Bayesian Statistics

Fabio Sigrist

ETH Zurich, Autumn Semester 2019

Today's topics

- Course organization
- Bayes formula
- Introduction to the Bayesian approach
- Interpretations of probability

Course organization

Course organization

- **Homepage**:
  http://stat.ethz.ch/lectures/as19/bayesian-statistics.php
  - Course material and announcements
- A **script** is available on the homepage
- **Exercises**: will be provided on the homepage. No exercise lessons
- **Books**
  - Christian Robert, The Bayesian Choice, 2nd edition, Springer 2007
  - A. Gelman et al., Bayesian Data Analysis, 3rd edition, Chapman & Hall (2013)
- Oral **exam**, 20 minutes

Overview of lecture

1. Introduction to Bayesian statistics
   - Bayes formula
   - Basics of Bayesian statistics and comparison to frequentist statistics
   - Interpretations of probability
   - Likelihood principle

2. Prior distributions
   - Conjugate priors
   - Non-informative priors
   - Expert priors

Overview of lecture

3. Hierarchical Bayes models
   - Hierarchical Bayes models
   - Empirical Bayes methods
   - Model selection in linear regression

4. Bayesian computation
   - Laplace approximation
   - Independent Monte Carlo methods
   - Basics of Markov chain Monte Carlo
   - Some advanced computational methods

Bayes formula

Bayes formula: discrete case

Assume that

- (A_i) is a finite or countable partition of Ω
- $B \in \Omega$ with $P(B) > 0^*$

Bayes formula is given by

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)}$$
$$= \frac{P(B \mid A_i)P(A_i)}{\sum_{k:P(A_k)>0} P(B \mid A_k)P(A_k)}$$

- *Clicker question*
- *See comment on blackboard*

This has **two interpretations**:

- Frequentist: relative frequency of A_i among those repetitions where B occurs
- Subjective: how to modify the prior degree of belief $P(A_i)$ in A_i after observing that B has occurred

*We assume that all sets are measurable

Joint and marginal density

- Let $\mathbf{X} = (X_1, X_2)$ be a two-dimensional random vector with **joint density** f_{X_1,X_2}. I.e., for any (measurable) subset $A \subseteq \mathbb{R}^2$

$$\mathbb{P}((X_1, X_2) \in A) = \int_A f_{X_1,X_2}(x_1, x_2) dx_1 dx_2$$

  Heuristically, we have

$$\mathbb{P}(x_1 \leq X_1 \leq x_1 + dx_1, x_2 \leq X_2 \leq x_2 + dx_2) = f(x_1, x_2) dx_1 dx_2$$

- The **marginal densities** are

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) dx_2 \quad f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) dx_1$$

Conditional density

- We call
$$f_{X_2|X_1}(x_2 \mid x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)}$$
the **conditional** density of X_2 at x_2 given $X_1 = x_1$

- *See blackboard for a formal justification*

- We can write
$$f_{X_2|X_1}(x_2 \mid x_1) \propto f(x_1, x_2)$$
since the denominator $f_{X_1}(x_1)$ does not depend on x_2. I.e., $f_{X_1}(x_1)$ is ”just” a normalizing constant that depends on x_1 but not on x_2

Bayes formula: continuous case

Applying the formula for the conditional density twice in two directions, we obtain **Bayes formula**

$$f_{X_1|X_2}(x_1 \mid x_2) = \frac{f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1)}{\int_{-\infty}^{\infty} f_{X_2|X_1}(x_2 \mid x_1') f_{X_1}(x_1') dx_1'}$$
$$\propto f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1)$$

Comments

- The densities above need not be with respect to Lebesgue measure, any product measure can be used
- The existence of densities for the marginal P_{X_1} of X_1 is not needed

The Bayesian approach to statistics

Goal of inferential statistics

- In inferential statistics (both Bayesian and frequentist), one usually assumes that
  - The observations x are realizations of a random variable $X \in \mathbf{X}$. Often, $\mathbf{X} = \mathbb{R}^n$
  - One further assumes that $X \sim P_\theta$ where P_θ belongs to a set of distributions $\mathcal{P} = (P_\theta; \theta \in \Theta)$ parametrized by θ
    - In parametric statistics, Θ is a subset of $\mathbb{R}^p$
- The goal of inferential statistics is to make inference on the parameter θ by using the observed data x

Bayesian model

In Bayesian statistics, we assume:

1. The observations x are realizations of a random vector X which is distributed according to some parametric distribution P_θ with density $f(x \mid \theta)$

$$X = (X_1, \ldots, X_n) \sim f(x \mid \theta)$$

- $f(x \mid \theta)$ is called the **likelihood**

2. The parameter θ is distributed according to a distribution with density $\pi(\theta)$

- This density $\pi(\theta)$ is called the **prior***

- The prior has an epistemic interpretation. It describes our beliefs about possible values of θ before we see the data

