Pour se remettre dans le bain

- Expliquer l'intérêt de la conjugaison, et en donner la définition.
- Quel est l'intérêt d'utiliser une vraisemblance issue d'une famille exponentielle naturelle ?
- Donner des exemples de lois conjuguées.
- Donner deux exemples de méthodes de construction de priors non-informatifs. Quelles sont les limites de ces méthodes ?

Bayesian estimators

Bayesian paradigm is based on the posterior distribution.

Bayesian estimators

Bayesian paradigm is based on the posterior distribution.

Many estimators may be derived from the posterior.

MAP

The maximum a posteriori estimator is defined as

$$\arg\max_{\theta} f(x|\theta)\pi(\theta)$$

(penalized likelihood estimator).

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors and corresponding MAP estimators:

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors and corresponding MAP estimators:

$$\pi(\theta) = \frac{1}{\mathcal{B}(1/2,1/2)}\theta^{-1/2}(1-\theta)^{-1/2} \quad \bigg|$$

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors and corresponding MAP estimators:

$$\pi(\theta) = \frac{1}{\mathcal{B}(1/2,1/2)}\theta^{-1/2}(1-\theta)^{-1/2} \;\Big|\; \hat{\theta} = \max\left(\frac{x-1/2}{n-1}, 0\right)$$

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors and corresponding MAP estimators:

$\pi(\theta) = \frac{1}{\mathcal{B}(1/2,1/2)}\theta^{-1/2}(1-\theta)^{-1/2}$	$\hat{\theta} = \max\left(\frac{x-1/2}{n-1}, 0\right)$
$\pi(\theta) = 1$	

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors and corresponding MAP estimators:

$\pi(\theta) = \frac{1}{\mathcal{B}(1/2,1/2)}\theta^{-1/2}(1-\theta)^{-1/2}$	$\hat{\theta} = \max\left(\frac{x-1/2}{n-1}, 0\right)$
$\pi(\theta) = 1$	$\hat{\theta} = x/n$

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors and corresponding MAP estimators:

$\pi(\theta) = \frac{1}{\mathcal{B}(1/2,1/2)}\theta^{-1/2}(1-\theta)^{-1/2}$	$\hat{\theta} = \max\left(\frac{x-1/2}{n-1}, 0\right)$
$\pi(\theta) = 1$	$\hat{\theta} = x/n$
$\pi(\theta) = \theta^{-1}(1-\theta)^{-1}$	

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors and corresponding MAP estimators:

$\pi(\theta) = \frac{1}{\mathcal{B}(1/2,1/2)} \theta^{-1/2}(1-\theta)^{-1/2}$	$\hat{\theta} = \max\left(\frac{x-1/2}{n-1}, 0\right)$
$\pi(\theta) = 1$	$\hat{\theta} = x/n$
$\pi(\theta) = \theta^{-1}(1-\theta)^{-1}$	$\hat{\theta} = \max\left(\frac{x-1}{n-2}, 0\right)$

Not always appropriate!

Consider $f(x|\theta) = \frac{1}{\pi}\left(1 + (x-\theta)^2\right)^{-1}$ and $\pi(\theta) = \frac{1}{2}\exp(-|\theta|)$.

Not always appropriate!

Consider $f(x|\theta) = \frac{1}{\pi}\left(1 + (x-\theta)^2\right)^{-1}$ and $\pi(\theta) = \frac{1}{2}\exp(-|\theta|)$.

The MAP is $\hat{\theta} = 0$!

Other possible estimators

- Mean: $\hat{\theta} = \mathbb{E}_{\theta \sim \pi(\cdot|x)} \theta = \int \theta \pi(\theta|x) \mathrm{d}\theta$.

Other possible estimators

- Mean: $\hat{\theta} = \mathbb{E}_{\theta \sim \pi(\cdot|x)} \theta = \int \theta \pi(\theta|x) \mathrm{d}\theta$.

- Median: $\hat{\theta} = \mathrm{med}(\pi(\cdot|x))$.

Other possible estimators

- Mean: $\hat{\theta} = \mathbb{E}_{\theta \sim \pi(\cdot|x)} \theta = \int \theta \pi(\theta|x) \mathrm{d}\theta$.
- Median: $\hat{\theta} = \mathrm{med}(\pi(\cdot|x))$.
- Realization: $\hat{\theta} \sim \pi(\cdot|x)$.

Other possible estimators

- Mean: $\hat{\theta} = \mathbb{E}_{\theta \sim \pi(\cdot|x)} \theta = \int \theta \pi(\theta|x) \mathrm{d}\theta$.
- Median: $\hat{\theta} = \mathrm{med}(\pi(\cdot|x))$.
- Realization: $\hat{\theta} \sim \pi(\cdot|x)$.
- ...

