Evaluation / sélection de modèles

I / Cadre statistique

- notre but est d'approcher f, défini de la façon suivante :
  → Soit P une mesure de probabilité sur $\mathbb{R}^p$:
  $$\min_{f \in \mathcal{F}} \mathbb{E}[\ell(Z, f)] = \min_{f \in \mathcal{F}} \int \ell(Z, f)\, d P(Z)$$
  → ℓ : fonction de perte : "loss function"
  → $(f, \mathcal{F})$: une fonction de regression
  un classifieur
  une densité associée à une certaine loi

- exemple : régression (linéaire)
  → $\ell(Z, f) = (y - f(x))^2$ avec $Z = (y, x)$
  $\mathcal{F} = \{ x \longmapsto \beta^T x = \beta \in \mathbb{R}^d \}$

- on peut considérer d'autres "loss functions"
  → ex : $\ell(y, f) = |y - f(x)|$
  non dérivable, va poser problème pour résoudre problème d'optim, mais moins sensible aux outliers (utile lorsque la queue de distribution est plus importante qu'une gaussienne)
  → d'autres classes de fonctions :
  $\mathcal{F} = \{ x \longmapsto \sum \alpha_i K(x, x_i) : \alpha \in \mathbb{R}^n \}$
  SVR : support vector regression, noyau SVM généralisé à la régression
  → $\mathcal{F} = \{ x \longmapsto f_\theta(x) = f_\theta$ la fonction d'un certain réseau de neurones $\}$
  → classification (similaire à la régression)
  → estimation de densité
  $\mathcal{F} = \{ f_\theta : \theta \in \Theta \}$: une collection de densités
  $\ell(Z, f) = -\log(f(Z))$ (voir distance de Kullback)

$$\min_{\ell} -\mathbb{E}[\log(f(x))] \Leftrightarrow \max_{\ell} \mathbb{E}[\log(f) - \log(f_0)]$$

avec f_0 la densité de $\mathbb{P}$

$$\Leftrightarrow \max_{\ell} \mathbb{E}(\log(f(Z))] \Leftrightarrow \max_{\ell} \int \log \frac{f(x)}{f_0(x)} f_0(x)\,dx$$

- On a le résultat suivant :

$$\forall f, \int \log(f(x)/f_0(x))\, f_0(x)\,dx \leqslant 0$$

avec égalité si et seulement si $f = f_0$.

- On suppose observer un échantillon $(Z_i)_{i=1\dots n}$ de variables indépendantes et identiquement distribuées sous $\mathbb{P}$.

→ on applique la "minimisation du risque empirique"

$$\hat{f} \in \underset{f \in \mathcal{F}_\ell}{\operatorname{argmin}} \boxed{\frac{1}{n} \sum_{i=1}^{n} \ell(Z_i, f)}$$

→ def : $\boxed{R(f) = \mathbb{E}[\ell(Z, f)]}$

$$\boxed{\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i, f)}$$

- théorie de l'estimation : $\hat{f} \approx f$?

→ $\mathbb{E}(\hat{f}) \approx f$

→ $\mathrm{Var}(\hat{f}) \leqslant ?$

→ questions classiques, mais ici on veut poser une question différente : celle de l'évaluation de la qualité de $\hat{f}$.

II) Correction du biais

Approche : la qualité de $\hat{f}$ sera évaluée par :
- $\hat{R}(\hat{f})$? non car on valorise juste le plus grand modèle, celui avec toutes les variables actives.
- $R(\hat{f})$: oui !

problème : $R(\hat{f})$ est une quantité difficile à évaluer.
→ l'estimateur naturel $\hat{R}(\hat{f})$ est biaisé.
→ intuition : pour tout $f \in \mathcal{F}$: $\mathbb{E}[\hat{R}(f)] = R(f)$
pour tout $f \in \mathcal{F}$, $\hat{R}(f)$ est non biaisé
mais $\hat{R}(\hat{f}) = \min_{f \in \mathcal{F}} \hat{R}(f)$, on attend le résultat suivant :
$\hat{R}(\hat{f}) \leq R(\hat{f})$ la tendance

Il est difficile de vérifier cela dans un cadre général. Vérifions le pour les moindres carrés :

