Séance 7: Évaluation de la qualité de prédiction

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Septième partie VII

Évaluation de la qualité de prédiction

Objectifs

- Comment évaluer la performance d'un modèle statistique ?
- On dispose de données $\mathcal{D}$ « étiquetées »$\mathcal{D} = (X_1, Y_1) \dots (X_n, Y_n)$
- Évaluer de façon fiable la performance (fiabilité, confiance) d'un modèle est important pour ensuite pouvoir *choisir* le meilleur.
- On propose généralement trois stratégies :
  - une possibilité de partager $\mathcal{D}$ en deux parties : l'une pour l'apprentissage du modèle, l'autre pour le test

$$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}$$

  - une pénalisation du biais par la complexité du modèle lors de la phase d'ajustement (apprentissage) du modèle
  - un usage intensif de calcul par simulations statistiques complexes
- Quel modèle pour obtenir la meilleure fiabilité de prédiction
- Cas idéal : on possède deux échantillons, *train* et *test*
- Cas standard : on possède des données non séparées en *train* et *test*. Décomposition de $\mathcal{D}$ ou stratégie de pénalisation
- Cas plus pénible : on possède peu d'échantillons d'apprentissage : utilisation de simulations

Objectifs

Plusieurs considérations :

- Consistence du modèle : un modèle est consistant si lorsque la taille des données tend vers $+\infty$, l'erreur d'apprentissage tend vers l'erreur de test en probabilité.
- Vitesse de convergence : Évaluation de la faculté de généralisation de l'ensemble d'apprentissage lorsque sa taille augmente.
- Contrôle du modèle : Estimation de la capacité de généralisation du modèle lorsque le nombre d'exemples est fixé.

On notera que le « meilleur »modèle au sens prédictif n'est pas forcément

- celui qui s'ajuste le mieux aux données d'apprentissage (*overfitting*)
- le vrai modèle si la variance des estimations est trop importante

Définitions

X : variables prédictives et Y la variable à prédire.
On suppose donné $\mathcal{D}$ et qu'il y a une loi jointe F entre X et Y. Le modèle statistique s'écrit

$$Y = \phi(X) + \epsilon$$

On suppose ϵ centré, indépendant de X et on note $Var(\epsilon) = \sigma^2$.
Erreur de prédiction : L'erreur de prédiction du modèle est définie par

$$\mathcal{E}(\mathcal{D}) = \mathbb{E}_F \left[Q(Y, \hat{\phi}(X)) \right]$$

où Q est la fonction de perte.
Interprétation : C'est l'erreur mesurée par Q si les observations (X, Y) étaient générées par la loi jointe F alors que le modèle appris sur $\mathcal{D}$ est $\hat{\phi}$. Penser à l'exemple simple de la régression linéaire...

Fonctions de perte

- Cas quantitatif : (pour de la régression) la variable Y est par exemple réelle, dans ce cas, on prend en général la fonction de perte quadratique :

$$Q(y, \hat{y}) = (y - \hat{y})^2$$

Il est également possible de choisir

$$Q(y, \hat{y}) = |y - \hat{y}|$$

moins sensible aux valeurs extrêmes mais plus complexes à manipuler théoriquement.

- L'erreur de prédiction est dans ce cas la fonction de perte quadratique moyenne observée.
- Cas qualitatif : (pour de la classification) la variable Y vaut par exemple $\{0; 1\}$ et dans ce cas :

$$Q(y, \hat{y}) = \mathbf{1}_{\mathbf{y} \neq \hat{\mathbf{y}}}$$

- L'erreur de prédiction est dans ce cas l'erreur de classification moyenne.

Rappel de la décomposition Biais/Variance

- Dans le cas quadratique (c'est également vrai dans un cadre plus général), on a :

$$\mathcal{E} = \sigma^2 + Biais^2 + Variance^2$$

- Plus la famille de fonctions ϕ est riche, plus le biais est réduit.
- Mais la variance augmente(!) avec le nombre de paramètres à estimer.
- D'où le compromis à effectuer : accepter de biaiser un peu l'estimation de ϕ pour réduire la variance (régression ridge par exemple).

Estimation de l'erreur pour n grand

- On considère tout d'abord la qualité d'ajustement du modèle sur l'échantillon observé.
- C'est une estimation optimiste qui est biaisée puisque liée aux observations sur lesquelles on a calculer l'estimateur.
- On note cette estimation :

$$\hat{\mathcal{E}}_P = \frac{1}{n}\sum_{i=1}^{n} Q(Y, \hat{\phi}(X_i))$$

- Pour mesurer le biais sans erreur, le plus simple est de disposer de trois ensembles d'échantillons :

$$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{valid} \cup \mathcal{D}_{test}$$

  - $\hat{\mathcal{E}}_P(\mathcal{D}_{train})$ utilisée pour déterminer $\hat{\phi}$.
  - $\hat{\mathcal{E}}_P(\mathcal{D}_{valid})$ utilisée pour choisir le meilleur modèle parmi une famille de modèles
  - $\hat{\mathcal{E}}_P(\mathcal{D}_{test})$ utilisée pour calculer la "vraie" erreur de la méthode

Estimation de l'erreur pour n grand

Problème de cette méthode lorsque n n'est pas assez grand :

- Pour bien « conceptualiser » : penser au modèle de régression où il faut déterminer le degré optimal de l'estimateur.
- Si n petit, l'estimateur peut ne pas avoir assez d'échantillon dans $\mathcal{D}_{train}$ pour être performant.
- La variance de l'estimateur ($\hat{\phi}$) est inaccessible.
- Lorsque n est petit, on peut au moins supprimer l'échantillon de validation et utiliser des technique de simulations.

