Distribution-Dependent PAC-Bayes Priors

Guy Lever [1] François Laviolette[2] John Shawe-Taylor[1]

[1] Uiversity College London
Centre for Computational Statistics and Machine Learning

[2] Université Laval
Département d'informatique

22 March, 2010

Overview

- PAC-Bayes prior informed by data-generating distribution (Catoni's "localization")
- Investigate localization in a variety of methodologies:
- Gibbs-Boltzmann (original setting)
  - Sharp risk analysis
  - Investigate (controlling) function class complexity
  - Encode assumptions about interaction between classifiers and data geometry
- Gaussian Processes (new setting)
  - Practical
  - Sharp risk analysis
- Significant reduction in KL divergence

Preliminaries- Typical PAC-Bayes Analysis

- Distribution D over $\mathcal{X} \times \mathcal{Y}$
- Sample $\mathcal{S} \sim D^m$
- Class $\mathcal{H}$ of hypotheses $h : \mathcal{X} \to \mathcal{Y}$
- prior P, posterior Q over $\mathcal{H}$
- Recall PAC-Bayes bound

Theorem (Seeger's bound)

For any D, any set $\mathcal{H}$ of classifiers, any distribution P on $\mathcal{H}$, for all Q on $\mathcal{H}$ and any $\delta \in (0, 1]$, with probability at least $1 - \delta$

$$\mathrm{kl}(\widehat{\mathrm{risk}}_{\mathcal{S}}(G_Q), \mathrm{risk}(G_Q)) \leq \frac{1}{m}\left(\mathrm{KL}(Q||P) + \ln \frac{\xi(m)}{\delta}\right)$$

$$\text{where } \xi(m) = \mathcal{O}(\sqrt{m})$$

- Dominant quantity is KL divergence - can be large...

Localization - Motivation

Typically...

- P not informed by data-generating distribution
  - Prior weight assigned to high risk classifiers
  - If Q "good" then $D(Q||P)$ large
- Choice of Q constrained by need to minimize divergence

Localization...

- Key observation: P can be informed by D
- e.g. high prior mass only to classifiers with low true risk

$$p(h) = \frac{1}{Z'} e^{-\gamma \mathrm{risk}(h)}$$

- P unknown
- Choose Q such that $\mathrm{KL}(Q||P)$ estimated

Localization 2 - Our interpretation

- We consider exponential families

$$p(h) := \frac{1}{Z'} e^{-F_p(h)} \qquad q(h) := \frac{1}{Z} e^{-\widehat{F}_q(h)}$$

- To obtain risk analysis we just need to bound $\mathrm{KL}(Q||P)$

Lemma

$$\mathrm{KL}(Q||P) \quad \leq \quad (\mathbb{E}_{h\sim Q} - \mathbb{E}_{h\sim P})[F_p(h) - \widehat{F}_q(h)]$$

- Choose $\widehat{F}_q$ to estimate F_p from the sample $\mathcal{S}$
- $\mathrm{KL}(Q||P) \leq \sup_{h\in\mathcal{H}} |F_p(h) - \widehat{F}_q(h)|$
- Lemma is "recursive"
- Establish convergence: KL decays with the sample

Stochastic ERM 1 - Risk Bound

- P and Q are Gibbs-Boltzmann distributions

$$p(h) := \frac{1}{Z'} e^{-\gamma \mathrm{risk}(h)} \qquad q(h) := \frac{1}{Z} e^{-\gamma \widehat{\mathrm{risk}}_{\mathcal{S}}(h)}$$

- We must bound $(\mathbb{E}_{h\sim Q} - \mathbb{E}_{h\sim P})[\gamma \mathrm{risk}(h) - \gamma \widehat{\mathrm{risk}}_{\mathcal{S}}(h)]$

Lemma

With probability at least $1-\delta$,

$$\mathrm{KL}(Q\|P) \;\leq\; \frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{2\xi(m)}{\delta}} + \frac{\gamma^2}{4m}\,.$$