$\hat{\theta} \in \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\theta\|_2^2$ avec $Y \in \mathbb{R}^n$ $X \in \mathbb{R}^{n \times p}$

on sait $X^T X \hat{\theta} = X^T Y$

$\Leftrightarrow X^T (X\hat{\theta} - Y) = 0$

$\Leftrightarrow X^T (X(\hat{\theta} - \theta^*) + (X\theta^* - Y)) = 0$

$\Leftrightarrow \boxed{X^T (X(\hat{\theta} - \theta^*)) = X^T \varepsilon}$

Regardons le comportement de $\hat{R}(\hat{f})$ dans ce cadre.

$$\hat{R}(\hat{f}) = \frac{\|Y - X\hat{\theta}\|_2^2}{n} = \frac{\|\varepsilon + X(\theta^* - \hat{\theta})\|_2^2}{n} = \frac{\|\varepsilon\|_2^2 + 2\langle \varepsilon, X(\theta^* - \hat{\theta})\rangle + \|X(\theta^* - \hat{\theta})\|_2^2}{n}$$

$= -2\langle \varepsilon, X(\hat{\theta} - \theta^*)\rangle$; $= \langle \varepsilon, X(\hat{\theta} - \theta^*)\rangle$

(car $\|X(\theta^* - \hat{\theta})\|_2^2 = (\theta^* - \hat{\theta})^T X^T X (\theta^* - \hat{\theta}) = -(\theta^* - \hat{\theta})^T X^T \varepsilon = \langle \varepsilon, X(\hat{\theta} - \theta^*)\rangle$)

$$= \frac{\|\varepsilon\|_2^2 - \langle \varepsilon, X(\hat{\theta} - \theta^*)\rangle}{n}$$

On obtient :

$$= \frac{\|\varepsilon\|_2^2 - \|X(\theta^* - \hat{\theta})\|_2^2}{n}$$

le risque empirique est biaisé inférieurement

$X, \hat{\theta}$

On cherche à estimer $R(\hat{f}) = \mathbb{E}_Z[\ell(Z, \hat{f})] = \mathbb{E}_{Z}[(Y - \hat{f}(X))^2]$

$= \mathbb{E}[(\varepsilon_1 + X_1^T(\theta^* - \hat{\theta}))^2] = \mathbb{E}[\varepsilon_1^2] + 2\underbrace{\mathbb{E}[\varepsilon_1 X_1^T](\theta^* - \hat{\theta})}_{=0} + \mathbb{E}[(X(\theta^* - \hat{\theta}))^2]$

$= \mathbb{E}[\varepsilon_1^2] + \mathbb{E}[(X_1^T(\theta^* - \hat{\theta}))^2]$

cadre asymptotique : $n \to \infty$

$\frac{\|X(\hat{\theta}-\theta^*)\|_2^2}{n} = (\hat{\theta}-\theta^*)\hat{\Sigma}_n(\hat{\theta}-\theta^*)$ avec $\hat{\Sigma}_n = \frac{X^TX}{n}$

$= \|\hat{\Sigma}_n^{1/2}(\hat{\theta}-\theta^*)\|_2^2$

Théorie MCO : $\sqrt{n}(\hat{\theta}-\theta^*) \to \mathcal{N}(0, \sigma^2\Sigma^{-1})$ $\Sigma = \mathbb{E}(X_1X_1^T)$

Slutsky $\Rightarrow \hat{\Sigma}_n^{1/2}\sqrt{n}(\hat{\theta}-\theta^*) \to \mathcal{N}(0, \sigma I_d)$

Théorème de continuité :

$\|\hat{\Sigma}_n^{1/2}\sqrt{n}(\hat{\theta}-\theta^*)\|_2^2 \to \|\mathcal{N}(0,\sigma^2 I)\|_2^2 = \sigma^2 \chi_d^2$

Ainsi, il y a bien un biais asymptotique car :

$\mathbb{E}[\chi_d^2] = \mathbb{E}[\sum_{i=1}^{d} Z_i^2] = d$

Le C_p de Mallows est défini par le biais $\hat{R}(\hat{f}) - R(\hat{f})$:

