Pour se remettre dans le bain

Pour se remettre dans le bain

1. Quelle est la différence entre *statistical learning* et *machine learning* ?
2. Donner la définition d'un algorithme d'apprentissage.
3. Quels sont les quatre grands types d'apprentissage ?
4. Que définissent *fat data* et *tall data* ?
5. Comment compare-t-on les performances d'algorithmes d'apprentissage ?
6. Quelle notion est souvent cruciale au moment de choisir une bonne fonction de perte ? En donner la définition.
7. Donner quelques exemples de fonctions de perte.
8. Enoncer le théorème de Bayes.
9. Quel est le rôle de la distribution *a priori* ?
10. Donner la définition d'un modèle bayésien.
11. Quelles sont les quatre distributions importantes en bayésien ?

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- Operates conditional upon the observations.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- Operates conditional upon the observations.
- Integrates simultaneously prior knowledge and information brought by data.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- Operates conditional upon the observations.
- Integrates simultaneously prior knowledge and information brought by data.
- Coherent and complete inferential scope and unique motor of inference.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- Operates conditional upon the observations.
- Integrates simultaneously prior knowledge and information brought by data.
- Coherent and complete inferential scope and unique motor of inference.
- Usually known up to a constant! $m(\mathbf{x})$ may be intractable.

Prior distributions

There is no such thing as *the* prior distribution!

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ.

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ.

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ.

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Improper priors: $\int \pi(\theta) \mathrm{d}\theta = +\infty$.

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ.

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Improper priors: $\int \pi(\theta)\mathrm{d}\theta = +\infty$.

A prior on θ may depend on additional parameters: those are called hyperparameters.

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

A family $\mathcal{F}$ of probability distributions is *conjugate* for a likelihood $f(x|\theta)$ if for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

A family $\mathcal{F}$ of probability distributions is *conjugate* for a likelihood $f(x|\theta)$ if for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

Only of interest when $\mathcal{F}$ is parameterized: switching from the prior to the posterior is reduced to an update of parameters.

Advantages

- Limited/finite information conveyed by data x

Advantages

- Limited/finite information conveyed by data x
- Preservation of the structure of the prior $\pi(\theta)$

Advantages

- Limited/finite information conveyed by data x
- Preservation of the structure of the prior $\pi(\theta)$
- Exchangeability

Advantages

- Limited/finite information conveyed by data x
- Preservation of the structure of the prior $\pi(\theta)$
- Exchangeability
- Allows for generation of "virtual observations"

Advantages

- Limited/finite information conveyed by data x
- Preservation of the structure of the prior $\pi(\theta)$
- Exchangeability
- Allows for generation of "virtual observations"
- Most importantly: tractability and simplicity

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*.

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*. When

$$f(x|\theta) = h(x)\exp(-\theta x - \psi(\theta))$$

the family is said to be *natural*.

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*. When

$$f(x|\theta) = h(x)\exp(-\theta x - \psi(\theta))$$

the family is said to be *natural*.

Main interest: allow for conjugate priors

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\exp(\theta\mu - \lambda\psi(\theta)), \quad \lambda > 0.$$

Classical exponential families and conjugate priors

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal	Normal	Normal
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$
		$\rho^{-1} = \sigma^2 + \tau^2$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal	Normal	Normal
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$
		$\rho^{-1} = \sigma^2 + \tau^2$
Poisson	Gamma	Gamma
$\mathcal{P}(\theta)$	$\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + x, \beta + 1)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + x, \beta + n - x)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative binomial $\mathrm{Neg}(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + m, \beta + x)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative binomial $\mathrm{Neg}(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	Dirichlet $\mathcal{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative binomial $\mathrm{Neg}(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	Dirichlet $\mathcal{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + 1/2, \beta + (\mu - x)^2/2)$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial\theta^2}\right].$$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial\theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial\theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial\theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- Relates to information theory

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial\theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- Relates to information theory
- Parameterization invariant

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- Relates to information theory
- Parameterization invariant
- Suffers from dimensionality curse

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial\theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- Relates to information theory
- Parameterization invariant
- Suffers from dimensionality curse
- Depends on data: incoherence with the likelihood principle

Example

If $x \sim \mathcal{B}(n, \theta)$, Jeffreys' prior is

$$\pi(\theta) \propto \mathcal{B}e(1/2, 1/2).$$

If $n \sim \mathrm{Neg}(x, \theta)$, Jeffreys' prior is

$$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$$

Non-informative priors: Laplace priors

With a finite set $\{\theta_1, \ldots, \theta_p\}$, uniform prior $\pi(\theta_i) = 1/p$.

Non-informative priors: Laplace priors

With a finite set $\{\theta_1, \ldots, \theta_p\}$, uniform prior $\pi(\theta_i) = 1/p$.

Continuous extension: $\pi(\theta) \propto 1$. This is no longer a probability distribution yet if $\int f(x|\theta)\mathrm{d}\theta < +\infty$, the posterior is well-defined as a probability distribution. Modeling is crucial. Weakness: lack of reparameterization invariance.