Apprentissage: cours 3
Validation croisée
Consistance uniforme
Théorème No Free Lunch

Guillaume Obozinski

27 février 2013

1 Validation croisée

1.1 Sélection de l'algorithme d'apprentissage

Données d'entrainement : $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$

Algorithme d'apprentissage : $\mathscr{A} : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathcal{F}$

Famille d'algorithmes d'apprentissage : $(\mathscr{A}_m)_{m \in \mathcal{M}}$

Famille de prédicteurs $\left(\widehat{f}_m\right)_{m \in \mathcal{M}}$

Exemples :
- k-plus proches voisins pour différent k
- Nadaraya-Watson avec différent noyaux et différentes largeurs de bande
- régression polynomiale de différent degrés
- histogrammes pour différentes partition
- régression linéaire sur la base de plusieurs sous-ensembles de variables

Dans ce cours par abus de notation on écrira souvent $\widehat{f}$ pour $\mathscr{A}$ et $\widehat{f}(D_n)$ pour $\mathscr{A}(D_n)$. Pour être rigoureux, il faudrait toujours utiliser $\widehat{f}_{D_n} := \mathscr{A}(D_n)$.

Excès de risque : $\mathcal{R}\left(\widehat{f}_m(D_n)\right) - \mathcal{R}\left(f^\star\right)$

Problème Sélection de l'algorithme d'apprentissage, sélection des *hyperparamètres*, sélection du modèle, méta-apprentissage.

Enjeu Compromis entre sur-apprentissage et sous-apprentissage.

1.2 Validation simple

Soit $\widehat{f}$ un prédicteur. On cherche à estimer $\mathcal{R}\left(\widehat{f}(D_n)\right)$, à l'aide des données D_n uniquement (estimation dont on se servira ensuite pour résoudre le problème de sélection). On sépare les données D_n en deux ensembles non-vides :

Définition 1 (Données d'entraînement *vs* données de validation)**.** Soit I_n^v un sous-ensemble de $\{1, \dots, n\}$ tel que $0 < n_v := |I^e| < n$ et I^e son complémentaire, avec $n_e = |I^e|$ On définit

Données d'entraînement $D_n^e = \{(X_i, Y_i)\}_{i \in I^e}$

Données de validation $D_n^v = \{(X_i, Y_i)\}_{i \in I^v}$

Définition 2 (Validation simple)**.** On définit l'estimateur par validation simple du risque :

$$\widehat{\mathcal{R}}^{\text{val}}(\widehat{f}; D_n; I^v) := \frac{1}{|I^v|} \sum_{i \in I^v} \ell\left(\widehat{f}_{D_n^e}(X_i), Y_i\right) \quad \text{avec} \quad D_n^e = \{(X_i, Y_i)\}_{i \notin I^v}$$

1.3 Validation croisée

Définition 3 (Validation croisée)**.** Si pour $j \in \{1, \dots, B\}$, I_j^v est un sous-ensemble propre de $\{1, \dots, n\}$, on définit l'estimateur par validation croisée :

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}; D_n; (I_j^v)_{1 \leqslant j \leqslant B}\right) := \frac{1}{B} \sum_{j=1}^{B} \widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}; D_n; I_j^v).$$

Définition 4 (Validation croisée k-fold)**.** Si $(B_j)_{1 \leqslant j \leqslant V}$ est une partition de $\{1, \dots, n\}$,

$$\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}; D_n; (B_j)_{1 \leqslant j \leqslant k}\right) := \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}; D_n; (B_j)_{1 \leqslant j \leqslant k}\right)$$

On sous-entend généralement que la partition est uniforme de sorte que $\lfloor n/k \rfloor \leqslant |B_j| \leqslant \lceil n/k \rceil$.

Définition 5 (Leave-one-out)**.**

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{f}; D_n\right) := \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}; D_n; (\{j\})_{1 \leqslant j \leqslant n}\right)$$

1.4 Propriétés de l'estimateur par validation croisée du risque

Biais

Proposition 1 (Espérance d'un estimateur par validation croisée du risque)**.** *Soit $\widehat{f}$ un algorithme d'apprentissage et $I_1^e, \dots, I_B^e$ des sous-ensembles propres de $\{1, \dots, n\}$ de même cardinal n_e. Alors,*

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}; D_n; (I_j^v)_{1 \leqslant j \leqslant B}\right)\right] = \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_{D_n^e}\right)\right] \tag{1}$$

où D_{n_e} désigne un ensemble de n_e observations indépendantes de même loi P que les $(X_i, Y_i) \in D_n$.

Variance

- Pour la validation simple :

$$\mathrm{var}\left(\widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}; D_n; I^v)\right) = \frac{1}{n_v}\mathbb{E}\left[\mathrm{var}\left(\ell(\widehat{f}_{D_n^e}(X), Y)\right) \,\middle|\, D_n^e\right] + \mathrm{var}\left(\mathcal{R}\left(\widehat{f}_{D_n^e}\right)\right)$$

- Facteurs de variabilité : taille n_v de l'ensemble de validation (l'augmenter fait diminuer la variance, à n_e fixe du moins), "stabilité" de $\mathcal{A}$ (pour un ensemble de taille n_e), nombre B de découpages considéré.
- En général, la variance est difficile à quantifier précisément, car n_e et n_v sont toujours liés ($n_e + n_v = n$), et parfois B leur est lié également (e.g., k-fold).

