Statistical Methods - Bayesian Inference

Lecture 1

Mattias Villani

Sveriges Riksbank and Stockholm University

March 30, 2010

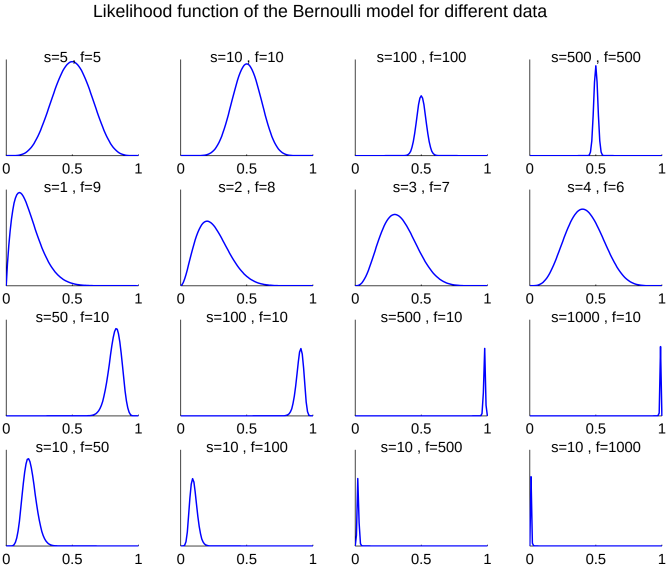
Overview

- Likelihood
- Bayesian inference
- Examples: The Bernoulli and Normal models

The likelihood function

- Bernoulli trials:
$$x_1, ..., x_n|\theta \overset{iid}{\sim} Bern(\theta).$$
- Likelihood:
$$\begin{aligned} p(x_1, ..., x_n|\theta) &= p(x_1|\theta) \cdots p(x_n|\theta) \\ &= \theta^s (1-\theta)^f, \end{aligned}$$
where $s = \sum_{i=1}^{n} x_i$ is the number of successes in the Bernoulli trials and $f = n - s$ is the number of failures.
- Given the data $x_1, ..., x_n$, we may plot $p(x_1, ..., x_n|\theta)$ as a function of θ.

The likelihood function from Bernoulli trials



- Will the likelihood give us un idea of which values of θ that should be regarded as probable (in some sense)? Kind of, but ... No!
- In order to say that one value of θ is more probable than another we clearly must think of θ as random. But θ may be something that we know is non-random, *e.g.* a fixed natural constant.
- Bayesian: doesn't matter if θ is fixed or random. What matters is whether or not You know the value of θ. If θ is uncertainty to You, then You can assign a probability distribution to θ which reflects Your knowledge about θ. Subjective probability.

Learning from data - Bayes' theorem

- Given that you have formulated a distribution for θ, $p(\theta)$, how can we learn from data? That is, how do we make the transition from $p(\theta) \rightarrow p(\theta|Data)$? Bayes' theorem is the key.
- One form of Bayes' theorem reads (A and B are events)

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

  So that Bayes' theorem 'reverses the conditioning', i.e. takes us from $p(B|A)$ to $p(A|B)$.
- Let $A = \theta$ and $B = Data$

$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)}.$$

- Interpreting the likelihood function as a probability density for θ is just as wrong as ignoring the factor $p(A)/p(B)$ in Bayes' theorem.

Generalized Bayes' theorem

- From your basic statistics textbook:

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{p(B)} = \frac{p(B|A_i)p(A_i)}{\sum_{i=1}^{k} p(B|A_i)p(A_i)}.$$

- Let $\theta_1, ..., \theta_k$ be k different values on a parameter θ. Bayes' Theorem:

$$p(\theta_i|Data) = \frac{p(Data|\theta_i)p(\theta_i)}{p(Data)} = \frac{p(Data|\theta_i)p(\theta_i)}{\sum_{i=1}^{k} p(Data|\theta_i)p(\theta_i)}.$$

