Pour se remettre dans le bain

Pour se remettre dans le bain

1. Citer les quatre estimateurs bayésiens les plus couramment utilisés.
2. Qu'est-ce qu'une région de crédibilité ?
3. Comment prédire quand on est bayésien(ne) ?
4. Décrire l'approche quasi-bayésienne.
5. Pourquoi l'approche quasi-bayésienne peut-elle être vue comme une généralisation de l'apprentissage bayésien ?
6. Illustrer la provenance du quasi-posterior au moyen d'une formulation variationnelle.
7. Citer les quatre estimateurs quasi-bayésiens les plus couramment utilisés.
8. Rappeler le principe de l'ERM et de l'agrégation à poids convexe (EWA). Quel est le lien entre EWA et apprentissage quasi-bayésien ?

Assessing the performance: the oracle approach

Oracle:

$$\phi^\star \in \operatorname*{arg\,min}_{\phi \in \mathcal{Y}^{\mathcal{X}}} R(\phi).$$

Ultimate goal: do almost as well as the oracle.

Assessing the performance: the oracle approach

Oracle:

$$\phi^\star \in \operatorname*{arg\,min}_{\phi \in \mathcal{Y}^\mathcal{X}} R(\phi).$$

Ultimate goal: do almost as well as the oracle.

Excess risk:

$$\mathcal{E}(\cdot) = R(\cdot) - R^\star \geq 0, \qquad R^\star = R(\phi^\star).$$

PAC oracle inequalities

PAC oracle inequalities

Let $R^\star$ denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

PAC oracle inequalities

Let $R^\star$ denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

Key argument: concentration inequalities (*e.g.*, Bernstein, Hoeffding) + Legendre transform of the Kullback-Leibler divergence.

PAC oracle inequalities

Let $R^\star$ denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

Key argument: concentration inequalities (*e.g.*, Bernstein, Hoeffding) + Legendre transform of the Kullback-Leibler divergence.

Typical rates in the literature

- $\alpha = \frac{1}{2}$ (slow rate)
- $\alpha = 1$ (fast rate)

PAC oracle inequalities

Let $R^\star$ denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^\star + \frac{\Delta(\phi,\epsilon)}{n^\alpha} \right\}\right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

Key argument: concentration inequalities (*e.g.*, Bernstein, Hoeffding) + Legendre transform of the Kullback-Leibler divergence.

Typical rates in the literature

- $\alpha = \frac{1}{2}$ (slow rate)
- $\alpha = 1$ (fast rate)

Let $d = \dim(\mathcal{X})$

- $\Delta(\phi, \epsilon) \propto d + \log \frac{1}{\epsilon}$
- $\Delta(\phi, \epsilon) \propto \log d + \log \frac{1}{\epsilon}$

PAC oracle inequalities

Let $R^\star$ denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi\in\mathcal{F}} \left\{R(\phi) - R^\star + \frac{\Delta(\phi,\epsilon)}{n^\alpha}\right\}\right) \geq 1-\epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

Key argument: concentration inequalities (*e.g.*, Bernstein, Hoeffding) + Legendre transform of the Kullback-Leibler divergence.

Typical rates in the literature

- $\alpha = \frac{1}{2}$ (slow rate)
- $\alpha = 1$ (fast rate)

Let $d = \dim(\mathcal{X})$

- $\Delta(\phi,\epsilon) \propto d + \log\frac{1}{\epsilon}$
- $\Delta(\phi,\epsilon) \propto \log d + \log\frac{1}{\epsilon}$

The remainder term grows with d and the size of $\mathcal{F}$. It decreases with n.

Hoeffding inequality

Let $V_1, \ldots, V_n$ be independent real-valued random variables such that $a_i \leq V_i \leq b_i$ a.s. Let $\bar{V}_n = \frac{1}{n} \sum_{i=1}^{n} V_i$.

Hoeffding inequality

Let $V_1, \ldots, V_n$ be independent real-valued random variables such that $a_i \leq V_i \leq b_i$ a.s. Let $\bar{V}_n = \frac{1}{n}\sum_{i=1}^n V_i$. Then

$$\mathbb{P}(\bar{V}_n - \mathbb{E}\bar{V}_n > t) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \forall t > 0.$$

Lemma (Csiszar, 1975 ; Catoni, 2004)

Let $(A, \mathcal{A})$ be a measurable space. For any probability μ on $(A, \mathcal{A})$ and any measurable function $h : A \to \mathbb{R}$ such that $\int (\exp \circ h) \mathrm{d}\mu < \infty$,

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

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of μ, the supremum with respect to m on the right-hand side is reached for the Gibbs distribution g given by

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

The PAC-Bayesian theory

The PAC-Bayesian theory

...consists in producing PAC inequalities of quasi-Bayesian estimators.

