Théorie PAC-Bayésienne

Benjamin Guedj
http://www.lsta.upmc.fr/doct/guedj/index.html

LSTA, UPMC & LTCI, Telecom ParisTech

14 février 2013

Les grands principes

- Apprentissage statistique (machine learning).

Les grands principes

- Apprentissage statistique (machine learning).
- Ensemble de techniques issues de l'approche PAC et de la communauté bayésienne.

Les grands principes

- Apprentissage statistique (machine learning).
- Ensemble de techniques issues de l'approche PAC et de la communauté bayésienne.
- PAC : probably approximately correct. Liens avec l'approche oracle : choisir un prédicteur parmi un ensemble d'alternatives tel que, avec grande probabilité, les prédictions soient aussi précises que possible.

Les grands principes

- Apprentissage statistique (machine learning).
- Ensemble de techniques issues de l'approche PAC et de la communauté bayésienne.
- PAC : probably approximately correct. Liens avec l'approche oracle : choisir un prédicteur parmi un ensemble d'alternatives tel que, avec grande probabilité, les prédictions soient aussi précises que possible.
- Analyse bayésienne : une distribution a priori sur ces alternatives permet d'attribuer à des portions de l'espace des paramètres une masse d'autant plus grande qu'elle est consistante avec l'échantillon d'apprentissage.

Références

Approche PAC

- McAllester (1999); Shawe-Taylor and Williamson (1997)

Références

Approche PAC

- McAllester (1999); Shawe-Taylor and Williamson (1997)

Formalisation PAC-Bayésienne, classification

- Catoni (2004, 2007)

Références

Approche PAC

- McAllester (1999); Shawe-Taylor and Williamson (1997)

Formalisation PAC-Bayésienne, classification

- Catoni (2004, 2007)

Régression

- Alquier (2006, 2008); Audibert (2004a,b); Audibert and Catoni (2010, 2011)

Références

Approche PAC

- McAllester (1999); Shawe-Taylor and Williamson (1997)

Formalisation PAC-Bayésienne, classification

- Catoni (2004, 2007)

Régression

- Alquier (2006, 2008); Audibert (2004a,b); Audibert and Catoni (2010, 2011)

Régression sparse

- Alquier and Lounici (2011); Dalalyan and Salmon (2012); Dalalyan and Tsybakov (2008, 2012); Rigollet and Tsybakov (2012)

Références

Approche PAC

- McAllester (1999); Shawe-Taylor and Williamson (1997)

Formalisation PAC-Bayésienne, classification

- Catoni (2004, 2007)

Régression

- Alquier (2006, 2008); Audibert (2004a,b); Audibert and Catoni (2010, 2011)

Régression sparse

- Alquier and Lounici (2011); Dalalyan and Salmon (2012); Dalalyan and Tsybakov (2008, 2012); Rigollet and Tsybakov (2012)

Extension à des modèles particuliers

- Alquier and Biau (2013); Alquier, Li, and Wintenberger (2012); Guedj and Alquier (2013); Suzuki (2012)

Construction des estimateurs PAC-Bayésiens I

Notons $(\Theta, \mathcal{T})$ un espace de paramètres, et $\mathcal{D}_n = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ un n-échantillon.

Soit π une distribution *a priori* sur $(\Theta, \mathcal{T})$ et $\lambda > 0$. On note R le risque associé à une certaine distance δ, et R_n le risque empirique correspondant basé sur $\mathcal{D}_n$.

La distribution *a posteriori* de Gibbs ρ_λ est alors définie sur $(\Theta, \mathcal{T})$ comme la mesure ayant pour densité

$$\frac{\mathrm{d}\rho_\lambda}{\mathrm{d}\pi}(\theta) \propto \exp(-\lambda R_n(\theta)).$$

On peut alors définir l'estimateur randomisé PAC-Bayésien

$$\hat{\theta} \sim \rho_\lambda$$

et l'agrégé

$$\bar{\theta} = \int_\Theta \theta \rho_\lambda(\mathrm{d}\theta).$$

Construction des estimateurs PAC-Bayésiens II

Notons $\mathcal{M}_\pi^1(\Theta)$ l'ensemble des mesures de probabilités sur $(\Theta, \mathcal{T})$ absolument continues par rapport à π, et $\mathcal{KL}(\cdot, \cdot)$ la divergence de Kullback-Leibler.

