On the Properties of Variational Approximations of Gibbs Posteriors

Pierre Alquier

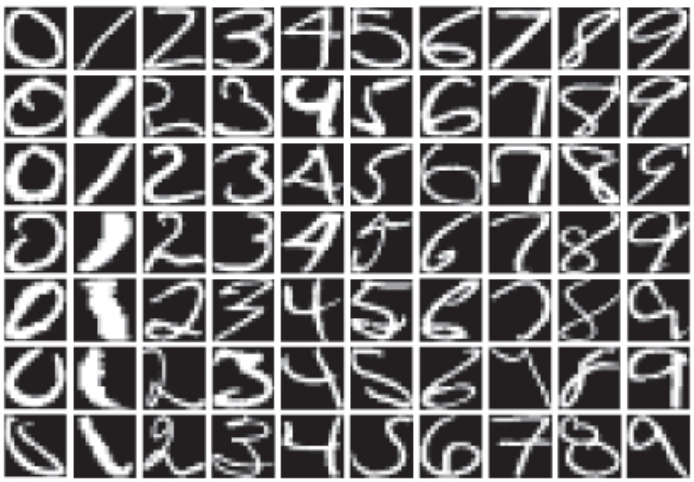
Big Data: Modeling, Estimation and Selection
Ecole Centrale de Lille - 10/06/2016

Learning vs. estimation

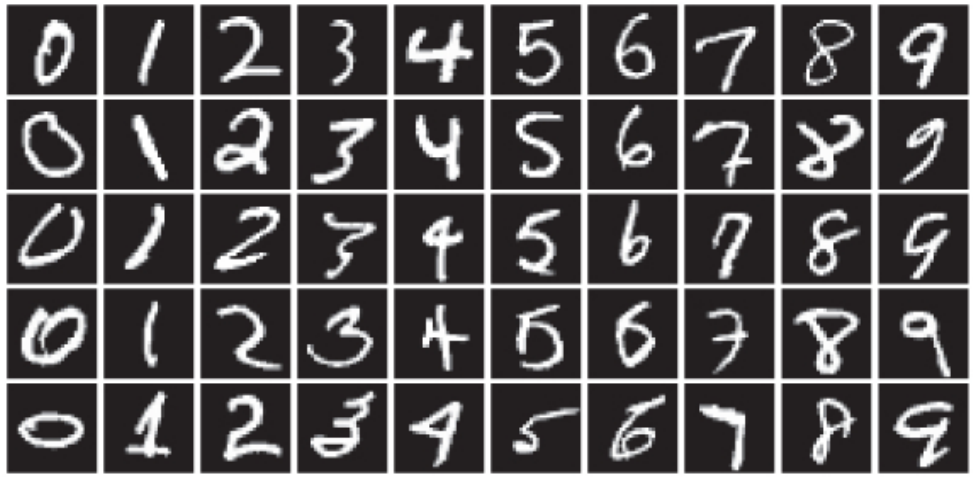
In many applications one would like to learn from a sample without being able to write the likelihood.

Learning vs. estimation

In many applications one would like to learn from a sample without being able to write the likelihood.



(a) USPS



(b) MNIST

Typical machine learning problem

Main ingredients :

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- a criterion of success, $R(\theta)$:

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- a criterion of success, $R(\theta)$:
  $\rightarrow$ for example $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$, $R(\theta) = \|\theta - \theta_0\|$ where θ_0 is a target parameter, ... we want $R(\theta)$ to be small. But note that it is unknown.

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- a criterion of success, $R(\theta)$:
  $\rightarrow$ for example $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$, $R(\theta) = \|\theta - \theta_0\|$ where θ_0 is a target parameter, ... we want $R(\theta)$ to be small. But note that it is unknown.
- an empirical proxy $r(\theta)$ for this criterion of success :

Typical machine learning problem

Main ingredients :

- observations object-label : (X_1, Y_1), (X_2, Y_2), ...
  $\rightarrow$ either given once and for all (batch learning), once at a time (online learning), upon request...
- a restricted set of predictors $(f_\theta, \theta \in \Theta)$.
  $\rightarrow$ $f_\theta(X)$ meant to predict Y.
- a criterion of success, $R(\theta)$:
  $\rightarrow$ for example $R(\theta) = \mathbb{P}(f_\theta(X) \neq Y)$, $R(\theta) = \|\theta - \theta_0\|$ where θ_0 is a target parameter, ... we want $R(\theta)$ to be small. But note that it is unknown.
- an empirical proxy $r(\theta)$ for this criterion of success :
  $\rightarrow$ for example $r(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(f_\theta(X_i) \neq Y_i)$.