The PAC-Bayesian theory

...consists in producing PAC inequalities of quasi-Bayesian estimators.

- Shawe-Taylor and Williamson (1997). A PAC analysis of a Bayes estimator, *COLT*
- McAllester (1998). Some PAC-Bayesian theorems, *COLT*
- McAllester (1999). PAC-Bayesian model averaging, *COLT*
- Catoni (2004). Statistical Learning Theory and Stochastic Optimization, Springer
- Audibert (2004). Une approche PAC-bayésienne de la théorie statistique de l'apprentissage, *Ph.D. thesis, Université Pierre & Marie Curie*
- Catoni (2007). PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, IMS
- Dalalyan and Tsybakov (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity, *Machine Learning*
- Alquier and Lounici (2011). PAC-Bayesian theorems for sparse regression estimation with exponential weights, *Electronic Journal of Statistics*

A flexible and powerful framework

A flexible and powerful framework

Numerous models addressed by the PAC-Bayesian literature.

- Alquier and Wintenberger (2012). Model selection for weakly dependent time series forecasting, *Bernoulli*
- Seldin, Laviolette, Cesa-Bianchi, Shawe-Taylor and Auer (2012). PAC-Bayesian inequalities for martingales, *IEEE Transactions on Information Theory*
- Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*
- Guedj and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*
- Guedj and Robbiano (2017). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- Alquier and Guedj (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*
- Li, Guedj and Loustau (2018). A Quasi-Bayesian Perspective to Online Clustering, *arXiv preprint*

A flexible and powerful framework

Numerous models addressed by the PAC-Bayesian literature.

- Alquier and Wintenberger (2012). Model selection for weakly dependent time series forecasting, *Bernoulli*
- Seldin, Laviolette, Cesa-Bianchi, Shawe-Taylor and Auer (2012). PAC-Bayesian inequalities for martingales, *IEEE Transactions on Information Theory*
- Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*
- Guedj and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*
- Guedj and Robbiano (2017). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- Alquier and Guedj (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*
- Li, Guedj and Loustau (2018). A Quasi-Bayesian Perspective to Online Clustering, *arXiv preprint*

Towards (almost) no assumptions to derive powerful results

- Bégin, Germain, Laviolette and Roy (2016). PAC-Bayesian bounds based on the Rényi divergence, *AISTATS*
- Alquier and Guedj (2017). Simpler PAC-Bayesian bounds for hostile data, *Machine Learning*

Sampling

Monte Carlo integration

Objective: approximation of an integral

$$\mathcal{I} = \int h(x) f(x) \mathrm{d}x.$$

Monte Carlo integration

Objective: approximation of an integral

$$\mathcal{J} = \int h(x) f(x) \mathrm{d}x.$$

Key idea: exploit the fact that $\mathcal{J} = \mathbb{E}_{X \sim f}[h(X)]$.

Monte Carlo principle

Monte Carlo principle

Sample a sequence $x_1, \ldots, x_m \sim f$.

Monte Carlo principle

Sample a sequence $x_1, \dots, x_m \sim f$.

Then use

$$\hat{\mathfrak{J}}_m = \frac{1}{m} \sum_{i=1}^{m} h(x_i)$$

as an estimator of $\mathfrak{J}$.

Monte Carlo principle

Sample a sequence $x_1, \ldots, x_m \sim f$.

Then use

$$\hat{\mathcal{J}}_m = \frac{1}{m} \sum_{i=1}^{m} h(x_i)$$

as an estimator of $\mathcal{J}$.

Justification: by the Strong Law of Large Numbers,

$$\hat{\mathcal{J}}_m \to \mathcal{J}.$$

Approximation evaluation

Approximation evaluation

Estimate the variance with

$$\nu_m = \frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(h(x_i) - \hat{\mathfrak{J}}_m)^2,$$

Approximation evaluation

Estimate the variance with

$$\nu_m = \frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(h(x_i) - \hat{\jmath}_m)^2,$$

and recall that for m large,

$$\frac{\hat{\jmath}_m - \mathbb{E}[h(X)]}{\sqrt{\nu_m}} \approx \mathcal{N}(0,1).$$

Importance sampling

Importance sampling

Simulating from f might not be the best idea: difficult/impossible, not optimal, ...

Importance sampling

Simulating from f might not be the best idea: difficult/impossible, not optimal, ...

An alternative to direct sampling is importance sampling, with the following trick:

Importance sampling

Simulating from f might not be the best idea: difficult/impossible, not optimal, ...