Theorem (Risk Bound for stochastic ERM)

With probability at least $1-\delta$,

$$\mathrm{kl}(\widehat{\mathrm{risk}}_{\mathcal{S}}(G_Q), \mathrm{risk}(G_Q)) \leq \frac{1}{m}\left(\frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{4\xi(m)}{\delta}} + \frac{\gamma^2}{4m} + \ln\frac{2\xi(m)}{\delta}\right)$$

Stochastic ERM 2 - Complexity

- Where is the dependence on function class complexity?
- Captured by γ: "inverse temperature" controls variance

$$p(h) := \frac{1}{Z'} e^{-\gamma \mathrm{risk}(h)} \qquad q(h) := \frac{1}{Z} e^{-\gamma \widehat{\mathrm{risk}}_{\mathcal{S}}(h)}$$

- If $\mathcal{H}$ is rich γ must be large to control $\mathbb{E}_{h \sim Q}[\widehat{\mathrm{risk}}_{\mathcal{S}}(h)]$
- New notion of complexity?

Regularized Stochastic ERM

- Add a regularization terms to control capacity

$$p(h) := \frac{1}{Z'} e^{-\gamma \mathrm{risk}(h) + \eta F_p(h)} \qquad q(h) := \frac{1}{Z} e^{-\gamma \widehat{\mathrm{risk}}_S(h) + \eta F_q(h)}$$

- e.g. RKHS regularization $F_p(h) = F_q(h) = ||h||^2_{\mathcal{H}}$.
- When $F_p = F_q$ we obtain same (unregularized) bound

Theorem (Risk Bound for Regularized Stochastic ERM)

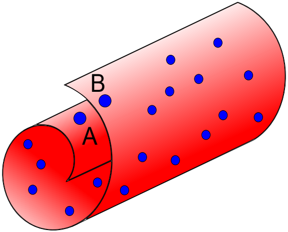
With probability at least $1 - \delta$,

$$\mathrm{kl}(\widehat{\mathrm{risk}}_S(G_Q), \mathrm{risk}(G_Q)) \leq \frac{1}{m}\left(\frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{4\xi(m)}{\delta}} + \frac{\gamma^2}{4m} + \ln\frac{2\xi(m)}{\delta}\right)$$

- But this should enable smaller γ

Regularization in Intrinsic Geometry of Data

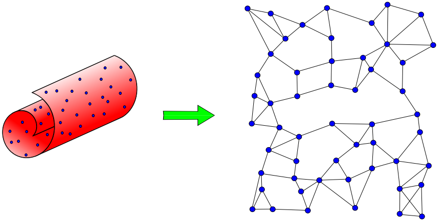
- Regularize w.r.t. interaction between hypotheses and geometry of data-generating distribution
- Data has its own *intrinsic* geometry



- e.g. intrinsic and extrinsic metrics can be very different
- Working assumption intrinsic geometry more suitable
- Correct setting for notions of function class complexity

Capturing Intrinsic Geometry of Data

- Intrinsic geometry learnt from random samples
- Given sample $\mathcal{S}$ of n points, form $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ on $\mathcal{S}$



- Define "smoothness" of h on $\mathcal{G}$

$$\widehat{U}_{\mathcal{S}}(h) := \frac{1}{n(n-1)} \sum_{ij} (h(X_i) - h(X_j))^2 W(X_i, X_j)$$

- Converges to smoothness w.r.t. data distribution (Hein et al.)
- Captures intuitions about how good classifiers interact with "true" structure of data
- Not possible without empirical geometry

Regularization in Intrinsic Geometry of Data

- Given $\mathcal{S} = \{(X_1, Y_1), ...(X_m, Y_m)\} \cup \{X_{m+1}, ...X_n\}$

$$p(h) := \frac{1}{Z'} e^{-\gamma \mathrm{risk}(h) + \eta U(h)} \quad q(h) := \frac{1}{Z} e^{-\gamma \widehat{\mathrm{risk}}_{\mathcal{S}}(h) + \eta \widehat{U}_{\mathcal{S}}(h)}$$

