Apprentissage: cours 11 (1ère partie)
Validation croisée

Sylvain Arlot

10 mai 2012

1 Validation croisée

1.1 Sélection/calibration d'une règle d'apprentissage

- Rappels : Échantillon $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$. Règle d'apprentissage : $\widehat{f} : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathcal{F}$.
- Problème de sélection d'une règle d'apprentissage : $\left(\widehat{f}_m \right)_{m \in \mathcal{M}}$ famille de règles d'apprentissage.
  Objectif : minimiser le risque $\mathcal{R}\left(\widehat{f}_m(D_n) \right) - \mathcal{R}\left(f^\star \right)$. Inégalité-oracle

$$\mathcal{R}\left(\widehat{f}_{\widehat{m}(D_n)}(D_n) \right) - \mathcal{R}\left(f^\star \right) \leqslant C \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}\left(\widehat{f}_m(D_n) \right) - \mathcal{R}\left(f^\star \right) \right\} + R_n$$

  Enjeux : compromis entre sur-apprentissage et sous-apprentissage.
- Exemples : sélection de modèles, choix de k (ou d'une distance) pour les k plus proches voisins, choix d'un noyau ou d'une largeur de bande pour les estimateurs de Nadaraya-Watson, choix du paramètre de régularisation ou d'un noyau, etc.

1.2 Estimateurs par validation croisée du risque

Soit $\widehat{f}$ une règle d'apprentissage. 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).

- Estimateur par validation simple (définition). Si I^e est un sous-ensemble propre de $\{1, \dots, n\}$ (c'est-à-dire, non-vide et de complémentaire non-vide),

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

  Échantillon d'entraînement D_n^e. Échantillon de validation $D_n^v = (X_j, Y_j)_{j \in I_e^c}$.

– Estimateur par validation croisée (définition générale). Si pour $j \in \{1, \dots, B\}$, I_j^e est un sous-ensemble propre de $\{1, \dots, n\}$,

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

– Exemples : leave-one-out

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

et V-fold : si $(B_j)_{1 \leqslant j \leqslant V}$ est une partition de $\{1, \dots, n\}$,

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

1.3 Propriétés de l'estimateur par validation croisée du risque

Biais

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

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

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

Variance

– Pour la validation simple :

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

– Facteurs de variabilité : taille n_v de l'échantillon de validation (l'augmenter fait diminuer la variance, à n_e fixe du moins), "stabilité" de $\mathcal{A}$ (pour un échantillon 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., V-fold).

1.4 Sélection d'estimateurs 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.

– Résultats typiques en régression (asymptotique $n \to +\infty$) :
  1. Si $n_e \sim n$, alors, on a une inégalité-oracle avec une constante $C = C_n \to 1$ quand $n \to \infty$.
  2. Si $n_e \sim \kappa n$ avec $\kappa \in]0;1[$, alors, on a une inégalité-oracle avec une constante $C = C_n \to C(\kappa) > 1$ quand $n \to \infty$, et l'excès de risque de l'estimateur sélectionné est sous-optimal (perte d'un facteur multiplicatif $C'(\kappa) > 1$).
  3. Si $n_e \ll n$, alors, on n'a pas d'inégalité-oracle avec une constante $C = \mathcal{O}(1)$ quand $n \to \infty$ (perte d'un facteur multiplicatif tendant vers l'infini avec n).
– 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-échantillons (entraînement, validation et test).

Exercice 1 (À faire pour le 24 mai). Programmer la validation croisée V-fold pour résoudre le problème de sélection de modèle donné dans le cours du 12 avril.
Comparer empiriquement $V = 2$, $V = 5$ et $V = 10$ (risque, temps de calcul).
Comparer à la performance obtenue avec la pénalisation ℓ_0.
Bonus : même chose pour le choix de k avec la méthode des k-plus proches voisins (où l'on ne dispose plus de la pénalisation ℓ_0).