Bayesian Learning - Lecture 1

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Course overview

- Five two-day **modules** with:
  - Lectures
  - Exercises
  - Labs

Course overview

- Five two-day **modules** with:
  - Lectures
  - Exercises
  - Labs
- **Examination**
  - Lab reports, 2 credits
  - Bayesian project report, 4 credits

Course overview

- Five two-day **modules** with:
  - Lectures
  - Exercises
  - Labs
- **Examination**
  - Lab reports, 2 credits
  - Bayesian project report, 4 credits
- **Bayesian project report**
  - Individual
  - Use one or several real datasets to illustrate Bayesian analysis of **several** models
  - Deadline **January 6**

Lecture overview

- The likelihood function
- Bayesian inference
- The Bernoulli model

The likelihood function

- Bernoulli trials:

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

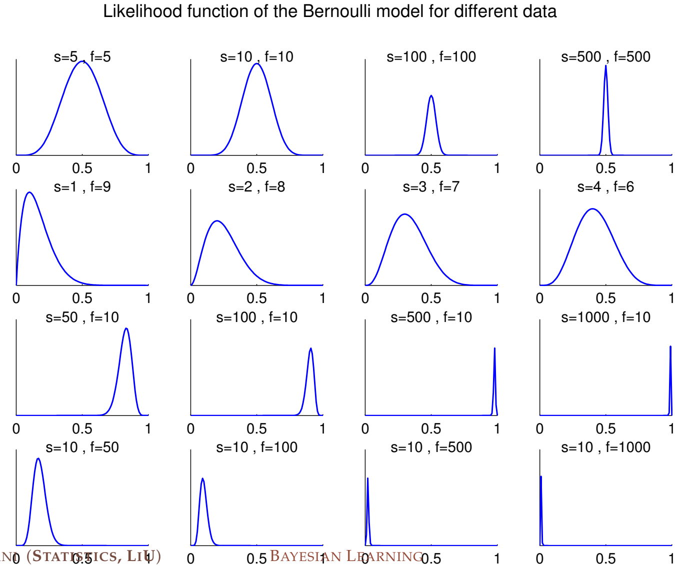
- Likelihood:

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

  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



Likelihood

- Two different roles played by $p(x_1, ..., x_n|\theta)$:
  - **a function of the data**, $x_1, ..., x_n$, for a *fixed* θ, it is a **probability distribution** for the data. Here the *data are random* and θ is fixed.
  - **a *deterministic function of*** θ for a ***fixed data*** sample. The **likelihood function**.
- The **likelihood principle**: Two experiments E_1 and E_2 that give rise to proportional likelihoods, i.e. $L_1(\theta) = c \cdot L_2(\theta)$ for all θ and some constant $c > 0$, should provide the same inference about θ.
- Many frequentist procedure violate the likelihood principle. Binomial vs inverse binomial sampling.
- **Birnbaum' theorem**: the likelihood principle follows directly from two universally accepted statistical principles.

Observed information

- The curvature of the likelihood is a measure of the informativeness (precision) of the data.
- The **observed information**

$$J_{\theta,\mathbf{x}} = -\frac{\partial^2 \ln L(\theta;\mathbf{x})}{\partial\theta^2}$$

- Asymptotic approximation of the likelihood function

$$N\left(\hat{\theta}, J^{-1}_{\hat{\theta},\mathbf{x}}\right),$$

  where $\hat{\theta}$ is the Maximum Likelihood Estimate (MLE) of θ.
- The normality can be proved heuristically by a second order Taylor expansion of the log-likelihood function.
- Example: Bernoulli data

$$J_{\hat{\theta},\mathbf{x}} = -\left.\frac{\partial^2 \ln L(\theta;\mathbf{x})}{\partial\theta^2}\right|_{\theta=\hat{\theta}} = \frac{s}{\hat{\theta}^2} + \frac{f}{(1-\hat{\theta})^2} = \frac{n}{\hat{\theta}(1-\hat{\theta})}.$$

Fisher information

- **Fisher information**

$$I_\theta = E_{\mathbf{x}|\theta}\left(J_{\theta,\mathbf{x}}\right),$$

where the expectation is with respect to the data distribution.

- The Fisher information is the information that can be **expected** before the data is observed.
- The **asymptotic distribution of the MLE**

$$\hat{\theta}|\theta \overset{approx}{\sim} N\left(\theta, \frac{1}{I_\theta}\right).$$

- Example: Bernoulli data

$$I_\theta = \frac{E(s)}{\theta^2} + \frac{E(f)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

Uncertainty and subjective probability

- 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**.
- Different types of prior information
  - Real **expert information**. Combo of previous studies and experience.
  - Vague prior information, or even **noninformative priors**.
  - **Reporting priors**
  - **Smoothness priors**. Regularization. Shrinkage. Big thing in modern statistics/machine learning.