*To simplify notation, we will denote by $\pi(\theta)$ both the prior distribution and the prior density

Bayesian model

- A Bayesian model thus consists of two parts: the likelihood $f(x \mid \theta)$ and the prior $\pi(\theta)$
- We interpret $f(x \mid \theta)$ as the conditional density of X given θ and the prior density $\pi(\theta)$ as the marginal density
- It follows that the **joint density $\pi(x, \theta)$ of (X, θ)** is given by

$$\pi(x, \theta) = \pi(\theta) f(x \mid \theta)$$

The posterior

- After observing the data x, the main object of interest in Bayesian inference is the **conditional density $\pi(\theta|x)$ of θ given $X = x$**. This is called the **posterior***

- Bayes formula tells us that the posterior density of θ is equal to

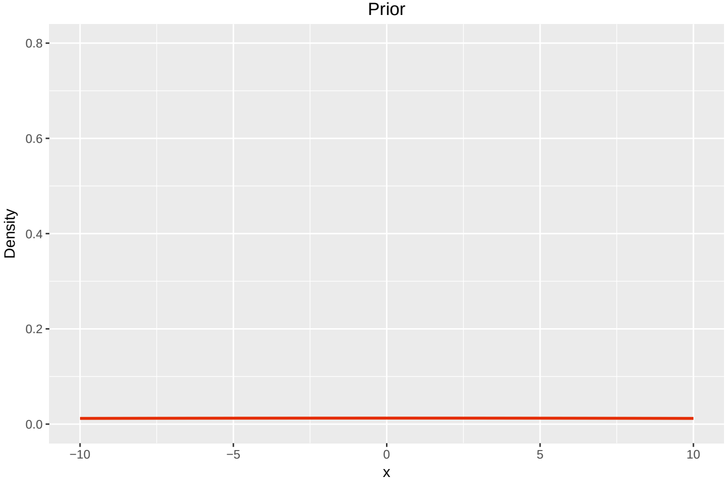
$$\pi(\theta|x) = \frac{\pi(\theta)f(x \mid \theta)}{\int_\Theta \pi(\theta')f(x \mid \theta')d\theta'} \propto \pi(\theta)f(x \mid \theta)$$

  - The posterior is proportional to the product of the prior and the likelihood
  - Interpretation: the posterior describes the uncertainty about θ after seeing the data

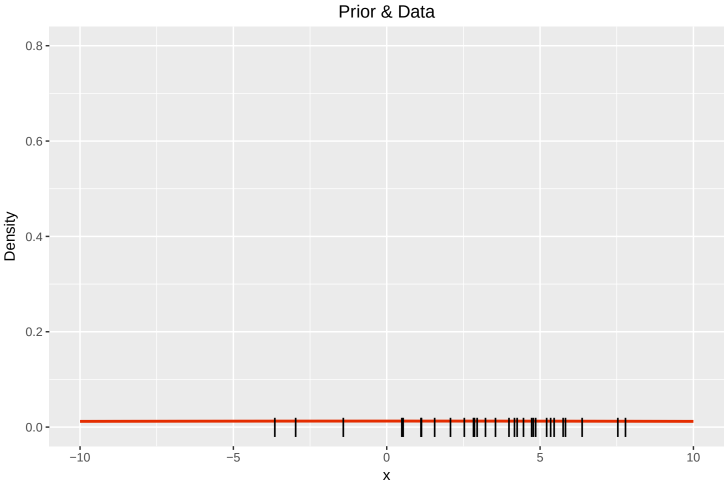
*Again, for the sake of notational simplicity, we will denote by $\pi(\theta|x)$ both the posterior distribution and the posterior density

Example of Bayesian inference: prior

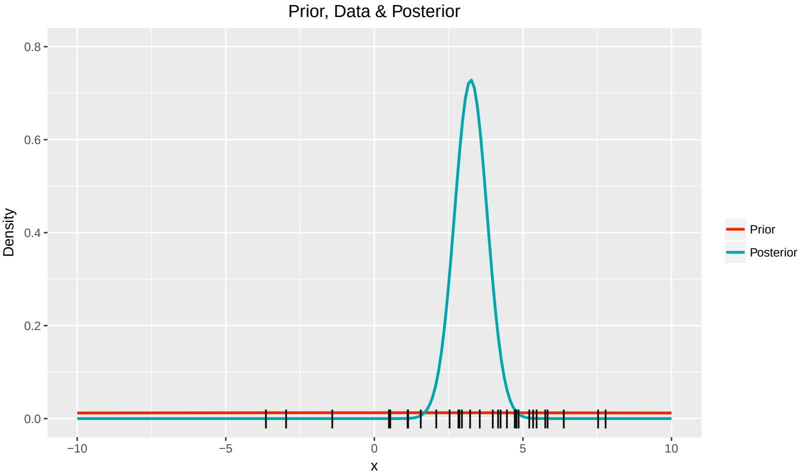
Example with a "flat" prior



Example of Bayesian inference: observe data



Example of Bayesian inference: update prior to obtain posterior



Challenges of Bayesian inference

Bayesian inference involves the following three main challenges:

