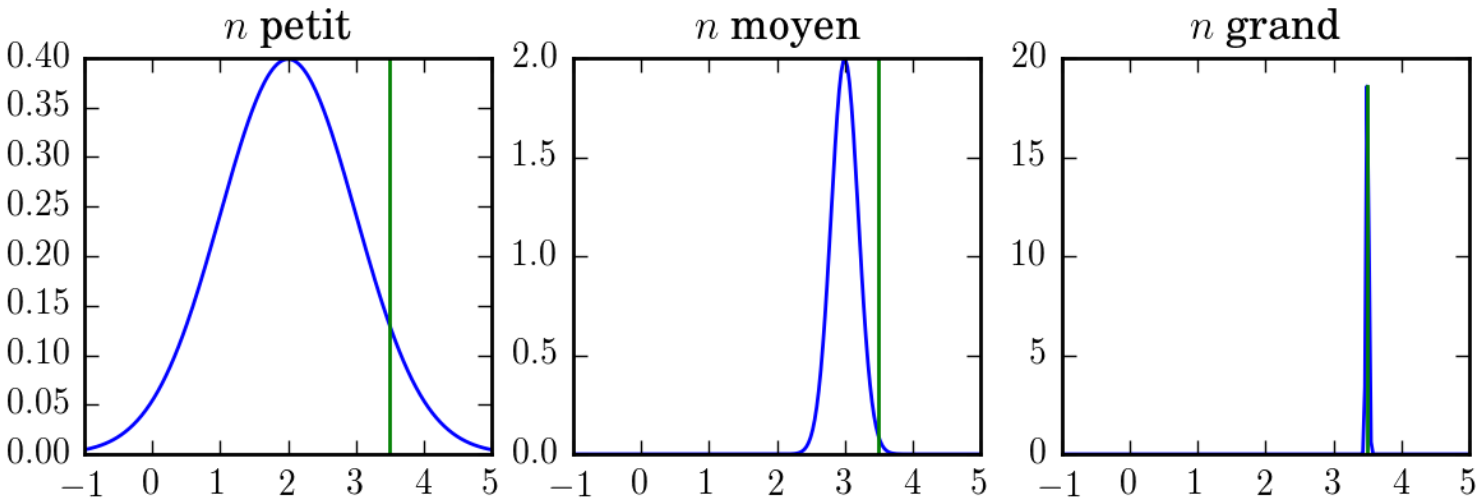
Un estimateur $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta_0}$ est admissible, si $\exists \mathbf{a}$ tel qu'il n'existe aucun autre estimateur ayant un risque plus faible en $\boldsymbol{\theta_0} = \mathbf{a}$.



Théorie fréquentiste la décision : Généralités sur l'estimation

Définition

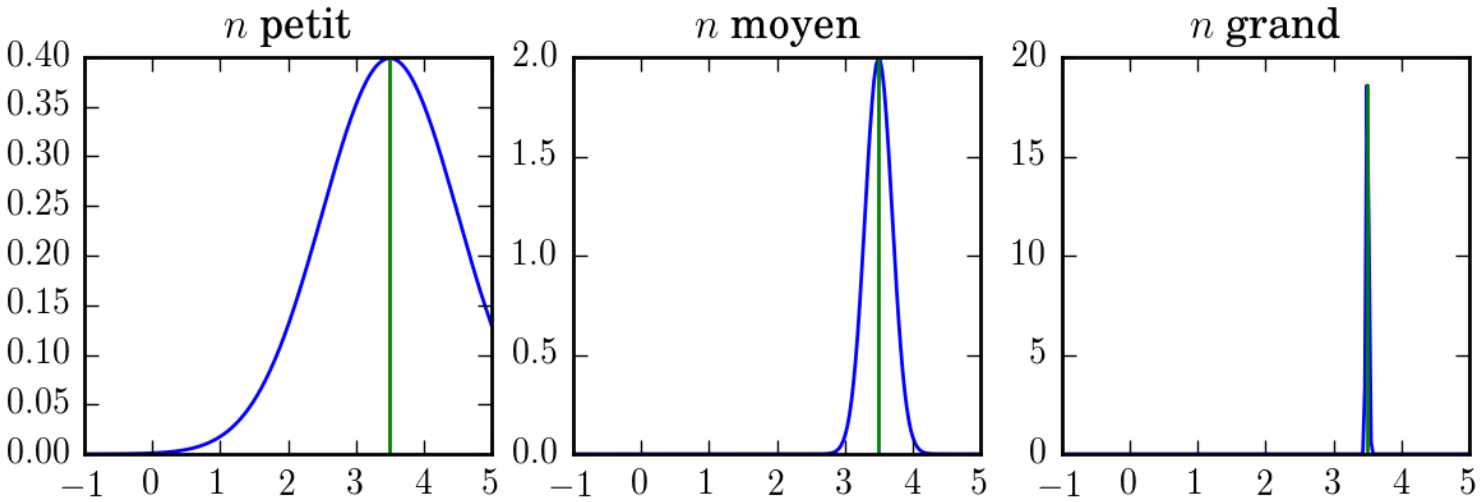
Un estimateur $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}_0$ est consistant, si $\lim_{\sharp\mathcal{D}\to\infty} \mathbb{P}\left(\left|\hat{\boldsymbol{\theta}}\left(\mathcal{D}\right) - \boldsymbol{\theta}_0\right| > \epsilon\right) = 0.$