La distribution de Gibbs est l'unique solution du problème d'optimisation convexe :

$$\underset{\rho \in \mathcal{M}_\pi^1(\Theta)}{\operatorname{argmin}} \left\{ \int_\Theta R_n(\theta)\rho(\mathrm{d}\theta) + \frac{\lambda}{n}\mathcal{KL}(\rho, \pi) \right\}.$$

Régression PAC-Bayésienne

Modèle : $Y = \theta^\star X + W$. Deux hypothèses :

1. $\forall k \in \mathbb{N}$, $\mathbb{E}[|W|^k] < \infty$, $\mathbb{E}[W|X] = 0$ et il existe deux constantes positives L et σ^2 telles que $\forall k \geq 2$, $\mathbb{E}[|W|^k|X] \leq \frac{k!}{2}\sigma^2 L^{k-2}$.
2. $\exists\ C > \max(1, \sigma)$ t.q. $\|\theta^\star\| \leq C$.

Les estimateurs PAC-Bayésiens sont contrôlés par des bornes PAC comme celle-ci.

Théorème (Alquier (2006); Alquier and Lounici (2011); Catoni (2004); Guedj and Alquier (2013))

Sous les hypothèses (1) et (2), avec probabilité au moins $1 - 2\varepsilon$,

$$\left.\begin{array}{l} R(\hat{\theta}) - R(\theta^\star) \\ R(\bar{\theta}) - R(\theta^\star) \end{array}\right\} \leq \text{cste} \times \inf_{\rho \in \mathcal{M}^1_\pi(\Theta)} \left\{ \int R(\theta)\rho(\mathrm{d}\theta) - R(\theta^\star) + \frac{\mathcal{KL}(\rho, \pi) + \log \frac{1}{\varepsilon}}{n} \right\},$$

pour tout $\varepsilon \in (0, 1)$.

Preuve du théorème précédent I

Lemme (Massart (2007))

Soit $(T_i)_{i=1}^n$ une collection de variables réelles indépendantes. Supposons qu'il existe deux constantes positives v et w t.q. pour tout entier $k \geq 2$,

$$\sum_{i=1}^{n} \mathbb{E}[(T_i)_+^k] \leq \frac{k!}{2} v w^{k-2}.$$

Alors pour tout $\gamma \in (0, \frac{1}{w})$,

$$\mathbb{E}\left[\exp\left(\gamma \sum_{i=1}^{n}(T_i - \mathbb{E}\,T_i)\right)\right] \leq \exp\left(\frac{v\gamma^2}{2(1-w\gamma)}\right).$$

Preuve du théorème précédent II

Lemme (Catoni (2004))

Soit $(A, \mathcal{A})$ *un espace mesurable. Pour toute mesure de probabilité* μ *sur* $(A, \mathcal{A})$ *et toute fonction mesurable* $h : A \to \mathbb{R}$ *t.q.* $\int (\exp \circ h) \mathrm{d}\mu < \infty$,

$$\log \int (\exp \circ h) \mathrm{d}\mu = \sup_{m \in \mathcal{M}^1_\pi(A, \mathcal{A})} \int h \mathrm{d}m - \mathcal{KL}(m, \mu),$$

avec la convention $\infty - \infty = -\infty$. *De plus, si* h *est majorée sur le support de* μ, *le supremum en* m *dans le terme de droite est atteint pour la distribution de Gibbs* g *définie par*

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp(h(a))}{\int (\exp \circ h) \mathrm{d}\mu}, \quad a \in A.$$

Preuve du théorème précédent III

Soit $\rho \in \mathcal{M}_{\pi}^{1}(\Theta)$ et $\theta \sim \rho$. Sous les hypothèses (1) et (2), soit $w = 8C\max(L, C)$, $\lambda \in (0, n/[w + 4(\sigma^2 + C^2)])$ et $\varepsilon \in (0, 1)$.