1. Finding a good model (both prior and likelihood)
2. Calculating the posterior
3. Assessing the fit of the model

Comments on prior distribution

- The choice of the prior can be difficult and it can influence the result of the analysis (see, e.g., the base rate problem)
- The prior cannot depend on the data
- Prior information, e.g., from past analyses or from subject specific knowledge can be used to specify prior distributions
- If there exists no explicit prior knowledge about the parameter θ, the prior should be chosen as non-informative as possible

  As we will se later in the course, it turns out that this is easier said than done

Comparison of Bayesian and frequentist approach

- In the **Bayesian approach**, once a (appropriate) model is specified and after the data is observed, the way to do inference is determined: calculate the posterior*

- In the **frequentist approach**, one starts with any inference method (e.g., a likelihood-based method, any other optimization method, method of moments, etc.), and asks oneself what would happen if one repeated the procedure many times with the data changing (i.e., sampled again from the same model) each time.

  One then compares the properties of the different estimators and uses the procedure (often maximum likelihood) which has good properties such as consistency, a good convergence rate, low (asymptotic) variance etc.

---

*Depending on the application, calculating the posterior might still be difficult from a computational point of view. Also sometimes one might prefer a point estimate instead of an entire distribution, and one has to decide which point estimate one should use.

Focus of this course

This course mainly focuses on the following topics:

- **Basics of Bayesian statistics** (definitions, point estimation, testing, Bayes factor, credible sets, asymptotics)
- How do we **choose prior distributions** that have ”good” properties (non-informative, good frequentist properties, computational tractability, etc.)?
- How can we **calculate or approximate the posterior** distribution?

Calculation of posterior: an example

- Assume $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{N}(\theta, 1)$, with θ unknown. Thus the likelihood is

$$f(x_1, \ldots, x_n \mid \theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

- As prior for θ, we choose a $\mathcal{N}\left(\mu, \tau^2\right)$-distribution, that is

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2\right)$$

- Determination of posterior $\pi(\theta|x_1, \ldots, x_n)$: *See blackboard*
  - *Clicker question*

Prior predictive distribution

- The denominator

$$f(x) = \int \underbrace{\pi(\theta)f(x \mid \theta)}_{=\pi(x,\theta)} d\theta$$

  in the Bayes formula for the posterior is called the **prior predictive density** $f(x)$

- It is also called the **marginal likelihood or marginal density** since $f(x)$ is obtained by marginalizing the joint density $\pi(x, \theta)$ of (X, θ) over θ

Posterior predictive distribution

- The density of a future observation Y from the same model conditional on the data x is called the **posterior predictive density** $f(y \mid x)$
- By the law of total probability, this can be written as

$$\begin{aligned} f(y \mid x) &= \int f(y, \theta \mid x) d\theta \\ &= \int f(y \mid x, \theta) \pi(\theta | x) d\theta \\ &= \int f(y \mid \theta) \pi(\theta | x) d\theta, \end{aligned}$$

where we have assumed that given θ, Y is independent of X

Prior and posterior predictive distribution: an example

- In the previous example (normal means), we can easily calculate the prior and the posterior predictive distribution

  See blackboard

Interpretations of probability

Two interpretations of probability

Aleatoric uncertainty (from Latin “alea” = dice)

- Uncertainty about outcomes of repeatable events
- Probabilities can be understood as mathematical idealizations of long-run relative frequencies, which we call the frequentist interpretation of probability

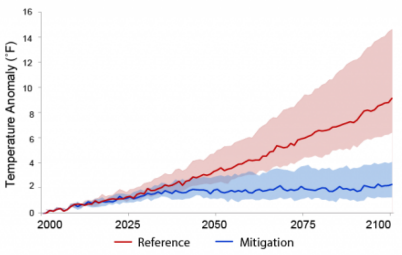
Epistemic uncertainty (from Greek “episteme” = knowledge)

- Uncertainty about unique events resulting from insufficient knowledge
- Probability statements are based on an evaluation of the available facts and information by an individual and thus become subjective degrees of belief

Examples of aleatoric and epistemic uncertainty

- **Aleatoric**: Throwing a dice $P(\text{dice} = 1)$

- **Epistemic**: Modeling climate change
  $P(\text{temperature increase in 2050} > 2^\circ)$



Clicker question