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

Simon Lacoste-Julien

2 octobre 2015

Résumé

On va voir la validation croisée pour faire la sélection de modèles. Un peu plus de théorie avec la notion de consistance *uniforme* et on définit la 'Sample complexity' d'un algorithme d'apprentissage. Finalement on voit un théorème 'no free lunch' qui dit en gros que sans faire des suppositions sur la loi qui génère les données, on ne peut apprendre de manière efficace (en particulier, le nombre de données n nécessaire pour avoir une performance raisonnable peut être arbitrairement grand).

1 Validation croisée

1.1 Sélection de l'algorithme d'apprentissage

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

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 $\widehat{f}_{D_n} := \mathscr{A}(D_n)$. $\widehat{f}(x; D_n)$ dénote $\widehat{f}_{D_n}$ évalué à x.

Excès de risque : $\mathcal{R}_P\left(\widehat{f}_m(D_n)\right) - \mathcal{R}_P(f^\star)$ $\quad - -$ $\quad \mathcal{R}_P(\cdot)$ rend la dépendance sur P explicite.

Risque : Le risque (au sens de Vapnik) donne l'*erreur de généralisation* de notre prédicteur – on veut le minimiser.

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

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

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

1.3 Validation croisée

Le problème avec la validation simple est que son estimation est trop variable car elle repose sur un choix arbitraire de découpage entre échantillons d'entraînement et de validation. Pour stabiliser l'estimateur, on peut faire une moyenne sur plusieurs découpages, ce que l'on appelle la *validation croisée*.

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

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

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

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

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)**.** (Équivalent à n-fold)

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

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 $I_1^v,\dots,I_k^v$ des sous-ensembles propres de $\{1,\dots,n\}$ de même cardinal n_v. Alors,*

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

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

Variance

— Pour la validation simple :

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

(Pour dériver cette équation, on utilise que $\mathrm{var}(X) = \mathbb{E}\,\mathrm{var}(X|Y) + \mathrm{var}\,\mathbb{E}[X|Y]$ avec $X = \widehat{\mathcal{R}}^{\mathrm{val}}(\dots)$ et $Y = D_n^e$.)

— 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) : $\mathrm{var}\left(\mathcal{R}\left(\widehat{f}_{D_n^e}\right)\right)$ diminue normalement avec n_e ; nombre k 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 k leur est lié également (e.g., k-fold).

1.5 Sélection d'hyperparamètre d'algorithme par validation croisée

— Définition :
$$\widehat{m} \in \arg\min_{m\in\mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}} \left(\widehat{f}_m; D_n; (I_j^v)_{1\leqslant j\leqslant k} \right) \right\}$$
— Pourquoi cela peut fonctionner :
Principe de l'estimation sans biais de l'espérance du risque (Proposition 1) ; analogue au principe de minimisation du risque empirique (cours 1).
— Choix d'une méthode de validation croisée : compromis entre temps de calcul et précision.
— Estimation du risque de l'estimateur final $\widehat{f}_{\widehat{m}}$: découpage en trois sous-ensembles (entraînement, validation et test).
— **Attention :** si $\mathcal{M}$ est trop grand, il y a encore danger de surapprentissage ! (Pourquoi ?)

2 Consistance uniforme *vs* universelle

Définition 6 (Consistance et consistance universelle)**.** 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. Pour la consistance *universellement uniforme*, l'algorithme d'apprentissage ne doit pas faire trop mal à chaque n pour *toutes* les distributions P.

2.1 Sample complexity

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 de* $\mathcal{P}$ pour l'algorithme $\widehat{f}$, le plus petit nombre $n(\mathcal{P}, \varepsilon, \widehat{f})$ tel que, pour tout $n \geqslant n(\mathcal{P}, \varepsilon, \widehat{f})$ on a
$$\sup_{P\in\mathcal{P}} \mathbb{E}_{D_n\sim P^{\otimes n}} \left[\mathcal{R}_P(\widehat{f}) \right] - \mathcal{R}_P(f_P^\star) < \varepsilon.$$

Entre d'autres termes, $n(\mathcal{P}, \varepsilon, \widehat{f})$ est la taille d'échantillon minimale nécessaire pour garantir un excès de risque en espérance inférieur à ε *pour n'importe quelle distribution* P *dans* $\mathcal{P}$.

La complexité en quantité de données *intrinsèque* de $\mathcal{P}$ est $n(\mathcal{P}, \varepsilon) := \inf_{\widehat{f}} n(\mathcal{P}, \varepsilon, \widehat{f})$, où l'infimum est pris sur tous les algorithmes d'apprentissage possibles.

Example : consistance universelle uniforme lorsque $\mathcal{X}$ est fini : $n(\mathcal{P}, \varepsilon, \widehat{f}) \leqslant \frac{|\mathcal{X}|}{\varepsilon^2}$ pour la règle de classification binaire qui prédit la classe la plus fréquente observée sur les données d'entraînement pour chaque x donné. (Voir théorème 2.1 dans les notes à http://www.di.ens.fr/~arlot/enseign/2009Centrale/ pour plus de détails).

Par contre, les théorèmes "No free lunch" – nous en verrons un aujourd'hui – 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
$$\forall P \in \mathcal{P}, \qquad \mathbb{E}_P \left[\mathcal{R}_P(\widehat{f}) \right] \leqslant \mathcal{R}_P(f_P^\star) + \varepsilon_n$$
où $\mathcal{P}$ sera l'ensemble des distributions possibles.

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 d'apprentissage de classification* $\widehat{f} : (\mathcal{X}\times\mathcal{Y})^n \mapsto \mathcal{F}$*,*

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

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

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

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

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

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,

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

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

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])**.** *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* $n \geqslant 1$*,*

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

1. On ne peut faire tendre K vers l'infini que si $\mathcal{X}$ est infini, d'où le besoin de cette condition. Pour $\mathcal{X}$ fini, il y a un déjeuner (trivial) gratuit (voir la section 2.1) !

Morale : La conclusion est donc que sans faire des suppositions sur la classe de distributions $\mathcal{P}$ qui pourrait générer les données, on ne peut obtenir des garanties sur l'erreur de généralisation de notre règle de classification pour un n fini donné. Nous allons voir dans le cours avec les bornes PAC-Bayes comment un a-priori sur $\mathcal{P}$ peut nous donner des garantis (et aussi motiver de la régularisation !).

Références

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