- If θ takes on a continuum of values

$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{\int_\theta p(Data|\theta)p(\theta)d\theta}.$$

The joy of ignoring a normalizing constant

- When *Data* is known, $p(Data)$ in Bayes' theorem is just a constant that makes $p(\theta|Data)$ integrate to one. Example: $x \sim N(\mu, \sigma^2)$

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

- We may write

$$p(x) \propto \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

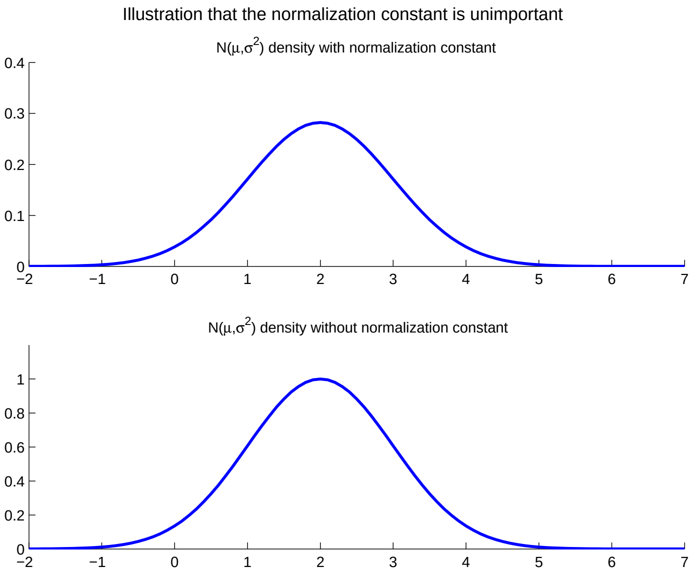
- Short form of Bayes' theorem

$$p(\theta|Data) \propto p(Data|\theta)p(\theta)$$

or

$$\text{Posterior} \propto \text{ Likelihood } \cdot \text{ Prior}$$

Normalization constant is not important



Bernoulli trials - Beta prior

- Model:

$$x_1, ..., x_n|\theta \overset{iid}{\sim} Bern(\theta)$$

- Prior:

$$\theta \sim Beta(\alpha, \beta)$$

$$x \sim Beta(\alpha, \beta) \Rightarrow p(x) = \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ for } 0 \le x \le 1.$$

- Posterior

$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto p(x_1, ..., x_n|\theta)p(\theta) \\ &= \theta^s(1-\theta)^f \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}. \end{aligned}$$

- But this is recognized as proportional to the $Beta(\alpha + s, \beta + f)$ density. That is, the prior-to-posterior mapping reads

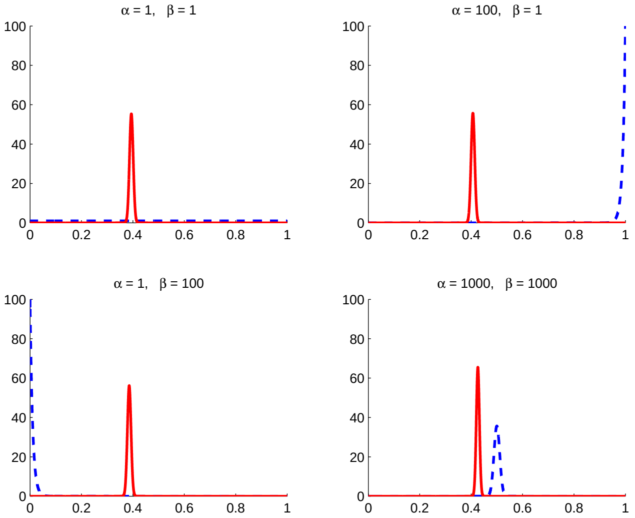
$$\theta \sim Beta(\alpha, b) \overset{x_1,...,x_n}{\Longrightarrow} \theta|x \sim Beta(\alpha + s, \beta + f).$$