Stratégie avec pénalisation : C_p de Mallows

- Historiquement : premier critère de pénalisation utilisé.
- Cadre : modèle linéaire où l'ajustement (ou R^2) n'est pas le seul critère retenu.
- L'« énergie »à minimiser s'écrit :

$$\mathcal{E}_p = \mathcal{E}(\mathcal{D}) + Penalite$$

- Le terme de pénalité corrige l'abus d'optimisme. On estime ce terme *via* la décomposition Biais/Variance.
- On a démontré (en 2001) qu'une bonne « énergie »peut être

$$\mathcal{E}_p = \mathcal{E}(\mathcal{D}) + \frac{2d}{n} s^2$$

- d est le nombre de paramètres du modèle (p+1 si on utilise un modèle linéaire sur p variables).
- n nombre d'observations de $\mathcal{D}_{train}$.
- s^2 estimation empirique sur $\mathcal{D}_{train}$ de la variance de l'erreur

Stratégie avec pénalisation : AIC et BIC

- On cherche à maximiser la vraisemblance $\mathcal{L}$ d'un modèle.
- On suppose que la famille de modèles possibles contient le « vrai »modèle.
- L'« énergie »à minimiser s'écrit :

$$AIC = -log(\mathcal{L}) + 2\frac{d}{n}$$

- On montre que dans le cas de modèles gaussiens, C_p et AIC sont équivalents.
- Dans le cas de petits échantillons, on préfère souvent utiliser la fonction de coût :

$$AIC = -log(\mathcal{L}) + 2\frac{n+d}{n-d-2}$$

- Un autre critère, BIC (Bayesian Information Criterion), utilise la fonction

$$BIC = -log(\mathcal{L}) + 2log(n)\frac{d}{n}$$

Stratégie avec pénalisation : choix de la pénalisation

- Pour n grand, BIC pénalise lourdement les modèles complexes.
- On montre que si $n \longmapsto \infty$, la probabilité que BIC sélectionne le « vrai »modèle tend vers 1.
- Ce n'est pas le cas pour AIC et C_p.
- Pour n petit, BIC se limite souvent a des modèles trop simples.
- Pour des modèles non-linéaires, on optimise souvent

$$Energie = f(Vraisemblance) + Penalisation(d)$$

ou

$$Energie = f(Erreurempirique) + Penalisation(d)$$

Validation croisée

- Très classique et utilisé, très simple à mettre en oeuvre
- Complexité de calcul conséquente
- Idée : simuler plusieurs échantillons de validation et calculer la moyenne des erreurs sur ces échantillons.
- Algorithme :
  - Découper $\mathcal{D}$ en k échantillon (« *k-folds cross validation* ») de tailles à peu près égales avec une loi uniforme sur $\mathcal{D}$. On obtient :

$$\mathcal{D}_1, \dots \mathcal{D}_k$$

  - Pour $i = 1 \longmapsto l$, isoler $\mathcal{D}_i$ et estimer le modèle sur les $k-1$ ensembles $(\mathcal{D} \setminus \mathcal{D}_i)$
  - Calculer l'erreur sur $\mathcal{D}_i$, modèle appris sur $\mathcal{D} \setminus \mathcal{D}_i$
  - Moyenne toutes ces erreurs
- $k = 5$ ou $k = 10$ sont le plus courant.
- Dans le cas où $k = n$, on obtient le « *Leave one out* »(Faisable si n est petit).
-

Estimation par bootrap

- Estimateur naïf :
  - On tire B échantillons par simulation uniforme sur $\mathcal{D}$: $\mathcal{D}_1, \dots \mathcal{D}_B$
  - Pour chaque échantillon, on détermine l'estimateur ϕ_i
  - On calcule l'erreur de ϕ_i en effectuant :

$$\hat{\mathcal{E}}_i = \frac{1}{n} \sum_{j=1}^{n} Q(Y_j, \hat{\phi}_i(X_j))$$

  - On effectue la moyenne sur les B échantillons bootstrap.
- Estimateur « *Out Of Bag* »
  - On tire B échantillons par simulation uniforme sur $\mathcal{D}$: $\mathcal{D}_1, \dots \mathcal{D}_B$
  - Pour chaque échantillon, on détermine l'estimateur ϕ_i
  - On calcule l'erreur de ϕ_i en effectuant

$$\hat{\mathcal{E}}_i = \frac{1}{n_i} \sum_{j \notin \mathcal{D}_i}^{n} Q(Y_j, \hat{\phi}_i(X_j))$$

  - On calcule la moyenne : $\mathcal{E}_{oob} = \frac{1}{B} \sum_{i=1}^{B} \hat{\mathcal{E}}_i$

Estimation par bootstrap

Estimateur Estimateur « $\mathcal{E}.632 - bootstrap$ »

- On tire B échantillons par simulation uniforme sur $\mathcal{D}$: $\mathcal{D}_1, \ldots \mathcal{D}_B$
- Pour chaque échantillon, on détermine l'estimateur ϕ_i
- On détermine alors $\mathcal{E}_{oob}$ et $\mathcal{E}_{intern}$
- On calcule :

$$\mathcal{E}_{.632-bootstrap} = 0.632\mathcal{E}_{oob} + 0.368\mathcal{E}_{intern}$$

- Compense l'excès de pessimisme de $\mathcal{E}_{oob}$ et l'excès d'optimisme de $\mathcal{E}_{intern}$.

En conclusion :

- Pour comparer 2 méthodes, toujours utiliser la même méthode d'estimation d'erreur !
- Se montrer prudent sur le caractère « absolu »d'une estimation d'erreur.