Empirical risk minimization (ERM)

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} r(\theta).$$

Empirical risk minimization (ERM)

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} r(\theta).$$

Theorem (Vapnik and Chervonenkis, in the 70's)

Vapnik, V. (1998). *Statistical Learning Theory*, Springer.

Classification setting. Let d_Θ denote the VC-dim. of Θ.

$$\mathbb{P}\left\{ R(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} R(\theta) + 4\sqrt{\frac{d_\Theta \log(n+1) + \log(2)}{n}} + \sqrt{\frac{\log(2/\varepsilon)}{2n}} \right\} \geq 1 - \varepsilon.$$

ERM with linear classifiers

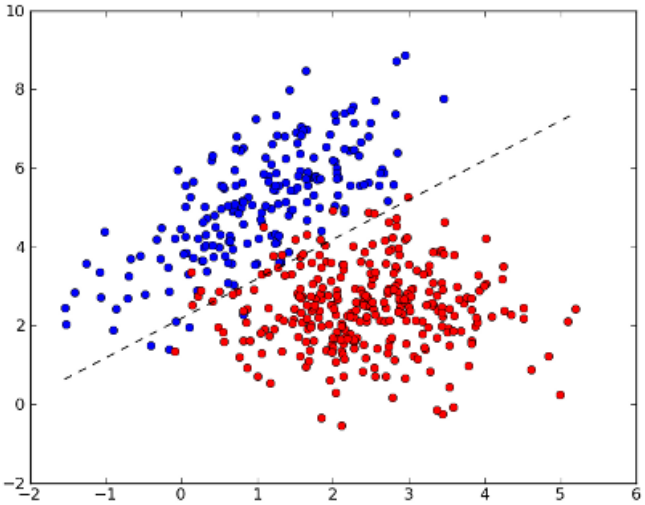
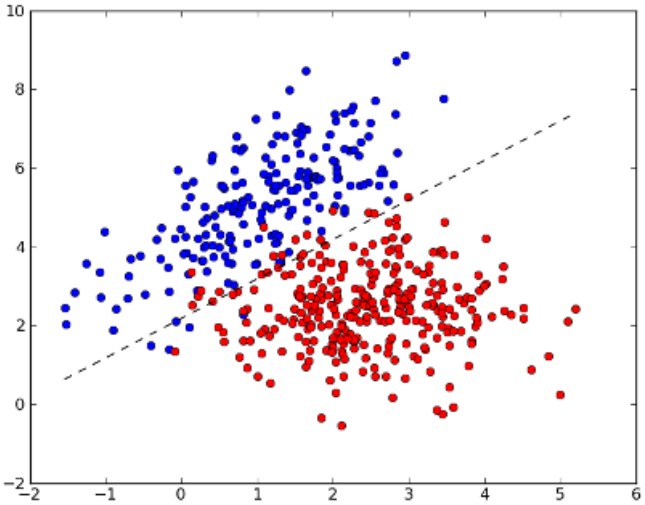


Table: Linear classifiers in $\mathbb{R}^p$: $d_\Theta = p + 1$. Source : http ://mlpy.sourceforge.net/

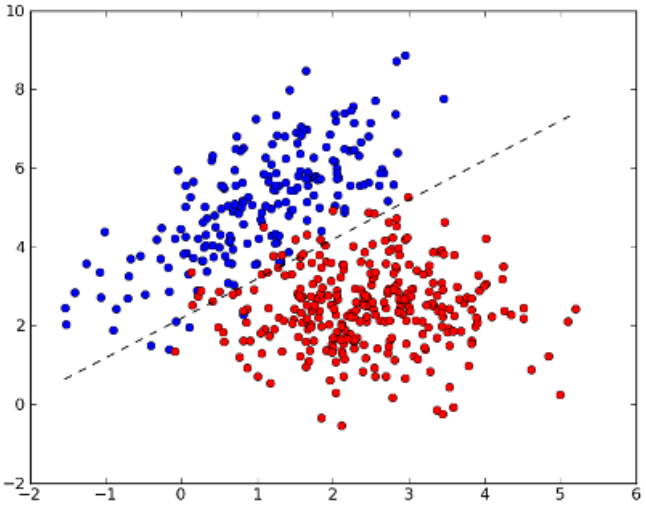
ERM with linear classifiers



Here $d_\Theta = 3$, $n = 500$.

