Sélection de modèles et sélection d'estimateurs pour l'Apprentissage statistique

Sylvain Arlot (Cnrs)

Cours Peccot 2011

Lieu

Salle de conférence, Collège de France, 3 rue d'Ulm (Paris 5e)

Dates

Lundi 10 Janvier 2011, 14h-16h
Lundi 17 Janvier 2011, 14h-16h
Lundi 24 Janvier 2011, 14h-16h
Lundi 31 Janvier 2011, 15h30-17h30

Résumé

Le problème de la prédiction est l'un des principaux problèmes de l'apprentissage statistique : étant donné un échantillon $(X_i, Y_i)_{i=1..n}$ de variables aléatoires indépendantes et identiquement distribuées, le but est de prédire une nouvelle réalisation Y_{n+1} de la variable d'intérêt à l'aide de X_{n+1} (les variables explicatives) uniquement.

De nombreux estimateurs (ou règles d'apprentissage) ont été proposés pour ce problème, chacun dépendant en général d'un ou plusieurs paramètres, dont la calibration précise est cruciale pour obtenir une performance optimale.

Cette série de quatre cours abordera le problème de la sélection d'estimateurs à l'aide des données uniquement, principalement dans le cadre de la prédiction. Ceci comprend en particulier le problème de la sélection de modèles (où chaque estimateur est obtenu par minimisation du risque empirique sur un modèle), et celui de la calibration de paramètres (par exemple, le paramètre de régularisation pour les méthodes de minimisation du risque empirique régularisé).

Nous considérerons deux approches principales : la pénalisation du risque empirique (avec une forme de pénalité déterministe ou estimée à l'aide des données), et la validation croisée.

Nous proposerons des réponses aux questions suivantes :
- Quels résultats mathématiques peuvent être prouvés pour les procédures de sélection existantes ?
- Comment ces résultats peuvent-ils aider à choisir en pratique une procédure de sélection pour un problème statistique donné ?
- Comment la théorie peut-elle guider la définition de nouvelles procédures de sélection ?

Plan prévisionnel

Lundi 10 Janvier, 14h00-16h00 :
Cadre de l'apprentissage statistique — Estimateurs classiques — Enjeux du problème de la sélection d'estimateurs et approches principales — Résultats mathématiques en jeu

Lundi 17 Janvier, 14h00-16h00 :
Pénalités minimales et calibration de pénalités (pour la régression par moindres carrés homoscédastique [7] ou hétéroscédastique [6, 10], pour les estimateurs linéaires en régression [3] ou pour l'estimation de densité [8, 9]).

Lundi 24 Janvier, 14h00-16h00 :
Pénalités par rééchantillonnage (pour la régression hétéroscédastique [2] ou l'estimation de densité [8]).

Lundi 31 Janvier, 15h30-17h30 :
Validation croisée [5] et pénalités reliées [1] — Application à la détection de ruptures [4].

Références

[1] Sylvain Arlot. V-fold cross-validation improved : V-fold penalization, February 2008. arXiv :0802.0566v2.

[2] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3 :557–624 (electronic), 2009.

[3] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

[4] Sylvain Arlot and Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, pages 1–20, 2010. 10.1007/s11222-010-9196-x.

[5] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4 :40–79, 2010.

[6] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10 :245–279 (electronic), 2009.

[7] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[8] Matthieu Lerasle. Optimal model selection in density estimation. hal-00422655, 2009.

[9] Adrien Saumard. Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models. hal-00512310, September 2010.

[10] Adrien Saumard. The slope heuristics in heteroscedastic regression. hal-00512306, September 2010.