1.5 Sélection d'algorithme par validation croisée

- Définition :

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (I_j^e)_{1 \leqslant j \leqslant B}\right)\right\}$$

- Pourquoi cela peut fonctionner :
  Principe de l'estimation sans biais du risque et Proposition 1.
- Choix d'une méthode de validation croisée : compromis entre temps de calcul et précision.
- Estimation du risque de l'estimateur final : découpage en trois sous-ensembles (entraînement, validation et test).

2 Consistance uniforme *vs* universelle

Définition 6 (Consistance et universelle consistance)**.** On dit qu'un algorithme d'apprentissage est consistant pour la loi P si

$$\lim_{n\to\infty} \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P(\widehat{f}) - \mathcal{R}_P(f_P^\star)\right] = 0.$$

On dit qu'il est universellement consistant s'il est consistant pour tout P.

Définition 7 (Consistance uniforme)**.** Soit $\mathcal{P}$ un ensemble de distributions sur les données. On dit qu'un algorithme d'apprentissage est *uniformément consistant* sur $\mathcal{P}$ si

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P(\widehat{f}) - \mathcal{R}_P(f_P^\star)\right] = 0.$$

La différence entre les consistances universelles et uniformes c'est essentiellement qu'on a échangé supremum et limite.

La difficulté de l'apprentissage pour une classe de distribution $\mathcal{P}$ est mesurée par sa complexité en quantité de données ou *sample complexity*.

Définition 8. (Sample complexity) Soit $\varepsilon > 0$, on appelle *complexité en quantité de données*, le plus petit nombre $n(\mathcal{P},\varepsilon)$ tel que, pour tout $n \geqslant n(\mathcal{P},\varepsilon)$ on a

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P(\widehat{f})\right] - \mathcal{R}_P(f_P^\star) < \varepsilon.$$

Les théorèmes "No free lunch" – nous en verrons un dans la suite de ce cours – prouvent qu'il n'y pas de consistance universellement uniforme dès que le problème d'apprentissage est suffisamment riche, typiquement dès que $\mathcal{X}$ est infini.

On ne pourra donc pas montrer d'inégalité du type

$$\forall P\in\mathcal{P}, \qquad \mathbb{E}_P\left[\mathcal{R}_P(\widehat{f})\right] \leqslant \mathcal{R}_P(f_P^\star) + \varepsilon_n$$

pour $\mathcal{P}$ sera l'ensemble des distributions possibles.

En revanche, si on se donne un séquence de modèles $\mathcal{F}_m$ (ou espace d'hypothèses) tel que $\mathcal{F} := \cup_{m\geqslant 1}\mathcal{F}_m$ est soit $\mathcal{Y}^\mathcal{X}$ ou un ensemble très grand de fonctions, on pourra définir une penalité $\mathrm{pen}(m,n)$ et montrer une inégalité oracle

Définition 9. (Inégalité oracle) Soient $\mathcal{F}_m$ une séquence de modèles et $f^*_{m,P}$ la fonction cible dans $\mathcal{F}_m$ pour des données distribuées selon P. On appelle *inégalité oracle* une inégalité de la forme

$$\mathbb{E}_P\left[\mathcal{R}_P(\widehat{f})\right] - \mathcal{R}_P(f_P^\star) \leqslant C_n \inf_m \Big(\mathcal{R}_P(f^\star_{m,P}) - \mathcal{R}_P(f_P^\star) + \mathrm{pen}(m,n)\Big) + \varepsilon_n$$

Construire une suite de prédicteurs dans des modèles de plus en plus grand suivant la logique des inégalités oracle s'appelle la méthode de Grenander ou *method of sieves*.

3 Un théorème no free lunch en classification

Référence : Chapitre 7 de [DGL96].

Théorème 1. *On considère la perte* $0-1$ $\ell(f;(x,y)) = \mathbb{1}_{f(x)\neq y}$ *en classification binaire supervisée, et l'on suppose que* $\mathcal{X}$ *est infini. Alors, pour tout* $n\in\mathbb{N}$ *et toute règle de classification* $\widehat{f} : (\mathcal{X}\times\mathcal{Y})^n \mapsto \mathcal{F}$*,*

$$\sup_P \left\{ \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}\left(\widehat{f}\,(D_n)\right) - \mathcal{R}\left(f^\star\right)\right]\right\} \geqslant \frac{1}{2} > 0 \ , \tag{2}$$

le sup *étant pris sur l'ensembles des mesures de probabilité sur* $\mathcal{X}\times\mathcal{Y}$*. En particulier, aucun algorithme de classification ne peut être uniformément universellement consistante lorsque* $\mathcal{X}$ *est infini.*

Démonstration. Soit $n, K \in \mathbb{N}$, $\widehat{f} : (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathcal{F}$ un algorithme de classification. L'espace $\mathcal{X}$ étant infini, à bijection près, on peut supposer que $\mathbb{N} \subset \mathcal{X}$.