Table: Linear classifiers in $\mathbb{R}^p$: $d_\Theta = p + 1$. Source : http ://mlpy.sourceforge.net/

ERM with linear classifiers



Here $d_\Theta = 3$, $n = 500$. With probability at least 90%,

$$R(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} R(\theta) + 0.842.$$

Table: Linear classifiers in $\mathbb{R}^p$: $d_\Theta = p + 1$. Source : http ://mlpy.sourceforge.net/

ERM with linear classifiers

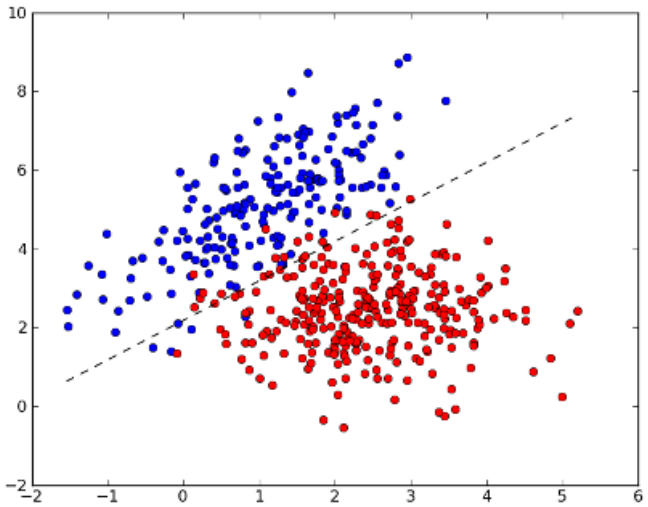


Table: Linear classifiers in $\mathbb{R}^p$: $d_\Theta = p + 1$. Source : http ://mlpy.sourceforge.net/

Here $d_\Theta = 3$, $n = 500$. With probability at least 90%,

$$R(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} R(\theta) + 0.842.$$

With $n = 5000$ we would have

$$R(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} R(\theta) + 0.301.$$

PAC-Bayesian bounds

One more ingredient :

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

EWA / pseudo-posterior / Gibbs estimator / ...

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

EWA / pseudo-posterior / Gibbs estimator / ...

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Example of context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

EWA / pseudo-posterior / Gibbs estimator / ...

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Example of context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- any $(f_\theta, \theta \in \Theta)$.

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

EWA / pseudo-posterior / Gibbs estimator / ...

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Example of context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- any $(f_\theta, \theta \in \Theta)$.
- $R(\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(Y, f_\theta(X))]$ for any *bounded* loss function $|\ell(\cdot,\cdot)| \leq B$.

PAC-Bayesian bounds

One more ingredient :

- a prior $\pi(\mathrm{d}\theta)$ on the parameter space.

EWA / pseudo-posterior / Gibbs estimator / ...

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Example of context :

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- any $(f_\theta, \theta \in \Theta)$.
- $R(\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(Y, f_\theta(X))]$ for any *bounded* loss function $|\ell(\cdot,\cdot)| \leq B$.
- $r(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$.

Catoni's bound for batch learning

Theorem

Catoni, O. (2007). *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 56 of Lecture Notes-Monograph Series, IMS.

$$\forall \lambda > 0, \quad \mathbb{P}\left\{ \int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \right.$$

$$\left. \leq \inf_{\rho} \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]\right\}$$

$$\geq 1-\varepsilon.$$

Catoni's bound for batch learning

Theorem

Catoni, O. (2007). *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 56 of Lecture Notes-Monograph Series, IMS.

$$\forall \lambda > 0, \quad \mathbb{P}\left\{ \int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \right.$$

$$\left. \leq \inf_{\rho} \left[\int R(\theta)\rho(\mathrm{d}\theta) + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]\right\}$$

$$\geq 1-\varepsilon.$$

improving on seminal work :

Shawe-Taylor, J. & Williamson, R. C. (1997). A PAC Analysis of a Bayesian Estimator. *COLT'97*.

McAllester, D. A. (1998). Some PAC-Bayesian Theorems. *COLT'98*.

Application : finite set of predictors $\theta_1, \ldots, \theta_M$

With π the uniform distribution on $\{\theta_1, \ldots, \theta_M\}$ we get

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)$$

$$\leq \inf_{\rho=\delta_{\theta_i}} \left[\int R \mathrm{d}\rho + \frac{\lambda B^2}{n} + \frac{2}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right]$$

Application : finite set of predictors $\theta_1, \ldots, \theta_M$

With π the uniform distribution on $\{\theta_1, \ldots, \theta_M\}$ we get

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)$$

$$\leq \inf_{\rho=\delta_{\theta_i}} \left[\int R \mathrm{d}\rho + \frac{\lambda B^2}{n} + \frac{2}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right]$$

$$\leq \inf_{1 \leq i \leq M} \left[R(\theta_i) + \frac{\lambda B^2}{n} + \frac{2}{\lambda} \left[\log(M) + \log\left(\frac{2}{\varepsilon}\right) \right] \right]$$