- $\widehat{U}_{\mathcal{S}}(h) := \frac{1}{n(n-1)} \sum_{ij} (h(X_i) - h(X_j))^2 W(X_i, X_j)$, "smoothness" on $\mathcal{G}$
- $U(h) := \mathbb{E}_{\mathcal{S}}[\widehat{U}_{\mathcal{S}}(h)]$
- To bound $\mathrm{KL}(Q||P)$ we must bound $(\mathbb{E}_{h\sim Q} - \mathbb{E}_{h\sim P})[U(h) - \widehat{U}_{\mathcal{S}}(h)]$
- $\widehat{U}_{\mathcal{S}}(h)$ is a U-statistic of order 2
- We need PAC-Bayes concentration of U-process...

PAC-Bayes *U*-process concentration

- $U_{\mathcal{S}}(h) := \frac{1}{n(n-1)} \sum_{i \neq j} f_h(X_i, X_j)$

Theorem (PAC-Bayes concentration for *U*-processes)

For all t, with probability at least $1 - \delta$

$$\mathbb{E}_{h\sim Q}[\widehat{U}_{\mathcal{S}}(h) - U(h)] \leq \frac{1}{t}\left(\mathrm{KL}(Q||P) + \frac{t^2(b-a)^2}{2n} + \ln\left(\frac{1}{\delta}\right)\right)$$

where $a \leq f_h(X, X') \leq b$

Proof.

Germain et. al's general recipe for PAC-Bayes bounds
Hoeffding's decomposition into martingales
Hoeffding's lemma recursively (as in Azuma/McDiarmid) □

Bound for Intrinsic Regularization

- Putting everything together we obtain a bound for the case,

$$p(h) := \frac{1}{Z'} e^{-\gamma \mathrm{risk}(h) + \eta U(h)} \quad q(h) := \frac{1}{Z} e^{-\gamma \widehat{\mathrm{risk}}_S(h) + \eta \widehat{U}_S(h)}$$

Theorem (Risk Bound for Intrinsic Regularization)

For $\eta < \sqrt{n}$, with probability at least $1 - \delta$

$$\mathrm{kl}(\widehat{\mathrm{risk}}_S(G_Q), \mathrm{risk}(G_Q)) \leq \frac{1}{m}\left(A^2 + B + A\sqrt{2B + A^2} + \ln \frac{\xi(m)}{\delta}\right)$$

$$A := \frac{\gamma\sqrt{n}}{2\sqrt{m}(\sqrt{n} - \eta)}$$

$$B := \frac{\sqrt{n}}{\sqrt{n} - \eta}\left(\gamma\sqrt{\frac{2}{m}\ln\frac{4\xi(m)}{\delta}} + \frac{2\eta}{\sqrt{n}}\left(32b^4w^2 + \ln\frac{4}{\delta}\right)\right)$$

- Controlling function class complexity in this way is unusual
- Flexibility of PAC-Bayes and localization

Gaussian Process Prediction

- Extend localization to Gaussian processes
- Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- RKHS $\mathcal{H} := \overline{\mathrm{span}\{K(\boldsymbol{x}, \cdot) \ : \ \boldsymbol{x} \in \mathcal{X}\}}$
- $h(\boldsymbol{x}) := \langle h, K(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}}$

$$p(h) := \frac{1}{Z'} e^{-\frac{\gamma}{2}||h-\mu||^2_{\mathcal{H}}} \quad q(h) := \frac{1}{Z} e^{-\frac{\gamma}{2}||h-\mu_{\mathcal{S}}||^2_{\mathcal{H}}}$$

where

$$\mu_{\mathcal{S}} := \operatorname*{argmin}_{h \in \mathcal{H}} \{\widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h) + \lambda ||h||^2_{\mathcal{H}}\} \quad \mu := \mathbb{E}_{\mathcal{S}}[\mu_{\mathcal{S}}].$$

