PAC-Bayesian Theory Meets Bayesian Inference

Pascal Germain[†], Francis Bach[†], Simon Lacoste-Julien[†], Alexandre Lacoste[‡]

[†] INRIA Paris - École Normale Supérieure
[‡] Google

NIPS 2016 spotlight

Dans la vie, l'essentiel est de porter sur tout des jugements a priori.

— Boris Vian

PAC-Bayesian Theory

The PAC-Bayesian theory claims to provide "PAC guarantees to Bayesian algorithms" (McAllester, 1999).

Assumption

The training set (X, Y) contains n **i.i.d. samples** from a **data distribution** $\mathcal{D}$.

Probably Approximately Correct (PAC) bound

With probability at least "$1-\delta$", the loss of predictor f is less than "ε",

$$\Pr_{X,Y\sim\mathcal{D}^n}\left(\mathcal{L}_{\mathcal{D}}(f) \leq \varepsilon(\widehat{\mathcal{L}}_{X,Y}(f), n, \delta, \ldots)\right) \geq 1-\delta\,.$$

Bayesian Flavor

Given a prior π and a posterior $\hat{\rho}$ over a class of predictors $\mathcal{F}$,

$$\Pr_{X,Y\sim\mathcal{D}^n}\left(\mathop{\mathbf{E}}_{f\sim\hat{\rho}}\mathcal{L}_{\mathcal{D}}(f) \leq \varepsilon(\mathop{\mathbf{E}}_{f\sim\hat{\rho}}\widehat{\mathcal{L}}_{X,Y}(f), n, \delta, \mathrm{KL}(\pi\|\hat{\rho}), \ldots)\right) \geq 1-\delta\,.$$

A PAC-Bayesian Theorem for [0,1]-losses*

* In the paper, we consider unbounded losses

Given a loss function $\ell(f,x,y) \in [0,1]$, $\quad \mathcal{L}_{\mathcal{D}}(f) \;:=\; \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}} \ell(f,x,y)\,.$

Theorem (adapted from Catoni, 2007)

With probability at least "$1-\delta$",

$$\forall\, \hat\rho \text{ on } \mathcal{F}: \quad \mathop{\mathbf{E}}_{f\sim\hat\rho} \mathcal{L}_{\mathcal{D}}(f) \leq \frac{1}{1-e^{-1}} \left(\mathop{\mathbf{E}}_{f\sim\hat\rho} \widehat{\mathcal{L}}_{X,Y}(f) + \frac{1}{n}\left[\mathrm{KL}(\hat\rho\|\pi) + \ln\frac{1}{\delta}\right] \right),$$

The bound suggests to minimize the following trade-off :

$$n \mathop{\mathbf{E}}_{f\sim\hat\rho} \widehat{\mathcal{L}}_{X,Y}(f) + \mathrm{KL}(\hat\rho\|\pi)\,.$$

Optimal posterior

$$\hat\rho^*(f) \;=\; \frac{1}{Z_{X,Y}}\,\pi(f)\, e^{-n\,\widehat{\mathcal{L}}_{X,Y}(f)}\,.$$

PAC-Bayesian Theory Meets Bayesian Inference

Negative log-likelihood loss function

Given a Bayesian likelihood $p(Y|X,\theta)$, let $\ell_{\rm nll}(\theta,x,y) = \ln \frac{1}{p(y|x,\theta)}$.

The PAC-Bayesian and Bayesian posteriors align :

$$\underbrace{\hat{\rho}^*(\theta) = \frac{\pi(\theta)\, e^{-n\,\widehat{\mathcal{L}}_{X,Y}^{\ell_{\rm nll}}(\theta)}}{Z_{X,Y}}}_{\text{PAC-Bayesian posterior}} = \underbrace{\frac{p(\theta)\, p(X,Y|\theta)}{p(Y|X)} = p(\theta|X,Y)}_{\text{Bayesian posterior}}\,.$$

The normalization constant $Z_{X,Y}$ corresponds to the Bayesian *marginal likelihood*

$$Z_{X,Y} = p(Y|X) = \int_\Theta \pi(\theta)\, e^{-n\,\widehat{\mathcal{L}}_{X,Y}^{\ell_{\rm nll}}(\theta)}\, d\theta\,.$$

Moreover,

$$-\ln Z_{X,Y} = n\, \mathbf{E}_{\theta\sim\hat{\rho}^*}\, \widehat{\mathcal{L}}_{X,Y}^{\ell_{\rm nll}}(\theta) + \mathrm{KL}(\hat{\rho}^*\|\pi)\,.$$

Farewell

Take home message !

The Bayesian marginal likelihood minimizes (some) PAC-Bayesian Bounds.

Our paper also contains :

- PAC-Bayesian theorems for unbounded (sub-gamma) loss functions,
- Study of Bayesian model selection techniques (model evidence),
- Bayesian linear regression experiments.