Credible regions

Natural confidence region: highest posterior density (HPD).

$$\mathcal{C}_\alpha^\pi = \{\theta; \pi(\theta|x) > \alpha\}.$$

Prediction

Reminder: if $x \sim f(\cdot|\theta)$ and $z \sim g(\cdot|x,\theta)$, the *predictive* distribution is

$$g^\pi(z|x) = \int g(z|x,\theta)\pi(\theta|x)\mathrm{d}\theta.$$

Example: normal prediction

Assume that $(x_1, \dots, x_n) \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$ and

$$\pi(\mu, \sigma^2) \propto (\sigma^2)^{-\lambda_\sigma - 3/2} \exp\left(\frac{-\lambda_\mu(\mu - \xi)^2 + \alpha}{2\sigma^2} \right).$$

The posterior is

$$\mathcal{N}\left(\frac{\lambda_\mu \xi + n\bar{x}_n}{\lambda_\mu + n}, \frac{\sigma^2}{\lambda_\mu + n} \right) \times \mathcal{IG}\left(\lambda_\sigma + n/2, \frac{\alpha + s_x^2 + \frac{n\lambda_\mu(\bar{x}_n - \xi)^2}{\lambda_\mu + n}}{2} \right).$$

(where $s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$)

$$
\begin{aligned}
g^{\pi}(z|x_1,\ldots,x_n) \propto & \int (\sigma^2)^{-\lambda_\sigma - 2 - n/2} \exp(-(z-\mu)^2/2\sigma^2) \\
& \times \exp\left(-(\lambda_\mu + n)\left(\mu - \frac{\lambda_\mu \xi + n\bar{x}_n}{\lambda_\mu + n}\right)^2 + \alpha + s_x^2 + \frac{n\lambda_\mu(\bar{x}_n - \xi)^2}{\lambda_\mu + n}\right) / 2\sigma^2 \\
& \quad \times \mathrm{d}(\mu, \sigma^2) \\
\propto & \left[\alpha + s_x^2 + \frac{n\lambda_\mu(\bar{x}_n - \xi)^2}{\lambda_\mu + n} + \frac{\lambda_\mu + n + 1}{\lambda_\mu + n}\left(z - \frac{\lambda_\mu \xi + n\bar{x}_n}{\lambda_\mu + n}\right)^2\right]^{-(2\lambda_\sigma + n + 1)/2}
\end{aligned}
$$

$$
\begin{aligned}
g^\pi(z|x_1,\dots,x_n) \propto \int & (\sigma^2)^{-\lambda_\sigma-2-n/2}\exp(-(z-\mu)^2/2\sigma^2) \\
& \times \exp\left(-(\lambda_\mu+n)\left(\mu-\frac{\lambda_\mu\xi+n\bar{x}_n}{\lambda_\mu+n}\right)^2+\alpha+s_x^2+\frac{n\lambda_\mu(\bar{x}_n-\xi)^2}{\lambda_\mu+n}\right)/2\sigma^2 \\
& \quad \times \mathrm{d}(\mu,\sigma^2) \\
\propto & \left[\alpha+s_x^2+\frac{n\lambda_\mu(\bar{x}_n-\xi)^2}{\lambda_\mu+n}+\frac{\lambda_\mu+n+1}{\lambda_\mu+n}\left(z-\frac{\lambda_\mu\xi+n\bar{x}_n}{\lambda_\mu+n}\right)^2\right]^{-(2\lambda_\sigma+n+1)/2}
\end{aligned}
$$

Student t distribution with mean $\frac{\lambda_\mu\xi+n\bar{x}_n}{\lambda_\mu+n}$ and $2\lambda_\sigma+n$ degrees of freedom.

Quasi-Bayesian learning, and a foretaste of PAC-Bayesian theory

The quasi-Bayesian approach

The quasi-Bayesian approach

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

The quasi-Bayesian approach

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

Quasi-posterior

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right)\pi(\cdot),$$

for some inverse temperature $\lambda > 0$.

The quasi-Bayesian approach

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

Quasi-posterior

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right)\pi(\cdot),$$

for some inverse temperature $\lambda > 0$.

Key fact!

The quasi-Bayesian approach

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

Quasi-posterior

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right)\pi(\cdot),$$

for some inverse temperature $\lambda > 0$.

Key fact! In general, $\exp\left(-\lambda R_n(\cdot)\right)$ is not a likelihood, hence the term quasi-Bayesian.

