Apprentissage: cours 9 (2ème partie)
Pénalisation ℓ_0

Sylvain Arlot

12 avril 2012

Références : le premier cours de [Arl11] (pour une analyse théorique poussée de la sélection d'estimateurs des moindres carrés par pénalisation en régression) et les chapitres 12–13 de [DGL96] (pour la construction de pénalités via la dimension de Vapnik).

1 Correction de l'exercice 2 du partiel du 2 avril

- Cadre de la régression sur un plan d'expérience fixe, mesure de risque associée, excès de risque (question 3) :

$$Y = F + \varepsilon \in \mathbb{R}^n \qquad \mathcal{R}(t) - \mathcal{R}(F) = \frac{1}{n}\mathbb{E}\left[\|t - F\|^2\right] \qquad \widehat{\mathcal{R}}_n(t) = \frac{1}{n}\|t - Y\|^2$$

- Régressogramme sur une partition régulière à D morceaux (question 4) :

$$\widehat{F}_D = A_D Y \in \operatorname{argmin}_{t \in \mathcal{C}_D}\left\{\widehat{\mathcal{R}}_n(t)\right\}$$

- Décomposition du risque : erreur d'approximation/erreur d'estimation (question 5) :

$$\mathcal{R}\left(\widehat{F}_D\right) - \mathcal{R}(F) = \mathcal{R}\left(\widehat{F}_D\right) - \mathcal{R}(F_D) + \mathcal{R}(D_D) - \mathcal{R}(F)$$
$$\text{avec} \qquad F_D = A_D F \in \operatorname{argmin}_{t \in \mathcal{C}_D} \mathcal{R}(t)$$

- Étude de l'erreur d'approximation en fonction de D (question 6).
- Étude de l'erreur d'estimation en fonction de D (question 7) :

$$\mathcal{R}\left(\widehat{F}_D\right) - \mathcal{R}(F_D) = \frac{1}{n}\|A_D\varepsilon\|^2$$

- Quelle valeur du risque asymptotique si l'on choisit D au mieux (choix oracle $D^\star$, question 8) ?
- Que se passe-t-il si l'on choisit mal D (question 9) ?
- Quel est l'ordre de grandeur de $D^\star$ (question 10) ?

2 Pénalisation ℓ_0 pour la régression linéaire

2.1 Estimateur par projection

- Estimateur par projection (minimisation du risque empirique sur un sous-espace vectoriel) :
$$\widehat{F}_S = \Pi_S Y \in \operatorname{argmin}_{t\in S} \widehat{\mathcal{R}}_n(t)$$
où S est un sous-espace vectoriel de R^n, et Π_S est la matrice de projection orthogonale sur S.
- Exemples :
  - régressogrammes (réguliers ou pas),
  - base de $\mathbb{R}^n$ tronquée (Fourier, ondelettes, etc.) : si Y_i correspond à l'observation d'un signal en t_i $(i = 1, \dots, n)$, et si $(\phi_j)_{j\in\mathbb{N}}$ est une famille de vecteurs de $L^2([0,1])$, cela revient à prendre le sous-espace $S = S_D$ de $\mathbb{R}^n$ généré par $\{(\phi_j(t_i))_{1\leqslant i\leqslant n}\,,\, 1 \leqslant j \leqslant D\}$,
  - régression linéaire (on cherche $\widehat{F}$ sous la forme $F_w = w^\top X$ avec X une matrice $n \times p$ donnée et $w \in \mathbb{R}^p$, donc $S = \left\{ w^\top X\,,\, w \in \mathbb{R}^p \right\}$ et $\Pi_S = \Pi(X) = X(X^\top X)^{-1}X^\top$ si $X^\top X$ est inversible, voir le premier cours).
- Remarque : un régressogramme sur une partition à D morceaux correspond à la régression linéaire avec $p = D$ et $X_{i,j} = \mathbb{1}_{t_i \in \text{ morceau } j \text{ de la partition}}$.
Question : et dans le cas d'une base de $\mathbb{R}^n$ tronquée ?
- Calcul de l'excès de risque d'un estimateur par projection (et de son espérance).

