Introduction into Bayesian statistics

Maxim Kochurov

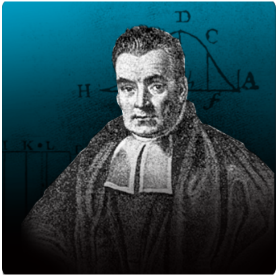
EF MSU

November 15, 2016

Content

Framework

- Treat everything as random variables
- Distribution over a variable is our ignorance about it
- Use bayes theorem



Notations

Bayes theorem

$$p(A|B) = \frac{p(B, A)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

Likelihood $= p(\mathcal{D}|\theta)$ - Measure of how well our data is described by the given model
Prior $= p(\theta)$ - Our prior beliefs about θ **before** seeing the data
Posterior $= p(\theta|\mathcal{D})$ - Our beliefs about θ **after** seeing the data
Evidence $= p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$ - Measure of how common is our data

Model specification

Bayesian model is specified by the given **data** $\mathcal{D}$, **likelihood function** $p(\mathcal{D}|\theta)$ - probability to see data under parameters θ, and **prior beliefs** $p(\theta)$ about parameters

Model specification

Bayesian model is specified by the given **data** $\mathcal{D}$, **likelihood function** $p(\mathcal{D}|\theta)$ - probability to see data under parameters θ, and **prior beliefs** $p(\theta)$ about parameters

Example

Linear Regression

- $\mathcal{D} = (X, y) = \{(x_i, y_i) | i \in [1, \dots, N]\}$

Model specification

Bayesian model is specified by the given **data** $\mathcal{D}$, **likelihood function** $p(\mathcal{D}|\theta)$ - probability to see data under parameters θ, and **prior beliefs** $p(\theta)$ about parameters

Example

Linear Regression

- $\mathcal{D} = (X, y) = \{(x_i, y_i)|i \in [1, \dots, N]\}$
- $p(\mathcal{D}|\theta) = p(e|\beta); \quad e = y - X\beta; \quad e \sim \mathcal{N}(0, \sigma^2 * I)$

Model specification

Bayesian model is specified by the given **data** $\mathcal{D}$, **likelihood function** $p(\mathcal{D}|\theta)$ - probability to see data under parameters θ, and **prior beliefs** $p(\theta)$ about parameters

Example

Linear Regression

- $\mathcal{D} = (X, y) = \{(x_i, y_i)|i \in [1, \ldots, N]\}$
- $p(\mathcal{D}|\theta) = p(e|\beta); \quad e = y - X\beta; \quad e \sim \mathcal{N}(0, \sigma^2 * I)$
- $p(\theta) = p(\beta); \quad \beta \sim \mathcal{N}(0, \sigma_\beta^2 * I)$

Model specification

Bayesian model is specified by the given **data** $\mathcal{D}$, **likelihood function** $p(\mathcal{D}|\theta)$ - probability to see data under parameters θ, and **prior beliefs** $p(\theta)$ about parameters

Example

Linear Regression

- $\mathcal{D} = (X, y) = \{(x_i, y_i) | i \in [1, \dots, N]\}$
- $p(\mathcal{D}|\theta) = p(e|\beta); \quad e = y - X\beta; \quad e \sim \mathcal{N}(0, \sigma^2 * I)$
- $p(\theta) = p(\beta); \quad \beta \sim \mathcal{N}(0, \sigma_\beta^2 * I)$
- $p(\beta|\mathcal{D}) = \frac{p(\mathcal{D}|\beta)p(\beta)}{p(\mathcal{D})}$

Model specification

Bayesian model is specified by the given **data** $\mathcal{D}$, **likelihood function** $p(\mathcal{D}|\theta)$ - probability to see data under parameters θ, and **prior beliefs** $p(\theta)$ about parameters

Example

Linear Regression

- $\mathcal{D} = (X, y) = \{(x_i, y_i) | i \in [1, \dots, N]\}$
- $p(\mathcal{D}|\theta) = p(e|\beta); \quad e = y - X\beta; \quad e \sim \mathcal{N}(0, \sigma^2 * I)$
- $p(\theta) = p(\beta); \quad \beta \sim \mathcal{N}(0, \sigma_\beta^2 * I)$
- $p(\beta|\mathcal{D}) = \frac{p(\mathcal{D}|\beta)p(\beta)}{p(\mathcal{D})}$
- Maximizing $p(\beta|\mathcal{D})$ by β gives a **MAP** solution. In this case it is called a ridge regression solution

Bayesian and Frequentist approach differences

	Frequentist	Bayesian
Randomness	Objective indefiniteness	Subjective ignorance
Variables	Random and Deterministic	Everything is random
Inference	ML Estimation	Posterior or MAP Estimation
Applicability	$n \gg 1$	$\forall n$

A story

A story

Consider you have to estimate θ in some problem1. All you have is your prior domain knowledge $p(\theta)$ and data $\mathcal{D}$.

A story

Consider you have to estimate θ in some problem1. All you have is your prior domain knowledge $p(\theta)$ and data $\mathcal{D}$.
Then you solve the problem and get posterior $p(\beta|\mathcal{D}_1)$, what's next?

A story

Consider you have to estimate θ in some problem1. All you have is your prior domain knowledge $p(\theta)$ and data $\mathcal{D}$.
Then you solve the problem and get posterior $p(\beta|\mathcal{D}_1)$, what's next?
Use $p(\beta|\mathcal{D}_1)$ in the future for problem2! Just treat posterior $p(\beta|\mathcal{D}_1)$ as a new prior.

A story

Consider you have to estimate θ in some problem1. All you have is your prior domain knowledge $p(\theta)$ and data $\mathcal{D}$.
Then you solve the problem and get posterior $p(\beta|\mathcal{D}_1)$, what's next?
Use $p(\beta|\mathcal{D}_1)$ in the future for problem2! Just treat posterior $p(\beta|\mathcal{D}_1)$ as a new prior.

$$p(\beta|\mathcal{D}_2) = \frac{p(\mathcal{D}_2|\beta)p(\beta|\mathcal{D}_1)}{p(\mathcal{D}_2)}$$