Théorie fréquentiste la décision : Généralités sur l'estimation

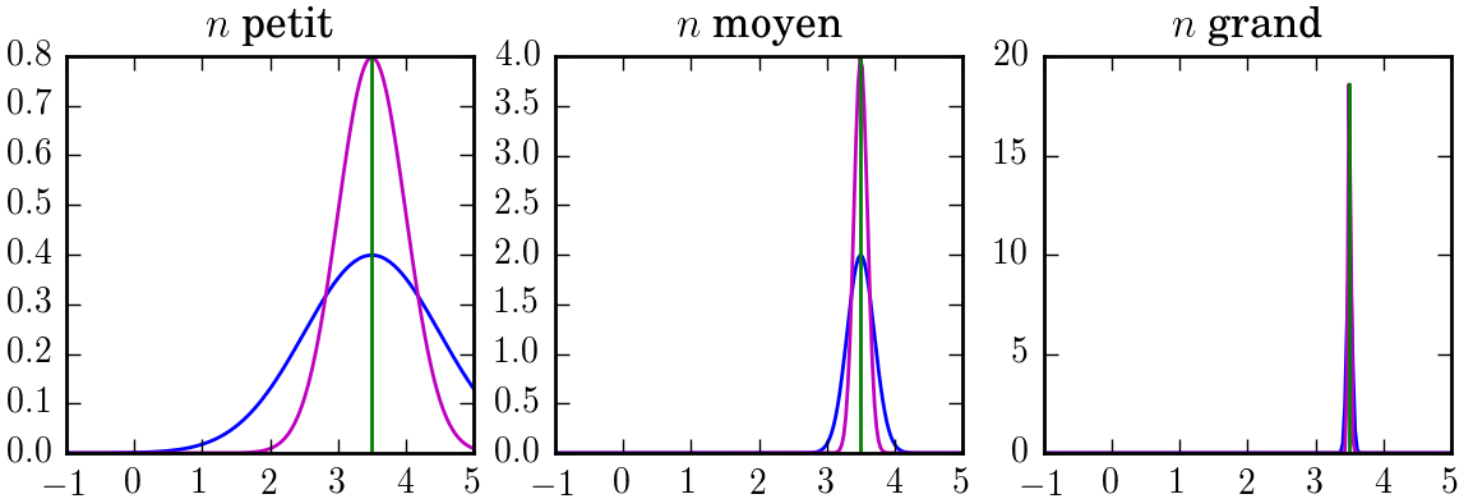
Définition

Un estimateur $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}_0$ est non biaisé, si $\mathbb{E}_{\mathcal{D}|\boldsymbol{\theta}_0}\left[\hat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}_0$.