$\mathbb{E}[\hat{R}(\hat{f}) - R(\hat{f})] \approx \mathbb{E}[\|\varepsilon\|_2^2 - \|X(\theta^*-\hat{\theta})\|_2^2 - \|\varepsilon\|_2^2 - \|X^T(\theta^*-\hat{\theta})\|_2^2] = \frac{-2\|X^T(\theta^*-\hat{\theta})\|_2^2}{n}$

$= \frac{-2\chi_d^2\sigma^2}{n}$

$C_p = \frac{\|Y - X\hat{\theta}\|_2^2}{n} + \frac{2d}{n}\hat{\sigma}^2$ ⟸

Autre critère similaire : AIC, BIC, GIC

- AIC : similaire à Mallows mais dans un contexte plus général de vraisemblance :

$\hat{\theta}_n \in \operatorname{argmax}_\theta Lik(\theta) = \operatorname{argmax}_\theta \sum_{i=1}^{n} \log(f_\theta(z_i))$

$\Rightarrow AIC = 2\,Lik(\hat{\theta}) - 2\,\underbrace{length(\theta)}_{\text{dim du modèle}}$ nbr de paramètres de programme à optim

- BIC, GIC : tout comme le Mallows, BIC et GIC ont un terme correcteur du biais.

- $Lik(\theta) = \exp\left(-\frac{1}{2}(y - \beta^T x)(y - \beta^T x)/\sigma^2\right)$

III) Petite parenthèse sur l'utilisation de ces critères/méthodes

- $R(\hat{f})$ = comment $\hat{f}$ résout notre problème) model assessment
  → quantités importantes en elle-même

- Ces estimations de $R(f)$ permettent de faire de la sélection en comparant différents modèles. On préfère $\hat{f}_1$ à $\hat{f}_2$ si $R(\hat{f}_1) \leq R(\hat{f}_2)$

- exemple : → sélection forward ; on compare chaque modèle correspondant à un jeu de variables sélectionnées.
  → dès qu'on a un hyperparamètre à régler ou définir par exemple Lasso $\|Y - X\hat{\theta}\|_2^2 + \lambda\|\theta\|_1$
  support (variable active) qui permet de calculer le Mallows
  ⚠ le choix | du critère
  de la méthode d'estimation de $R(\hat{f})$ est importante
  Exc : $\hat{R}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_2(x_i))$ $\hat{f}_2$ = estimateur à plus proches voisins
  argmin $\hat{R}(\hat{f}) = 1$ car ici $\hat{f}(x_i) = x_i$

IV) Validation croisée

- le principe : séparer les données en 2 parties :
  → une pour apprendre le modèle $\hat{f}$
  → une pour évaluer le risque $R(\hat{f})$

- utilisation : plus large que AIC, BIC, etc ... car libre de tout modèle. Ici on peut régler n'importe quel hyperparamètre

a) le hold-out

- Soit $\begin{cases} S \subset \{1, \dots, n\} & \text{l'échantillon de validation} \\ S^c & \text{le complémentaire de } S \text{ dans } \{1 \dots n\} : \text{l'échantillon d'apprentissage} \end{cases}$

On définit : $\hat{f}_{S^c} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i \in S^c} \ell(Z_i, f)$

- L'estimateur "Hold-out" du risque de $\hat{f}$, $R(\hat{f})$, est $\hat{R}_{HO} = \frac{1}{|S|} \sum_{i \in S} \ell(Z_i, \hat{f}_{S^c})$ $\quad |S|$: cardinalité de S.

- 1er résultat : $\mathbb{E}[\hat{R}_{HO}] = \frac{1}{|S|} \sum_{i \in S} \mathbb{E}\left[\ell(Z_i, \hat{f}_{S^c})\right]$