Bernoulli example: spam emails

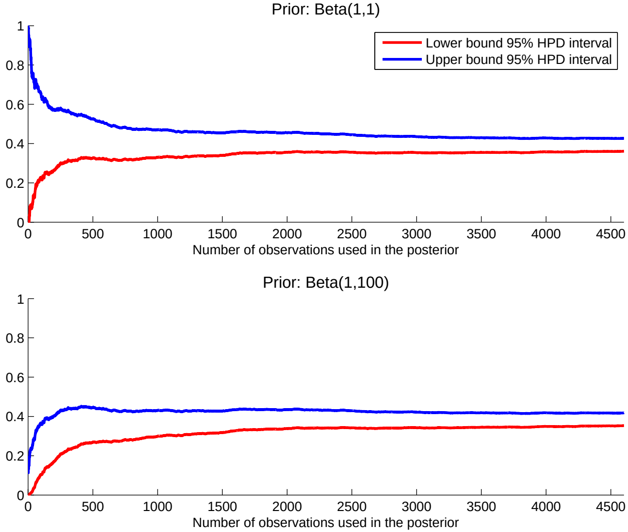
- George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam.
- Let $x_i = 1$ if i:th email is spam. Assume $x_i|\theta \overset{iid}{\sim} Bernoulli(\theta)$ and $\theta \sim \mathrm{Beta}(\alpha, \beta)$.
- Posterior

$$\theta|x \sim Beta(\alpha + 1813, \beta + 2788)$$

Spam data: The effect of different priors



Spam data: Posterior convergence



Bayesian updating

- Suppose: you already have $x_{1,}x_2, ..., x_n$ data points, and the corresponding posterior $p(\theta|x_1, ..., x_n)$
- Now, a fresh additional data point x_{n+1} arrive.
- The posterior based on all available data is

$$p(\theta|x_{1,}..., x_{n+1}) \propto p(x_{n+1}|\theta, x_1, ..., x_n)p(\theta|x_1, ..., x_n).$$

- The following is thus equivalent:
  - Analyzing the likelihood of all data $x_{1,}..., x_{n+1}$ with the prior based on no data $p(\theta)$
  - Analyzing the likelihood of the fresh data point x_{n+1} with the 'prior' equal to the posterior based on the old data $p(\theta|x_1, ..., x_n)$.
- Yesterday's posterior is today's prior.

Normal data with known variance - uniform prior

- Model:
$$x_1, ..., x_n|\mu, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2).$$
- Prior:
$$p(\theta) \propto c$$
- Likelihood (see Technical Appendix A):
$$\begin{aligned} p(x_1, ..., x_n|\theta, \sigma^2) &= \Pi_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(x_i - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2(\sigma^2/n)}(\theta - \bar{x})^2\right]. \end{aligned}$$
- Posterior
$$\theta \sim N(\bar{x}, \sigma^2/n)$$

Normal data with known variance - normal prior

- Prior

$$\theta \sim N(\mu_0, \tau_0^2)$$

- Posterior (see Technical Appendix A)

$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto p(x_1, ..., x_n|\theta, \sigma^2)p(\theta) \\ &\propto N(\theta|\mu_n, \tau_n^2), \end{aligned}$$

where

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2},$$

$$\mu_n = w\bar{x} + (1 - w)\mu_0,$$

and

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

Normal with known variance - normal prior, cont.

$$\theta \sim N(\mu_0, \tau_0^2) \overset{x_1,\ldots,x_n}{\Longrightarrow} \theta | x \sim N(\mu_n, \tau_n^2).$$

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$$

Posterior precision = Data precision + Prior precision

$$\mu_n = w\bar{x} + (1-w)\mu_0$$

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}$$

$$\text{Posterior mean} = \frac{\text{Data precision}}{\text{Posterior precision}}(\text{Data mean}) + \frac{\text{Prior precision}}{\text{Posterior precision}}(\text{Prior mean})$$