Application : finite set of predictors $\theta_1, \ldots, \theta_M$

With π the uniform distribution on $\{\theta_1, \ldots, \theta_M\}$ we get

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)$$

$$\leq \inf_{\rho=\delta_{\theta_i}} \left[\int R \mathrm{d}\rho + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]$$

$$\leq \inf_{1\leq i\leq M} \left[R(\theta_i) + \frac{\lambda B^2}{n} + \frac{2}{\lambda}\left[\log(M) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]$$

$$= \inf_{1\leq i\leq M} R(\theta_i) + 2B\sqrt{\frac{2\log(M)}{n}} + \log\left(\frac{2}{\varepsilon}\right)\sqrt{\frac{1}{2n\log(M)}}$$

for $\lambda = \dfrac{\sqrt{2n\log(M)}}{B}$.

How to sample from the Gibbs posterior ?

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

How to sample from the Gibbs posterior ?

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Popular method : MCMC $\rightarrow$ see Nial Friel's talk later today.

How to sample from the Gibbs posterior ?

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Popular method : MCMC $\rightarrow$ see Nial Friel's talk later today.

Dalalyan, A. and Tsybakov, A. (2011). Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Science*.

How to sample from the Gibbs posterior ?

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Popular method : MCMC $\rightarrow$ see Nial Friel's talk later today.

Dalalyan, A. and Tsybakov, A. (2011). Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Science*.

However : very hard to prove the convergence of the algorithm. Usually not possible to provide guarantees after a finite number of steps. See however

Joulin, A. & Ollivier, Y. (2010). Curvature, Concentration, and Error Estimates for Markov Chain Monte Carlo. *The Annals of Probability*.

Dalalyan, A. (2014). Theoretical Guarantees for Approximate Sampling from a Smooth and Log-Concave Density. *Preprint, to appear in JRSS-B*.

Variational Bayes methods

Idea from Bayesian statistics : approximate the posterior distribution $\pi(\theta|x)$. We fix a convenient family of probability distributions $\mathcal{F}$ and approximate the posterior by $\tilde{\pi}(\theta)$:

$$\tilde{\pi} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi(\cdot|x)).$$

Jordan, M. *et al* (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*.

Variational Bayes methods

Idea from Bayesian statistics : approximate the posterior distribution $\pi(\theta|x)$. We fix a convenient family of probability distributions $\mathcal{F}$ and approximate the posterior by $\tilde{\pi}(\theta)$:

$$\tilde{\pi} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi(\cdot|x)).$$

Jordan, M. *et al* (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*.

$\mathcal{F}$ is either parametric or non-parametric. In the parametric case, the problem boils down to an optimization problem :

$$\mathcal{F} = \{\rho_a, a \in \mathcal{A} \subset \mathbb{R}^d\} \dashrightarrow \min_{a \in \mathcal{A}} \mathcal{K}(\rho_a, \pi(\cdot|x)).$$

Variational Bayes methods

Idea from Bayesian statistics : approximate the posterior distribution $\pi(\theta|x)$. We fix a convenient family of probability distributions $\mathcal{F}$ and approximate the posterior by $\tilde{\pi}(\theta)$:

$$\tilde{\pi} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi(\cdot|x)).$$

Jordan, M. *et al* (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*.

$\mathcal{F}$ is either parametric or non-parametric. In the parametric case, the problem boils down to an optimization problem :

$$\mathcal{F} = \{\rho_a, a \in \mathcal{A} \subset \mathbb{R}^d\} \dashrightarrow \min_{a \in \mathcal{A}} \mathcal{K}(\rho_a, \pi(\cdot|x)).$$

Theoretical guarantees on the approximation ?

VB in PAC-Bayesian framework

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Then :

$$\begin{aligned}\mathcal{K}(\rho_a, \hat{\rho}_\lambda) &= \int \log\left[\frac{\mathrm{d}\rho_a}{\mathrm{d}\pi}\frac{\mathrm{d}\pi}{\mathrm{d}\hat{\rho}_\lambda}\right]\mathrm{d}\rho_a \\ &= \lambda\int r(\theta)\rho_a(\mathrm{d}\theta) + \mathcal{K}(\rho_a,\pi) + \log\int \exp[-\lambda r]\mathrm{d}\pi.\end{aligned}$$

VB in PAC-Bayesian framework

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r(\theta)\right]\pi(\mathrm{d}\theta).$$

Then :

$$\begin{aligned}\mathcal{K}(\rho_a, \hat{\rho}_\lambda) &= \int \log\left[\frac{\mathrm{d}\rho_a}{\mathrm{d}\pi}\frac{\mathrm{d}\pi}{\mathrm{d}\hat{\rho}_\lambda}\right]\mathrm{d}\rho_a \\ &= \lambda\int r(\theta)\rho_a(\mathrm{d}\theta) + \mathcal{K}(\rho_a,\pi) + \log\int \exp[-\lambda r]\mathrm{d}\pi.\end{aligned}$$