Théorie fréquentiste la décision : Généralités sur l'estimation

Un autre moyen de comparer la qualité d'un estimateur est la variance



Théorie fréquentiste la décision : Généralités sur l'estimation

Un autre moyen de comparer la qualité d'un estimateur est la variance.

On a le résultat suivant :

Inégalité de Cramer-Rao

Soit un estimateur non biaisé $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}_0$, si la distribution $p_{X|\boldsymbol{\theta}_0}$ est lisse, alors

$$\operatorname{var}\left[\hat{\boldsymbol{\theta}}\right] \geq \frac{1}{n\, I\left(\boldsymbol{\theta}_0\right)}. \tag{27}$$

$I\left(\boldsymbol{\theta}_0\right)$ est l'information de Fisher :

$$I\left(\boldsymbol{\theta}_0\right) = \mathbb{E}\left[-\frac{d^2}{d\theta^2}\log p_{\mathcal{D}|\theta=\boldsymbol{\theta}_0}\right], \tag{28}$$

$$= \mathbb{E}\left[\frac{d^2}{d\theta^2}\mathrm{NLL}\left(\theta=\boldsymbol{\theta}_0\right)\right]. \tag{29}$$

Théorie fréquentiste la décision pour le cas supervisé

- $\rightarrow$ Gros problème : $\boldsymbol{\theta}_0$ est inconnu et le risque n'est pas calculable !

Théorie fréquentiste la décision pour le cas supervisé

- $\rightarrow$ Gros problème : $\boldsymbol{\theta}_0$ est inconnu et le risque n'est pas calculable !
- .. sauf dans le cas particulier où $\boldsymbol{\theta}$ est observable comme dans un problème supervisé !

Théorie fréquentiste la décision pour le cas supervisé

- $\rightarrow$ Gros problème : $\boldsymbol{\theta}_0$ est inconnu et le risque n'est pas calculable !
- Retournons donc à notre problème d'apprentissage supervisé :

Théorie fréquentiste la décision pour le cas supervisé

- $\rightarrow$ Gros problème : $\boldsymbol{\theta}_0$ est inconnu et le risque n'est pas calculable !
- Retournons donc à notre problème d'apprentissage supervisé :
  - $\mathbf{x}$ est un exemple dont la valeur (ou classe) est y.

Théorie fréquentiste la décision pour le cas supervisé

- $\rightarrow$ Gros problème : $\boldsymbol{\theta}_0$ est inconnu et le risque n'est pas calculable !
- Retournons donc à notre problème d'apprentissage supervisé :
  - $\mathbf{x}$ est un exemple dont la valeur (ou classe) est y.
  - y est bien une v.a. (même à $\mathbf{x}$ fixé, il peut y avoir un aléa dans l'étiquetage).

Théorie fréquentiste la décision pour le cas supervisé

- $\rightarrow$ Gros problème : $\boldsymbol{\theta}_0$ est inconnu et le risque n'est pas calculable !
- Retournons donc à notre problème d'apprentissage supervisé :
  - $\mathbf{x}$ est un exemple dont la valeur (ou classe) est y.
  - y est bien une v.a. (même à $\mathbf{x}$ fixé, il peut y avoir un aléa dans l'étiquetage).
  - Nous cherchons une fonction f^* associe à chaque $\mathbf{x}$ individuellement un y de la meilleure manière possible.

Théorie fréquentiste la décision pour le cas supervisé

- $\rightarrow$ Gros problème : $\boldsymbol{\theta}_0$ est inconnu et le risque n'est pas calculable !
- Retournons donc à notre problème d'apprentissage supervisé :
  - $\mathbf{x}$ est un exemple dont la valeur (ou classe) est y.
  - y est bien une v.a. (même à $\mathbf{x}$ fixé, il peut y avoir un aléa dans l'étiquetage).
  - Nous cherchons une fonction f^* associe à chaque $\mathbf{x}$ individuellement un y de la meilleure manière possible.
  - Nous ne sommes pas intéressé par une fonction fonction f^* qui prendrait tout un dataset $\mathcal{D}$ en entrée car ce n'est pas l'usage souhaité.