- $\ell : \mathcal{Y} \times \mathcal{Y}$ convex, α-Lipschitz
- G_Q equivalent to Gaussian process $\{G_{\boldsymbol{x}}\}_{\boldsymbol{x} \in \mathcal{X}}$ on $\mathcal{X}$ with

$$\mathbb{E}[G_{\boldsymbol{x}}] = \mu_{\mathcal{S}}(\boldsymbol{x})$$

$$\mathbb{E}[(G_{\boldsymbol{x}} - \mathbb{E}[G_{\boldsymbol{x}}])(G_{\boldsymbol{x}'} - \mathbb{E}[G_{\boldsymbol{x}'}])] = \frac{1}{\gamma} K(\boldsymbol{x}, \boldsymbol{x}')$$

Gaussian Process Prediction 2 - Bounding the KL

- As usual to establish risk bound we bound $KL(Q||P)$

Lemma

$\mathrm{KL}(Q||P) = \frac{\gamma}{2}||\mu_{\mathcal{S}} - \mu||^2_{\mathcal{H}}$

Lemma

$$\mathbb{P}_{\mathcal{S}}\left(||\mu_{\mathcal{S}} - \mu||_{\mathcal{H}} \leq \frac{2\alpha\kappa}{\lambda}\sqrt{\frac{1}{m}\ln\frac{4}{\delta}}\right) \geq 1 - \delta$$

where $\kappa := \sup_{\boldsymbol{x}\in\mathcal{X}}\sqrt{K(\boldsymbol{x},\boldsymbol{x})}$

Proof.

Via bounded differences: consider

$$\mathcal{S} := \{(X_1, Y_1), ...(X_m, Y_m)\}$$
$$\mathcal{S}^{(i)} := \{(X_1, Y_1), ...(X_{i-1}, Y_{i-1}), (X_i', Y_i'), (X_{i+1}, Y_{i+1}), ...(X_m, Y_m)\}$$

By stability argument: $||\mu_{\mathcal{S}^{(i)}} - \mu_{\mathcal{S}}||_{\mathcal{H}} \leq \frac{\alpha\kappa}{\lambda m}$ then version of Azuma's inequality for Hilbert space-valued martingales □

Gaussian Process Prediction 3 - Risk bound

- recall

$$p(h) := \frac{1}{Z'} e^{-\frac{\gamma}{2}||h-\mu||^2_{\mathcal{H}}} \quad q(h) := \frac{1}{Z} e^{-\frac{\gamma}{2}||h-\mu_{\mathcal{S}}||^2_{\mathcal{H}}}$$

where

$$\mu_{\mathcal{S}} := \underset{h\in\mathcal{H}}{\text{argmin}}\{\widehat{\text{risk}}^{\ell}_{\mathcal{S}}(h) + \lambda||h||^2_{\mathcal{H}}\} \quad \mu := \mathbb{E}_{\mathcal{S}}[\mu_{\mathcal{S}}].$$

- Risk bound by putting all together

Theorem (Risk bound for Gaussian process prediction)

If $\ell(\cdot,\cdot)$ is α-Lipschitz, and $\mathcal{H}$ is separable then with probability at least $1-\delta$ over the draw of $\mathcal{S}$

$$\text{kl}(\widehat{\text{risk}}_{\mathcal{S}}(G_Q), \text{risk}(G_Q)) \leq \frac{1}{m}\left(\frac{\gamma\alpha^2\kappa^2}{\lambda^2 m}\log\frac{8}{\delta} + \ln\frac{2\xi(m)}{\delta}\right)$$

Conclusions

- Developed seemingly sharp risk analysis for Localization with Boltzmann prior/posterior
- Considered function class complexity and regularization
- Regularized w.r.t. interaction between hypotheses and data structure
- Extended the ideas to Gaussian Processes