We put

$$\tilde{a}_\lambda = \arg\min_{a\in\mathcal{A}}\left[\lambda\int r(\theta)\rho_a(\mathrm{d}\theta) + \mathcal{K}(\rho_a,\pi)\right] \text{ and } \tilde{\rho}_\lambda = \rho_{\hat{a}_\lambda}.$$

A PAC-Bound for VB Approximation

Theorem

Alquier, P., Ridgway, J. & Chopin, N. (2015). On the Properties of Variational Approximations of Gibbs Posteriors. *Preprint, accepted for publication in JMLR*.

$$\forall \lambda > 0, \quad \mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_\lambda(\mathrm{d}\theta) \right.$$

$$\left. \leq \inf_{a \in \mathcal{A}} \left[\int R(\theta)\rho_a(\mathrm{d}\theta) + \frac{\lambda}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho_a, \pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]\right\}$$

$$\geq 1 - \varepsilon.$$

A PAC-Bound for VB Approximation

Theorem

Alquier, P., Ridgway, J. & Chopin, N. (2015). On the Properties of Variational Approximations of Gibbs Posteriors. *Preprint, accepted for publication in JMLR*.

$$\forall \lambda > 0, \quad \mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_\lambda(\mathrm{d}\theta) \right.$$

$$\left. \leq \inf_{a \in \mathcal{A}} \left[\int R(\theta)\rho_a(\mathrm{d}\theta) + \frac{\lambda}{n} + \frac{2}{\lambda}\left[\mathcal{K}(\rho_a, \pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right]\right\}$$

$$\geq 1 - \varepsilon.$$

$\dashrightarrow$ if we can derive a tight oracle inequality from this bound, we know that the VB approximation is "at no cost".

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.
- $r(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[Y_i \neq f_\theta(X_i)]$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.
- $r(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[Y_i \neq f_\theta(X_i)]$.
- Gaussian prior $\pi = \mathcal{N}(0, \vartheta I)$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.
- $r(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[Y_i \neq f_\theta(X_i)]$.
- Gaussian prior $\pi = \mathcal{N}(0, \vartheta I)$.
- Gaussian approx. of the posterior :
  $\mathcal{F} = \left\{ \mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^d, \Sigma \text{ s. pos. def.} \right\}$.

Application to a linear classification problem

- (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) iid from $\mathbb{P}$.
- $f_\theta(x) = \mathbf{1}(\langle \theta, x \rangle \geq 0)$, $x, \theta \in \mathbb{R}^d$.
- $R(\theta) = \mathbb{P}[Y \neq f_\theta(X)]$.
- $r(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[Y_i \neq f_\theta(X_i)]$.
- Gaussian prior $\pi = \mathcal{N}(0, \vartheta I)$.
- Gaussian approx. of the posterior :
  $\mathcal{F} = \left\{ \mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^d, \Sigma \text{ s. pos. def.} \right\}$.

Optimization criterion :

$$\frac{\lambda}{n} \sum_{i=1}^n \Phi\left(\frac{-Y_i \langle X_i, \mu \rangle}{\sqrt{\langle X_i, \Sigma X_i \rangle}} \right) + \frac{\|\mu\|^2}{2\vartheta} + \frac{1}{2}\left(\frac{1}{\vartheta} \mathrm{tr}(\Sigma) - \log|\Sigma| \right)$$

using deterministic annealing and gradient descent.

Application of the main theorem

Corollary

Assume that, for $\|\theta\| = \|\theta'\| = 1$, $\mathbb{P}(\langle \theta, X \rangle \langle \theta', X \rangle) \leq c\|\theta - \theta'\|$ and take $\lambda = \sqrt{nd}$ and $\vartheta = 1/\sqrt{d}$. Then

$$\mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\theta R(\theta) + \sqrt{\frac{d}{n}}\left[\log(4n\mathrm{e}^2) + c\right] + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\sqrt{nd}} \right\} \geq 1 - \varepsilon.$$

Application of the main theorem

Corollary

Assume that, for $\|\theta\| = \|\theta'\| = 1$, $\mathbb{P}(\langle \theta, X \rangle \langle \theta', X \rangle) \leq c\|\theta - \theta'\|$ and take $\lambda = \sqrt{nd}$ and $\vartheta = 1/\sqrt{d}$. Then

$$\mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\theta R(\theta) + \sqrt{\frac{d}{n}}\left[\log(4n\mathrm{e}^2) + c\right] + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\sqrt{nd}} \right\} \geq 1 - \varepsilon.$$

N.B : under margin assumption, possible to obtain d/n rates...