Théorie fréquentiste la décision pour le cas supervisé

- $\rightarrow$ Gros problème : $\boldsymbol{\theta}_0$ est inconnu et le risque n'est pas calculable !
- Retournons donc à notre problème d'apprentissage supervisé :
  - $\mathbf{x}$ est un exemple dont la valeur (ou classe) est y.
  - y est bien une v.a. (même à $\mathbf{x}$ fixé, il peut y avoir un aléa dans l'étiquetage).
  - Nous cherchons une fonction f^* associe à chaque $\mathbf{x}$ individuellement un y de la meilleure manière possible.
  - Nous ne sommes pas intéressé par une fonction fonction f^* qui prendrait tout un dataset $\mathcal{D}$ en entrée car ce n'est pas l'usage souhaité.
  - Le risque s'écrit alors

$$\mathbb{E}_{\mathbf{x},Y}\left[L\left(y, f\left(\mathbf{x}\right)\right)\right] = \int_{\mathbb{X}}\int_{\mathbb{Y}} L\left(y, f\left(\mathbf{x}\right)\right) p_{\mathbf{x},Y}\left(\mathbf{x}, y\right) d\mathbf{x}dy, \quad (30)$$

$$= Err_{gen} \quad (31)$$

Théorie fréquentiste la décision pour le cas supervisé

- Mais au fond, c'est quoi la différence avec l'approche Bayésienne ?

$$\mathbb{E}_{\mathbf{x},Y}\left[L\left(y,f\left(\mathbf{x}\right)\right)\right] = \int_{\mathbb{X}}\int_{\mathbb{Y}} L\left(y,f\left(\mathbf{x}\right)\right) p_{\mathbf{x},Y}\left(\mathbf{x},y\right) d\mathbf{x}dy, \quad (32)$$

$$= \int_{\mathbb{X}}\int_{\mathbb{Y}} L\left(y,f\left(\mathbf{x}\right)\right) p_{Y|\mathbf{x}}\left(y\right) p_{\mathbf{x}}\left(\mathbf{x}\right) d\mathbf{x}dy, \quad (33)$$

$$= \int_{\mathbb{X}} \rho\left(f\left(\mathbf{x}\right)|\mathbf{x}\right) p_{\mathbf{x}}\left(\mathbf{x}\right) d\mathbf{x}. \quad (34)$$

Théorie fréquentiste la décision pour le cas supervisé

- Mais au fond, c'est quoi la différence avec l'approche Bayésienne ?

$$\mathbb{E}_{\mathbf{x},Y}\left[L\left(y,f\left(\mathbf{x}\right)\right)\right] = \int_{\mathbb{X}}\int_{\mathbb{Y}} L\left(y,f\left(\mathbf{x}\right)\right) p_{\mathbf{x},Y}\left(\mathbf{x},y\right) d\mathbf{x}dy, \quad (32)$$

$$= \int_{\mathbb{X}}\int_{\mathbb{Y}} L\left(y,f\left(\mathbf{x}\right)\right) p_{Y|\mathbf{x}}\left(y\right) p_{\mathbf{x}}\left(\mathbf{x}\right) d\mathbf{x}dy, \quad (33)$$

$$= \int_{\mathbb{X}} \rho\left(f\left(\mathbf{x}\right)|\mathbf{x}\right) p_{\mathbf{x}}\left(\mathbf{x}\right) d\mathbf{x}. \quad (34)$$

2 worlds collide !

Théorie fréquentiste la décision pour le cas supervisé

- Minimiser la perte attendue au sens bayésien et fréquentiste est équivalent dans notre cas !
- Cela reste valable pour tout risque intégré du type :

$$\int \text{Risque}\left(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}\right) p\left(\boldsymbol{\theta}_0\right) d\boldsymbol{\theta}_0. \tag{35}$$

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.
- Les approches Bayésiennes et fréquentistes sont équivalentes dans le contexte supervisé.

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

$$\begin{aligned} \mathbb{E}[B|X = \mathbf{x}_i] &= 0, \\ \Leftrightarrow \mathbb{E}[Y|X = \mathbf{x}_i] &= f_0(\mathbf{x}_i). \end{aligned} \tag{36}$$