An alternative to direct sampling is importance sampling, with the following trick:

$$\mathbb{E}_{X\sim f}[h(X)] = \int h(x)f(x)\mathrm{d}x = \int h(x)\frac{f(x)}{g(x)}g(x)\mathrm{d}x$$
$$= \mathbb{E}_{X\sim g}\left[h(X)\frac{f(X)}{g(X)}\right],$$

which allows us to use other distributions.

Importance sampling

Importance sampling

Sample a sequence $x_1, \ldots, x_m \sim g$.

Importance sampling

Sample a sequence $x_1, \ldots, x_m \sim g$.

Then use

$$\hat{\mathcal{J}}_m = \frac{1}{m} \sum_{i=1}^{m} \frac{f(x_i)}{g(x_i)} h(x_i)$$

as an estimator of $\mathcal{J}$.

Justification: by the Strong Law of Large Numbers,

$$\hat{\mathcal{J}}_m \to \mathcal{J}.$$

Justification

1. Converges for any choice of the distribution g as long as the support of g contains the support of f.

Justification

1. Converges for any choice of the distribution g as long as the support of g contains the support of f.

2. Instrumental distribution g may be chosen among distributions easy to simulate.

Justification

1. Converges for any choice of the distribution g as long as the support of g contains the support of f.

2. Instrumental distribution g may be chosen among distributions easy to simulate.

3. The same sample generated from g can be used repeatedly, not only for different functions h but also for different densities f.

Choice of importance function

Choice of importance function

The optimal choice is

$$g^{\star}(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)\mathrm{d}x},$$

but the integral is unknown (obviously).

Choice of importance function

The optimal choice is

$$g^\star(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)\mathrm{d}x},$$

but the integral is unknown (obviously).

In practice, pick a density g which is close enough to $|h|f$ and for which $|h|f/g$ is bounded.

Choice of importance function

The optimal choice is

$$g^{\star}(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)\mathrm{d}x},$$

but the integral is unknown (obviously).

In practice, pick a density g which is close enough to $|h|f$ and for which $|h|f/g$ is bounded.

Beware: importance sampling suffers from the curse of dimensionality.

Choice of importance function

The optimal choice is

$$g^\star(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)\mathrm{d}x},$$

but the integral is unknown (obviously).

In practice, pick a density g which is close enough to $|h|f$ and for which $|h|f/g$ is bounded.

Beware: importance sampling suffers from the curse of dimensionality.

If $\sup f/g = M < +\infty$, the accept-reject algorithm may be used.

Choice of importance function

The optimal choice is

$$g^\star(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)\mathrm{d}x},$$

but the integral is unknown (obviously).

In practice, pick a density g which is close enough to $|h|f$ and for which $|h|f/g$ is bounded.

Beware: importance sampling suffers from the curse of dimensionality.

If $\sup f/g = M < +\infty$, the accept-reject algorithm may be used.

The instrumental function may be π (the prior). But often inefficient if data informative, and impossible is π is improper...

Sampling random variables

The fundamental theorem of sampling:

Sampling random variables

The fundamental theorem of sampling:if $U \sim \mathcal{U}(0,1)$, let F denote a CDF, then $F^{-}(U)$ is a random variable with distribution $\mathrm{d}F$.

Sampling random variables

The fundamental theorem of sampling:if $U \sim \mathcal{U}(0,1)$, let F denote a CDF, then $F^{-}(U)$ is a random variable with distribution $\mathrm{d}F$.

In practice, this theorem has a very limited scope since the pseudo-inverse F^{-} is usually unknown/not analytically tractable.

Accept-reject

Goal: sample $x \sim f$.

Accept-reject

Goal: sample $x \sim f$.

Ingredients: a density g such that

- f/g is upper-bounded by M (known)
- sampling from g is easy

Accept-reject

Goal: sample $x \sim f$.

Ingredients: a density g such that

- f/g is upper-bounded by M (known)
- sampling from g is easy

The algorithm:

1. Sample $z \sim g$ and $u \sim \mathcal{U}(0, Mg(z))$
2. If $u \leq f(z)$, take $x = z$, otherwise go back to 1.

How should we choose g?

The mean number of samples to obtain 1 realization is M!
Efficient algorithm = M close to 1.

How should we choose g?

The mean number of samples to obtain 1 realization is M!
Efficient algorithm $= M$ close to 1.

In practice we pick a g similar to f.

How should we choose g?

The mean number of samples to obtain 1 realization is M!
Efficient algorithm = M close to 1.

In practice we pick a g similar to f.

Nice fact: no need to know the normalizing constant of f!