Lemme (1)

Avec probabilité au moins $1 - \varepsilon$

$$R(\theta) - R(\theta^\star) \leq \frac{1}{1 - \frac{4\lambda(\sigma^2 + C^2)}{n - w\lambda}} \left(R_n(\theta) - R_n(\theta^\star) + \frac{\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) + \log \frac{1}{\varepsilon}}{\lambda} \right).$$

Lemme (2)

Avec probabilité au moins $1 - \varepsilon$

$$\int R_n(\theta)\rho(\mathrm{d}\theta) - R_n(\theta^\star) \leq \left[1 + \frac{4\lambda(\sigma^2 + C^2)}{n - w\lambda} \right] \left[\int R(\theta)\rho(\mathrm{d}\theta) - R(\theta^\star) \right] + \frac{\mathcal{KL}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda}.$$

Preuve du théorème précédent IV

Lemme (3)

Avec probabilité au moins $1-\varepsilon$

$$\int R(\theta)\rho(\mathrm{d}\theta) - R(\theta^\star) \leq \frac{1}{1-\frac{4\lambda(\sigma^2+C^2)}{n-w\lambda}} \left(\int R_n(\theta)\rho(\mathrm{d}\theta) - R_n(\theta^\star) + \frac{\mathcal{KL}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda} \right).$$

Extension : le modèle additif sparse

Modèle : $Y = \sum_{j=1}^{p} f_j^\star(X_j) + W$.

Théorème (Guedj and Alquier (2013))

Sous les hypothèses (1) et (2), soit $w = 8C\max(L, C)$ *et* $\lambda = n\ell/[w + 4(\sigma^2 + C^2)]$, *pour* $\ell \in (0,1)$, *et* $\varepsilon \in (0,1)$. *Alors avec probabilité au moins* $1 - 2\varepsilon$,

$$\left.\begin{array}{l} R(\hat{\theta}) - R(\theta^\star) \\ R(\bar{\theta}) - R(\theta^\star) \end{array}\right\} \leq \text{cste} \times \inf_{m \in \mathcal{M}_\pi^1(\Theta)} \inf_{\theta \in \mathcal{B}_m^1(0,C)} \left\{ R(\theta) - R(\theta^\star) \right.$$

$$\left. + |S(m)| \frac{\log(p/|S(m)|)}{n} + \frac{\log(n)}{n} \sum_{j \in S(m)} m_j + \frac{\log(1/\varepsilon)}{n} \right\}.$$

Implémentation : package R `pacbpred`.
`http://cran.r-project.org/web/packages/pacbpred/index.html`

Références I

Pierre Alquier. *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. PhD thesis, Université Paris 6 - UPMC, December 2006.

Pierre Alquier. PAC-Bayesian Bounds for Randomized Empirical Risk Minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

Pierre Alquier and Gérard Biau. Sparse Single-Index Model. *Journal of Machine Learning Research*, 14:243–280, 2013.

Pierre Alquier and Karim Lounici. PAC-Bayesian Theorems for Sparse Regression Estimation with Exponential Weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

Pierre Alquier, Xiaoyin Li, and Olivier Wintenberger. Prediction of time series by statistical learning: General losses and fast rates. Submitted, 2012.

Jean-Yves Audibert. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré : Probabilités et Statistiques*, 40(6):685–736, 2004a.

Références II

Jean-Yves Audibert. *Théorie statistique de l'apprentissage : une approche PAC-Bayésienne*. PhD thesis, Université Paris 6 - UPMC, 2004b.

Jean-Yves Audibert and Olivier Catoni. Robust linear regression through PAC-Bayesian truncation. Submitted, 2010.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization*. École d'Été de Probabilités de Saint-Flour XXXI – 2001. Springer, 2004.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture notes – Monograph Series*. Institute of Mathematical Statistics, 2007.

Arnak S. Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4): 2327–2355, 2012.

Références III

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

Benjamin Guedj and Pierre Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7: 264–291, 2013.

Pascal Massart. *Concentration Inequalities and Model Selection*. École d'Été de Probabilités de Saint-Flour XXXIII – 2003. Springer, 2007.

David A. McAllester. Some PAC-Bayesian Theorems. *Machine Learning*, 37:355–363, 1999.

Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

Références IV

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9, 1997.

Taiji Suzuki. PAC-Bayesian Bound for Gaussian Process Regression and Multiple Kernel Additive Model. In *Proceedings of the 25th annual conference on Computational Learning Theory*, 2012.