Sketch of the proof

By the main theorem, with probability at least $1-\varepsilon$,

$$\int R \mathrm{d}\tilde{\rho}_\lambda \leq \inf_{\rho = \mathcal{N}(\theta, s^2 I)} \left[\int R \mathrm{d}\rho + \frac{\lambda}{n} + \frac{2}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right].$$

Sketch of the proof

By the main theorem, with probability at least $1-\varepsilon$,

$$\int R \mathrm{d}\tilde{\rho}_\lambda \leq \inf_{\rho = \mathcal{N}(\theta, s^2 I)} \left[\int R \mathrm{d}\rho + \frac{\lambda}{n} + \frac{2}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right].$$

As $\pi = \mathcal{N}(0, \vartheta I)$ we have

$$\mathcal{K}(\rho, \pi) = \frac{1}{2} \left[M \left(\frac{s^2}{\vartheta} - 1 + \log\left(\frac{\vartheta}{s^2}\right) \right) + \frac{\|\theta_0\|^2}{\vartheta} \right].$$

Sketch of the proof

By the main theorem, with probability at least $1-\varepsilon$,

$$\int R \mathrm{d}\tilde{\rho}_\lambda \leq \inf_{\rho = \mathcal{N}(\theta, s^2 I)} \left[\int R \mathrm{d}\rho + \frac{\lambda}{n} + \frac{2}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right].$$

As $\pi = \mathcal{N}(0, \vartheta I)$ we have

$$\mathcal{K}(\rho, \pi) = \frac{1}{2} \left[M \left(\frac{s^2}{\vartheta} - 1 + \log\left(\frac{\vartheta}{s^2}\right) \right) + \frac{\|\theta_0\|^2}{\vartheta} \right].$$

Then

$$\int R \mathrm{d}\rho \leq R(\theta) + \int 2c \|u - \theta\| \rho(\mathrm{d}u) \leq R(\theta) + 2c\sqrt{M}\sigma.$$

Sketch of the proof

By the main theorem, with probability at least $1-\varepsilon$,

$$\int R \mathrm{d}\tilde{\rho}_\lambda \leq \inf_{\rho = \mathcal{N}(\theta, s^2 I)} \left[\int R \mathrm{d}\rho + \frac{\lambda}{n} + \frac{2}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right].$$

As $\pi = \mathcal{N}(0, \vartheta I)$ we have

$$\mathcal{K}(\rho, \pi) = \frac{1}{2} \left[M \left(\frac{s^2}{\vartheta} - 1 + \log\left(\frac{\vartheta}{s^2}\right) \right) + \frac{\|\theta_0\|^2}{\vartheta} \right].$$

Then

$$\int R \mathrm{d}\rho \leq R(\theta) + \int 2c \|u - \theta\| \rho(\mathrm{d}u) \leq R(\theta) + 2c\sqrt{M}\sigma.$$

Chose adequate values for λ, ϑ and s^2 to conclude.

Test on real data

Dataset	Covariates	VB	SMC	SVM
Pima	7	21.3	22.3	30.4
Credit	60	33.6	32.0	32.0
DNA	180	23.6	23.6	20.4
SPECTF	22	06.9	08.5	10.1
Glass	10	19.6	23.3	4.7
Indian	11	25.5	26.2	26.8
Breast	10	1.1	1.1	1.7

Table: Comparison of misclassification rates (%). Last column : kernel-SVM with radial kernel. The hyper-parameters λ and ϑ are chosen by cross-validation.

Convexification of the loss

Can replace the 0/1 loss by a convex surrogate at "no" cost :

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*.

Convexification of the loss

Can replace the 0/1 loss by a convex surrogate at "no" cost :

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*.

- $R(\theta) = \mathbb{E}[(1 - Yf_\theta(X))_+]$ (hinge loss).
- $r(\theta) = \frac{1}{n}\sum_{i=1}^n (1 - Y_i f_\theta(X_i))_+$.
- Gaussian approx. : $\mathcal{F} = \left\{\mathcal{N}(\mu, \sigma^2 I), \mu \in \mathbb{R}^d, \sigma > 0\right\}$.

Convexification of the loss

Can replace the 0/1 loss by a convex surrogate at "no" cost :

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*.

- $R(\theta) = \mathbb{E}[(1 - Yf_\theta(X))_+]$ (hinge loss).
- $r(\theta) = \frac{1}{n}\sum_{i=1}^n (1 - Y_i f_\theta(X_i))_+$.
- Gaussian approx. : $\mathcal{F} = \left\{\mathcal{N}(\mu, \sigma^2 I), \mu \in \mathbb{R}^d, \sigma > 0\right\}$.