A generalization of Bayesian learning

The pseudo-likelihood term $\exp(-\lambda R_n(\cdot))$ is to be seen as a data fit term. However no model is attached to this representation! Quasi-Bayesian learning natively is a model-free learning paradigm.

A generalization of Bayesian learning

The pseudo-likelihood term $\exp(-\lambda R_n(\cdot))$ is to be seen as a data fit term. However no model is attached to this representation! Quasi-Bayesian learning natively is a model-free learning paradigm.

Tradeoff between interpretability (Bayesian modeling) and performance (quasi-Bayesian prediction). Echoes the celebrated similar tradeoff between ML and SL!

The missing link between machine learning and statistical learning?

Reminder:

- In ML, deterministic sequence $(\mathbf{x}_i, y_i)$,

$$\widehat{\phi}(\cdot) = \arg\min_m \left\{ \sum_{i=1}^n \ell(y_i, m(\mathbf{x}_i)) \right\}.$$

- In SL, random variables,

$$\widehat{\phi}(\cdot) = \arg\max_m \left\{ \sum_{i=1}^n \log \mathrm{d}P(Y_i; m(\mathbf{X}_i)) \right\}.$$

The missing link between machine learning and statistical learning?

Reminder:

- In ML, deterministic sequence $(\mathbf{x}_i, y_i)$,

$$\widehat{\phi}(\cdot) = \arg\min_{m} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\mathbf{x}_i) \right\}.$$

- In SL, random variables,

$$\widehat{\phi}(\cdot) = \arg\max_{m} \left\{ \sum_{i=1}^{n} \log \mathrm{d}P(Y_i; m(\mathbf{X}_i)) \right\}.$$

Quasi-Bayesian learning is a model-free approach yet relies on a stochastic assumption! Joining the best of two worlds.

A variational perspective

A variational perspective

With the classical quadratic loss $\ell\colon (a,b) \mapsto (a-b)^2$,

$$\widehat{\rho}_\lambda \in \operatorname*{arg\,inf}_{\rho \ll \pi} \left\{ \int_{\mathcal{F}} R_n(\phi)\rho(\mathrm{d}\phi) + \frac{\mathcal{K}(\rho,\pi)}{\lambda} \right\},$$

where $\mathcal{K}$ is the Kullback-Leibler divergence defined as

$$\mathcal{K}(\rho,\pi) = \begin{cases} \int_{\mathcal{F}} \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}\rho & \text{when } \rho \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

Typical quasi-Bayesian estimators

Typical quasi-Bayesian estimators

MAQP

$$\widehat{\phi}_\lambda \in \operatorname*{arg\,max}_{\phi \in \mathcal{F}} \widehat{\rho}_\lambda(\phi).$$

Typical quasi-Bayesian estimators

MAQP

$$\widehat{\phi}_\lambda \in \operatorname*{arg\,max}_{\phi \in \mathcal{F}} \widehat{\rho}_\lambda(\phi).$$

Mean

$$\widehat{\phi}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda} \phi = \int_{\mathcal{F}} \phi \widehat{\rho}_\lambda(\mathrm{d}\phi).$$

Typical quasi-Bayesian estimators

MAQP

$$\widehat{\phi}_\lambda \in \mathop{\mathrm{arg\,max}}_{\phi \in \mathcal{F}} \widehat{\rho}_\lambda(\phi).$$

Mean

$$\widehat{\phi}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda}\phi = \int_{\mathcal{F}} \phi \widehat{\rho}_\lambda(\mathrm{d}\phi).$$

Realization

$$\widehat{\phi}_\lambda \sim \widehat{\rho}_\lambda.$$

And so on.

Statistical aggregation revisited

Statistical aggregation revisited

Assume that $\mathcal{F}$ is finite.

Statistical aggregation revisited

Assume that $\mathcal{F}$ is finite.

The mean of the quasi-posterior $\widehat{\rho}_\lambda$ amounts to the celebrated exponentially weighted aggregate (EWA)

$$\widehat{\phi}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda}\phi = \sum_{i=1}^{\#\mathcal{F}} \omega_{\lambda,i}\phi_i$$

where

$$\omega_{\lambda,i} = \frac{\exp(-\lambda R_n(\phi_i))\pi(\phi_i)}{\sum_{j=1}^{\#\mathcal{F}} \exp(-\lambda R_n(\phi_j))\pi(\phi_j)}.$$

Guedj (2013). Agrégation d'estimateurs et de classificateurs : théorie et méthodes, *Ph.D. thesis, Université Pierre & Marie Curie*