Pour tout $r \in \{0,1\}^K$, notons P_r la distribution de probabilité sur $\mathcal{X} \times \mathcal{Y}$ définie par $\mathbb{P}_{(X,Y)\sim P_r}(X = j \text{ et } Y = r_j) = K^{-1}$ pour tout $j \in \{1,\dots,K\}$. Autrement dit, X est choisi uniformément parmi $\{1,\dots,K\}$, et $Y = r_X$ est une fonction déterministe de X. En particulier, pour tout r, $\mathcal{R}_{P_r}(f^\star) = 0$.

Pour tout $r \in \{0,1\}^K$ (déterministe), on pose

$$F(r) = \mathbb{E}_{(X_i,Y_i)_{1\leqslant i\leqslant n}\sim P_r^{\otimes n}}\left[\mathcal{R}_{P_r}\left(\widehat{f}(D_n)\right)\right] .$$

La remarque clé est que pour toute distribution de probabilité R sur $\{0,1\}^K$,

$$\sup_{r\in\{0,1\}^K}\{F(r)\} \geqslant \mathbb{E}_{r\sim R}[F(r)] .$$

Notons R la distribution uniforme sur $\{0,1\}^K$, de telle sorte que $r \sim R$ signifie que $r_1,\dots,r_K$ sont indépendantes et de même distribution Bernoulli $\mathcal{B}(1/2)$. Alors,

$$\begin{aligned}
\mathbb{E}_{r\sim R}[F(r)] &= \mathbb{P}\left(\widehat{f}(X;D_n) \neq Y\right)\\
&= \mathbb{P}\left(\widehat{f}(X;D_n) \neq r_X\right)\\
&= \mathbb{E}\left[\mathbb{P}_{(r_j)_{j\notin\{X_1,\dots,X_n\}}}\left(\widehat{f}(X;D_n) \neq r_X \;\middle|\; X, X_1,\dots,X_n, r_{X_1},\dots,r_{X_n}\right)\right]\\
&\geqslant \mathbb{E}\left[\mathbb{E}_{(r_j)_{j\notin\{X_1,\dots,X_n\}}}\left(\mathbb{1}_{\widehat{f}(X;D_n)\neq r_X}\mathbb{1}_{X\notin\{X_1,\dots,X_n\}} \;\middle|\; X, X_1,\dots,X_n, r_{X_1},\dots,r_{X_n}\right)\right]\\
&= \mathbb{E}_{X,X_1,\dots,X_n,r_{X_1},\dots,r_{X_n}}\left[\frac{\mathbb{1}_{X\notin\{X_1,\dots,X_n\}}}{2}\right]\\
&= \frac{1}{2}\left(1-\frac{1}{K}\right)^n .
\end{aligned}$$

Pour tout $n \in \mathbb{N}$ fixé, cette borne inférieure tend vers $1/2$ lorsque K tend vers 1, d'où le résultat. □

On verra plus tard dans le cours qu'en revanche la consistance universelle uniforme est possible en classification 0–1 lorsque $\mathcal{X}$ est fini.

Un défaut du Théorème 1 est que la distribution P faisant échouer un algorithme de classification arbitraire $\widehat{f}$ change pour chaque taille d'échantillon. On pourrait donc imaginer qu'il est tout de même possible d'avoir une majoration de l'excès de risque de $\widehat{f}$ de la forme $c(P)u_n$ pour une suite $(u_n)_{n\geqslant 1}$ tendant vers 0 et une constante $c(P)$ fonction de la loi des observations. Le résultat suivant montre que ce n'est pas le cas, même avec une suite $(u_n)_{n\geqslant 1}$ tendant très lentement vers zéro.

Théorème 2 (Théorème 7.2 [DGL96], admis)**.** *On considère la perte* $0-1$ $\ell(f;(x,y)) = \mathbb{1}_{f(x)\neq y}$ *en classification binaire supervisée* ($\mathcal{Y} = \{0,1\}$)*, et l'on suppose que* $\mathcal{X}$ *est infini. Soit* $(a_n)_{n\geqslant 1}$ *une suite de réels positifs, décroissante, convergeant vers zéro, et telle que* $a_1 \leqslant 1/16$*. Alors, pour toute règle de classification* $\widehat{f} : \bigcup_{n\geqslant 1}(\mathcal{X}\times\mathcal{Y})^n \mapsto \mathcal{F}$*, il existe une distribution* P *sur* $\mathcal{X}\times\mathcal{Y}$ *telle que pour tout* $n \geqslant 1$*,*

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}(f^\star)\right] \geqslant a_n . \tag{3}$$

Références

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Verlag, 1996.