$\dashrightarrow$ the following criterion (which turns out to be convex !) :

$$\frac{1}{n}\sum_{i=1}^n (1 - Y_i \langle \mu, X_i \rangle) \Phi\left(\frac{1 - Y_i \langle \mu, X_i \rangle}{\sigma \|X_i\|_2}\right)$$

$$+\frac{1}{n}\sum_{i=1}^n \sigma \|X_i\| \varphi\left(\frac{1 - Y_i \langle \mu, X_i \rangle}{\sigma \|X_i\|_2}\right) + \frac{\|\mu\|_2^2}{2\vartheta} + \frac{d}{2}\left(\frac{\vartheta}{\sigma^2} - \log \sigma^2\right).$$

Application of the main theorem

Optimization with stochastic gradient descent on a ball of radius M. On this ball, the objetive function is L-Lipschitz. After k step, we have the approximation $\tilde{\rho}_\lambda^{(k)}$ of the posterior.

Corollary

Assume $\|X\| \leq c_x$ a.s., take $\lambda = \sqrt{nd}$ and $\vartheta = 1/\sqrt{d}$. Then

$$\mathbb{P}\left\{ \int R(\theta)\tilde{\rho}_\lambda^{(k)}(\mathrm{d}\theta) \leq \inf_\theta R(\theta) + \frac{LM}{\sqrt{1+k}} + \frac{c_x}{2}\sqrt{\frac{d}{n}}\log\left(\frac{n}{d}\right) + \frac{\frac{c_x^2+1}{2c_x} + 2c_x \log\left(\frac{2}{\varepsilon}\right)}{\sqrt{nd}} \right\} \geq 1-\varepsilon.$$

The PACVB package (James Ridgway)

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

PACVB: Variational Bayes (VB) Approximation of Gibbs Posteriors with Hinge Losses

Variational Bayesian approximations of Gibbs measures with hinge losses for classification and ranking.

Version:	1.1
Depends:	Rcpp, MASS
LinkingTo:	Rcpp, RcppArmadillo, BH
Published:	2016-02-04
Author:	James Ridgway
Maintainer:	James Ridgway <james.ridgway at bristol.ac.uk>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
NeedsCompilation:	yes
CRAN checks:	PACVB results

Downloads:

Reference manual:	PACVB.pdf
Package source:	PACVB_1.1.tar.gz
Windows binaries:	r-devel: PACVB_1.1.zip, r-release: PACVB_1.1.zip, r-oldrel: PACVB_1.1.zip
OS X Snow Leopard binaries:	r-release: PACVB_1.1.tgz, r-oldrel: not available
OS X Mavericks binaries:	r-release: PACVB_1.1.tgz

How to use PACVB ?

```
 > X
      [,1]        [,2]      [,3]
 [1,]    1 -1.48290060 0.3974124
 [2,]    1 -1.05599316 0.2554146
 [3,]    1  0.63464838 1.5370450
 [4,]    1 -0.36583539 1.5540228
 [5,]    1  0.08339866 0.9395758
...
 > Y
[1]  1 -1  1  1  1...
```

How to use PACVB ?

```
 > Sol = GDHinge(X,Y,lambda=25)
 > Sol
$m
          [,1]         [,2]      [,3]
[1,] 2.223396 -0.02744416 -1.205612

$s
[1] -1.406072

$bound
[1] 0.7990907
```

Application : collaborative filtering

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

Application : collaborative filtering

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	???
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

Application : collaborative filtering

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	👍
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

1-bit matrix completion

Object of interest : an $m_1 \times m_2$ matrix M, values in

$$\{👎, 👍\} = \{-1, +1\}.$$

1-bit matrix completion

Object of interest : an $m_1 \times m_2$ matrix M, values in

$$\{👎, 👍\} = \{-1, +1\}.$$

Entries $X_1 = (i_1, j_1), \ldots, (i_n, j_n)$ i.i.d from a distribution P, and $Y_\ell = M_{X_\ell}$.

1-bit matrix completion

Object of interest : an $m_1 \times m_2$ matrix M, values in

$$\{👎, 👍\} = \{-1, +1\}.$$

Entries $X_1 = (i_1, j_1), \ldots, (i_n, j_n)$ i.i.d from a distribution P, and $Y_\ell = M_{X_\ell}$.

Usual assumption : $\mathrm{rank}(M) = r \ll \min(m_1, m_2)$.