2.2 Choix d'un estimateur par projection

- Problème : $(S_m)_{m\in\mathcal{M}}$ famille de modèles disponibles, $\widehat{F}_m = \widehat{F}_{S_m}$ les estimateurs par projection associés, lequel choisir ?
- Exemples :
  - régressogrammes réguliers : choix de D.
  - régression linéaire : pour chaque sous-ensemble $m \subset \{1, \dots, p\}$ des variables composant X, qu'on note X_m la sous-matrice correspondante, et on a l'estimateur
$$\widehat{F}_m = \Pi(X_m)Y = \Pi_m Y \ .$$
Question : que vaut la dimension du sous-espace S_m correspondant ?
- Choix oracle : $m^\star \in \operatorname{argmin}_{m\in\mathcal{M}} \left\{ \mathcal{R}\left(\widehat{F}_m\right) \right\}$, mais il dépend de quantités inconnues. On cherche un choix $\widehat{m} = \widehat{m}(Y) \in \mathcal{M}$ dépendant des données uniquement.
Minimiser $\mathbb{E}\left[\mathcal{R}\left(\widehat{F}_m\right)\right]$: compromis biais-variance (sur-apprentissage/sous-apprentissage).
- Défaut du risque empirique $\widehat{\mathcal{R}}_n\left(\widehat{F}_m\right)$ comme critère de choix de m.
- Pénalisation :
$$\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} \left\{ \widehat{\mathcal{R}}_n\left(\widehat{F}_m\right) + \operatorname{pen}(m) \right\}$$
pour une fonction pen : $\mathcal{M} \mapsto \mathbb{R}$ (pénalité). Rôle de la pénalité : éviter le sur-apprentissage.

– Pénalité idéale : l'écart entre risque et risque empirique

$$\mathrm{pen}_{\mathrm{id}}(m) := \mathcal{R}\left(\widehat{F}_m\right) - \widehat{\mathcal{R}}_n\left(\widehat{F}_m\right) \quad .$$

– Pénalisation C_p de Mallows :

$$\mathrm{pen}(m) = \frac{2\sigma^2 \dim(S_m)}{n} \quad .$$

Proposition 1 (Calcul de la pénalité C_p de Mallows)**.** *Si $\varepsilon_1, \ldots, \varepsilon_n$ sont iid de moyenne nulle et de variance σ^2, alors*

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 \dim(S_m)}{n} - \sigma^2 \quad . \tag{1}$$

Exercice 1 (À faire à la maison, pour le 3 mai)*.* On considère la collection des modèles $(\mathcal{C}_D)_{1 \leqslant D \leqslant n}$ introduite dans le partiel, sans supposer ici que D divise nécessairement n. Les estimateurs correspondants sont des régressogrammes réguliers. Soit $F = (\, f(i/n)\,)_{1 \leqslant i \leqslant n}$ avec $\forall t \in [0,1]$, $f(t) = 4\sin(4\pi t) + 3\cos(6\pi t) - 2\sin(6\pi t)$. On observe

$$Y_i = F_i + \varepsilon_i \qquad \text{pour} \quad i \in \{\, 1, \ldots, n \,\}$$

avec ε_i i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

1. Programmer la génération d'un tel échantillon Y avec $\sigma = 3$ et $n = 200$.
2. Calculer les estimateurs "régressogrammes réguliers" $\widehat{F}_D$ pour $D = 1, \ldots, n$, et tracer en fonction de D : leur risque quadratique, leur risque empirique, leur risque empirique pénalisé par la pénalité C_p de Mallows.
3. Tracer l'estimateur $\widehat{F}_{\widehat{D}}$ de F obtenu par la pénalisation C_p (on fera comme si σ^2 était connu).
4. Calculer l'estimateur $\widehat{F}_{\widehat{D}(K)}$ sélectionné par la pénalité

   $$\mathrm{pen}(m, K) = \frac{K\sigma^2 D_m}{n}$$

   avec une constante $K > 0$ variable. En prenant une grille de valeurs de K (suffisamment fine dans l'intervalle $[1/2, 3]$), évaluer l'espérance du risque quadratique de $\widehat{F}_{\widehat{D}(K)}$, et le tracer en fonction de K. Commenter la courbe obtenue (lieu du minimum, comportement lorsque K est petit ou très grand, etc.).
5. Bonus : même chose en considérant les modèles S_D générés par les vecteurs $\{\, (\cos(k\pi i/n), \sin(k\pi i/n))_{1 \leqslant i \leqslant n}\,,\ 1 \leqslant k \leqslant D \,\}$. Comparer les résultats.