2 sources d'aléatoire : Z_i, $\hat{f}_{S^c} = T(S^c)$; indépendantes car $S \cap S^c = \emptyset$

$$= \frac{1}{|S|} \sum_{i \in S} \mathbb{E}\left[\mathbb{E}\left[\ell(Z_i, \hat{f}_{S^c}) \mid S^c\right]\right]$$

$$= \frac{1}{|S|} \sum_{i \in S} \mathbb{E}\left[\mathbb{E}_2\left[\ell(Z_i, \hat{f}_{S^c})\right]\right]$$

espérance par rapport à Z_i uniquement

$$= \frac{1}{|S|} \sum_{i \in S} \mathbb{E}\left[R(\hat{f}_{S^c})\right]$$

$$= \mathbb{E}\left[R(\hat{f}_{S^c})\right]$$

- intuition :

$$\hat{R}_{HO} - R(f) = \underbrace{\hat{R}_{HO} - \mathbb{E}[\hat{R}_{HO}]}_{\text{forme de variance car on peut calculer :}} + \underbrace{\mathbb{E}[\hat{R}_{HO}] - R(f)}_{\mathbb{E}[R(\hat{f}_{S^c}) - R(f)] \geq 0 \text{ si } S^c = \{1 \dots n\} \text{ full sample (décroissante quand } |S^c| \nearrow)}$$

$$\Downarrow$$

$$\operatorname{Var}(\hat{R}_{HO}) \approx O\left(\frac{1}{|S|}\right)$$

incompatibilité entre les 2 directions car $|S| + |S^c| = n$

- Il en résulte un trade-off entre $|S|$ et $|S^c|$

En pratique, il semble plus raisonnable de choisir $|S|$ petit par rapport à $|S^c|$ car l'apprentissage de $\hat{f}$ est souvent plus compliqué que celui de $\hat{R}(f)$ qui est une simple moyenne empirique.

En général, on choisit $|S| \approx 10\%$ de n

b) Algorithmique

- à partir du simple hold-out, on peut définir les algorithmes suivants, qui sont des classiques:

- K-fold : → on suppose que $n/K \in \mathbb{N}$
  → on définit $S_l = \{ z_{l(n/K)+1} \dots z_{(l+1)(n/l)} \}$ $l = 0 \dots K-1$

S_l : la $l^{ième}$ fold ; chaque fold contient n/K points et $S_i \cap S_j = \phi$

aussi : $\overset{K}{\underset{l=1}{\cup}} S_l = \{z_1 \dots z_n\}$ l'échantillon entier

→ $\hat{R}_{K\text{-fold}} = \frac{1}{K} \sum_{l=1}^{K} R_{HO}(S_l)$

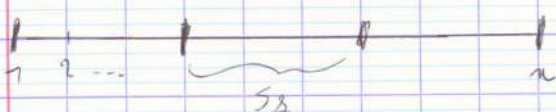
→ étape préliminaire à l'algo : mélange aléatoire uniforme afin d'éliminer une dépendance possible du classement

Cas particulier : leave one out : $\hat{R}_{LOO} = \frac{1}{n} \sum_{i=1}^{n} R_{HO}(z_i)$

problème : temps de calcul très long bien que dans certains cas comme celui de la régression linéaire, des simplifications existent ; on n'a alors pas besoin de calculer chaque fold.

- leave p-out :

→ $\hat{R}_{LPO} = \frac{1}{\binom{n}{p}} \sum_{S \in S_p} R_{HO}(S)$ avec S_p : ensemble des sous-échantillons de taille p.



Dans le LPO, on apprend sur $n-p$ points tout comme avec K-fold si $n - (n-1)(K/n) = p$. Il y a plus d'échantillons par rapport à K-fold.

- validation croisée par Monte Carlo
  → K-fold + tirage aléatoire sur les autres échantillons
  → si $(R_l)_{l=1 \dots K}$ est une suite de variables aléatoires identiquement distribuées (pas nécessairement indépendantes) alors :

$$Var\left(\frac{1}{K}\sum_{l=1}^{K} R_l\right) = \mathbb{E}\left[\left(\frac{1}{K}\sum_{l=1}^{K} R_l - \mathbb{E}(R_l)\right)^2\right] \leq \mathbb{E}\left[\frac{1}{K}\sum_{l=1}^{K}\left((R_l - \mathbb{E}(R_l))^2\right)\right] = \frac{1}{K}\sum_{l=1}^{K} Var(R_l) = Var(R_l)$$

Jensen

Monte Carlo : les exemples de validation sont tirés uniformément. Ex : si $|S| = 100$, et $K = 10$ alors un validation set sera de taille 10, et tiré aléatoirement avec Monte Carlo, comme par exemple : 2, 45, 79, 53, 37, 87, 91, 28, 48, 74