Prior and Gibbs posterior

Prior π

$$\underbrace{M}_{m_1 \times m_2} = \underbrace{L}_{m_1 \times K} \underbrace{R^T}_{K \times m_2},$$

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k), \quad \frac{1}{\gamma_k} \sim \Gamma(a, b).$$

Prior and Gibbs posterior

Prior π

$$\underbrace{M}_{m_1 \times m_2} = \underbrace{L}_{m_1 \times K} \underbrace{R^T}_{K \times m_2},$$

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k), \quad \frac{1}{\gamma_k} \sim \Gamma(a, b).$$

Empirical hinge risk :

$$r(L,R) = \frac{1}{n} \sum_{\ell=1}^{n} \left(1 - Y_\ell (LR^T) x_\ell\right)_+ .$$

Prior and Gibbs posterior

Prior π

$$\underbrace{M}_{m_1 \times m_2} = \underbrace{L}_{m_1 \times K} \underbrace{R^T}_{K \times m_2},$$

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k), \quad \frac{1}{\gamma_k} \sim \Gamma(a, b).$$

Empirical hinge risk :

$$r(L,R) = \frac{1}{n} \sum_{\ell=1}^{n} \left(1 - Y_\ell (LR^T) x_\ell\right)_+ .$$

Gibbs posterior : $\hat{\rho}_\lambda(L,R) = \exp\left[-\lambda r(L,R)\right] \pi(L,R)$.

Variational Approximation of the Gibbs posterior

Here, family of approximation : $\rho_a = \rho_{(\mathcal{L},\mathcal{R},S,\Sigma,\alpha,\beta)}$

$$L_{i,k} \text{ indep. } \mathcal{N}(\mathcal{L}_{i,k}, S_{i,k}),\ R_{i,k} \text{ indep. } \mathcal{N}(\mathcal{R}_{i,k}, \Sigma_{i,k}),$$

$$\frac{1}{\gamma_k} \text{ indep. } \Gamma(\alpha_k, \beta_k).$$

Variational Approximation of the Gibbs posterior

Here, family of approximation : $\rho_a = \rho_{(\mathcal{L},\mathcal{R},S,\Sigma,\alpha,\beta)}$

$$L_{i,k} \text{ indep. } \mathcal{N}(\mathcal{L}_{i,k}, S_{i,k}),\ R_{i,k} \text{ indep. } \mathcal{N}(\mathcal{R}_{i,k}, \Sigma_{i,k}),$$

$$\frac{1}{\gamma_k} \text{ indep. } \Gamma(\alpha_k, \beta_k).$$

In this case, the $\int r \mathrm{d}\rho_a$ is not tractable but we prove that

$$\forall a \in \mathcal{A}, \quad \int r \mathrm{d}\rho_a + \frac{\mathcal{K}(\rho_a, \pi)}{\lambda} \leq r\left(\mathcal{L}\mathcal{R}^T\right) + \mathcal{B}_\lambda(a)$$

for some known and tractable $\mathcal{B}_\lambda(a)$.

Definition

$$\tilde{\rho} = \arg\min_{\rho_a} r\left(\mathcal{L}\mathcal{R}^T\right) + \mathcal{B}_\lambda(a).$$

Theoretical result

Theorem

Cottet, V. & Alquier, P. (2016). 1-bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation. *Preprint*.

With proba. at least $1-\varepsilon$ on the sample,

$$\mathbb{P}_{(L,R)\sim\tilde{\rho},(i,j)\sim P}[\mathrm{sign}((LR^T)_{i,j}) \neq M_{i,j}] \leq C\frac{r(m_1+m_2)\log(n)}{n}$$

for some (known) $C > 0$.

Theoretical result

Theorem

Cottet, V. & Alquier, P. (2016). 1-bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation. *Preprint*.

With proba. at least $1-\varepsilon$ on the sample,

$$\mathbb{P}_{(L,R)\sim\tilde{\rho},(i,j)\sim P}[\mathrm{sign}((LR^T)_{i,j}) \neq M_{i,j}] \leq C\frac{r(m_1+m_2)\log(n)}{n}$$

for some (known) $C > 0$.

- in practice, blockwise coordinate optimization gives with gradient descent good results to compute $\tilde{\rho}$.

Theoretical result

Theorem

Cottet, V. & Alquier, P. (2016). 1-bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation. *Preprint*.

With proba. at least $1-\varepsilon$ on the sample,

$$\mathbb{P}_{(L,R)\sim\tilde{\rho},(i,j)\sim P}[\mathrm{sign}((LR^T)_{i,j}) \neq M_{i,j}] \leq C\frac{r(m_1+m_2)\log(n)}{n}$$

for some (known) $C > 0$.

- in practice, blockwise coordinate optimization gives with gradient descent good results to compute $\tilde{\rho}$.
- in the paper, extention for noisy observations.

Thank you !