Question bonus : appliquer la pénalité C_p au problème de sélection de variables sur un jeu de données réelles parmi ceux qui sont disponibles sur `http://archive.ics.uci.edu/ml/datasets.html` (chercher un jeu de données pour la régression, avec des "attribute" (variables explicatives) numériques, pas trop nombreuses pour pouvoir

explorer les 2^p parties de l'ensemble des p variables explicatives ; par exemple, le jeu de données "wine quality" possède ces caractéristiques, voir http://archive.ics.uci.edu/ml/datasets/Wine+Quality). Pour l'estimation de σ^2, on pourra par exemple utiliser les résidus le modèle intégrant toutes les variables, ce qui est possible lorsque $p < n$.

3 Pénalisation dans un cadre général d'apprentissage

Note : Cette section sera traitée lors du cours du 3 mai.

3.1 Sélection de prédicteur

- Problème : $\left(\widehat{f}_m(D_n)\right)_{m\in\mathcal{M}}$ famille de prédicteurs, lequel choisir ?
- Objectif : minimiser le risque. 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.

3.2 Pénalisation

- Définition :

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}} \left\{\widehat{\mathcal{R}}_n\left(\widehat{f}_m(D_n)\right) + \mathrm{pen}(m)\right\}$$

  pour une fonction pen : $\mathcal{M} \mapsto \mathbb{R}$ (pénalité)
- Rôle de la pénalité : éviter le sur-apprentissage.
- Pénalité idéale :

$$\mathrm{pen}_{\mathrm{id}}(m) := \mathcal{R}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right)$$

- Heuristique :

$$\mathrm{pen}(m) \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right)\right]$$

  fournit une bonne procédure de sélection d'estimateur, si $\mathcal{M}$ n'est pas trop grand.

3.3 Construction de pénalités

- Début de justification (Vapnik) : si $\widehat{f}_m$ est un minimiseur du risque empirique sur S_m, alors

$$\mathrm{pen}_{\mathrm{id}}(m) \leqslant \sup_{f\in S_m} \left\{\mathcal{R}\left(f\right) - \widehat{\mathcal{R}}_n\left(f\right)\right\}$$

  quantité que nous avons déjà vue pour majorer l'erreur d'estimation (voir cours du 22 mars).

– Si S_m est fini et la fonction de perte ℓ à valeurs dans $[0,1]$, on peut appliquer le Théorème 1 du cours du 22 mars .
– Exemple : histogrammes en classification, sur une partition de taille $D_m < +\infty$: S_m est fini de cardinal 2^{D_m}, d'où la pénalité $\propto \sqrt{D_m/n}$. Cette pénalité est très pessimiste en pratique.
– On peut également faire mieux dans le cas zéro-erreur : voir l'Exercice 3.

Exercice 2 (Bonus). Tester une pénalité proportionnelle à $\sqrt{D_m/n}$ pour choisir le pas d'une partition régulière en classification dans l'exercice de programmation donné le 23 février.

Exercice 3 (Bonus). Si $\mathcal{R}(f^\star) = 0$ et $f^\star \in S$ fini, alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\mathcal{R}\left(\widehat{f}_S\right) \geqslant \varepsilon\right) \leqslant \operatorname{card}(S) \exp(-n\varepsilon) \tag{2}$$

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_S\right)\right] \leqslant \frac{1 + \ln(\operatorname{card}(S))}{n} \quad . \tag{3}$$

Références

[Arl11] Sylvain Arlot. Sélection de modèles et sélection d'estimateurs pour l'apprentissage statistique, January 2011. Cours Peccot. Collège de France. http://www.di.ens.fr/~arlot/peccot.htm.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.