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 $p(\theta) \to p(\theta|Data)$? 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

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

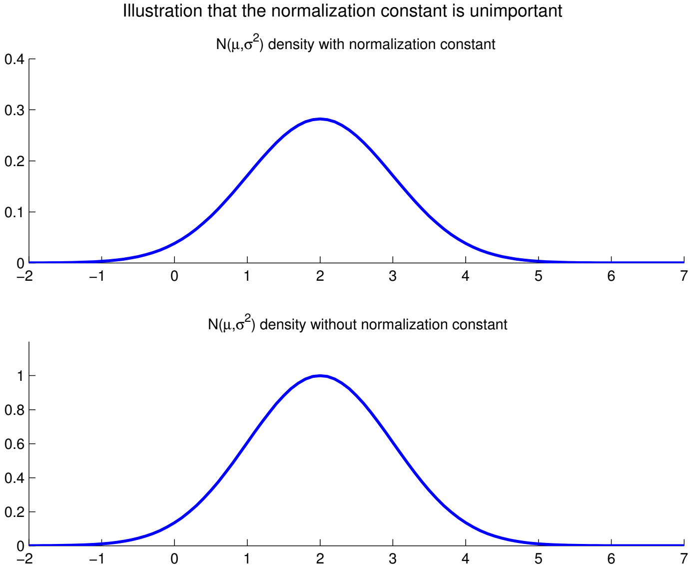
- Short form of Bayes' theorem

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

or

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

Normalization constant is not important



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} arrives.
- 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 are therefore 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.

Bernoulli trials - Beta prior

- Model:

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

- Prior:

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

$$p(y) = \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \text{ for } 0 \leq y \leq 1.$$

- Posterior

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

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

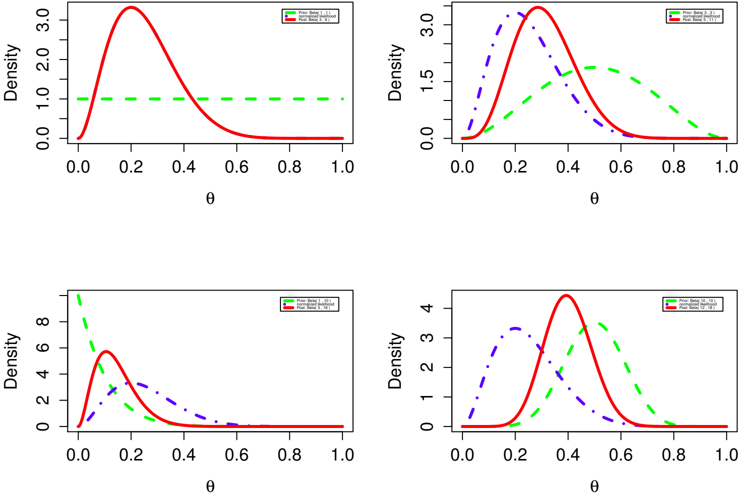
$$\theta \sim Beta(\alpha, \beta) \stackrel{x_1,...,x_n}{\Longrightarrow} \theta|x_1, ..., x_n \sim Beta(\alpha + s, \beta + f).$$

Bernoulli example: spam emails

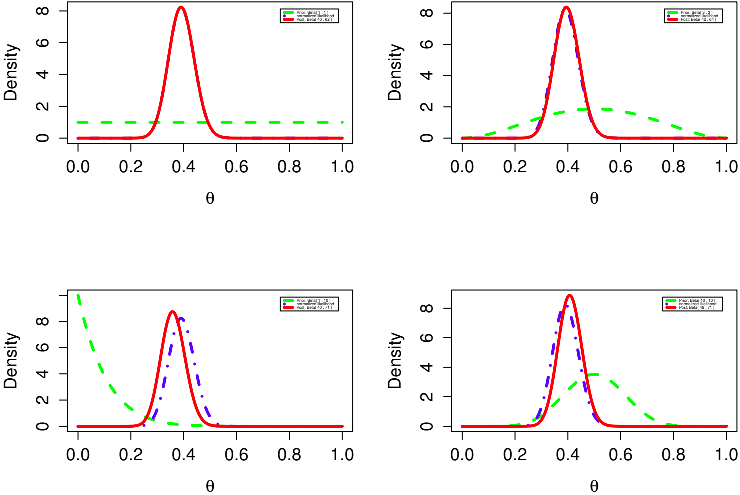
- George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam.
- Let $x_i = 1$ if i:th email is spam. Assume $x_i|\theta \overset{iid}{\sim} Bernoulli(\theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$.
- Posterior

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

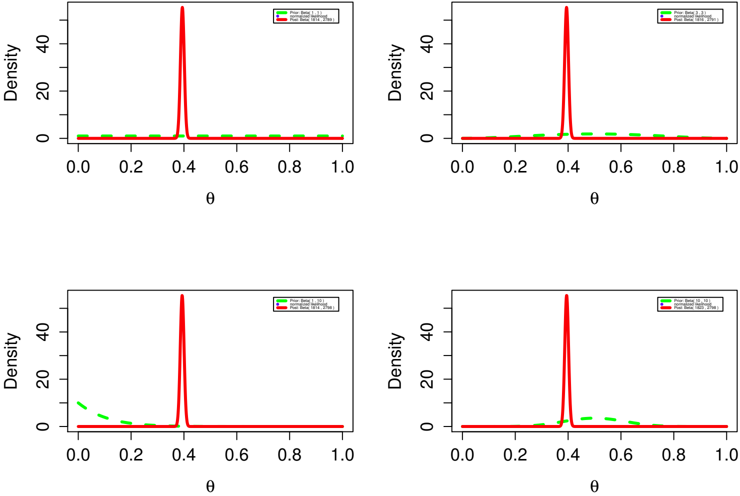
Spam data (n=10): prior sensitivity



Spam data (n=100): prior sensitivity



SPAM DATA (N=4601): PRIOR SENSITIVITY



Spam data: Posterior convergence