Introduction to Bayesian inference

Bayesian Computation 1 / 14

Guillaume Dehaene

Objectives

In this first course, we will:

1. Explain how the course is structured
2. Introduce Bayesian statistics
3. Recap Bayes' formula
4. Introduce the Bayesian approach to statistics
    1. Bayesian point estimation.
    2. Bayesian interval estimation.

1. Thesaurus

2. Introduction

3. Bayes' formula

4. Bayesian statistics

5. Examples

6. Conclusion

Thesaurus

Statistics

The object of statistics: **extract rational conclusions** from a dataset.

Classical approach to statistics: frequentist inference.

Key: probability statements are made *a-priori* of the data.

Alternatively, Bayesian inference. Possible arguments:

- Conditional statements are easy to interpret.
- Honest account of bias and modeling.
- Frequentist optimality.
- Subjective point of view on probability (probability as belief).
- Probability as rationality (philosophy / neuroscience).

Worst case: it's an important part of an "*homme du monde*" education in statistics.

Statistics

Like every field of statistics, Bayesian inference is supported on three pillars:

- **Probability theory:** my analyses are mathematically sound. (NB: not unambiguous: requires a definition of soundness).
- **Computation:** my computer can perform my analysis before the heatdeath of the universe.
- **Modeling:** my mathematical model of my data is a close match to the truth.

This course will focus on computation, but we will still discuss Theory and Modeling issues when appropriate.

Course objectives

The goal of the course is to give you **an overview of Bayesian inference:**

1. What it is and it isn't.
2. When to use it.
3. The strengths and the weaknesses.
4. **The computational methods** which make it **possible**.

You will not be experts but you should become fluent.

Structure of the course

Every week, the lecture will focus *mostly* on one topic / method.

The exercise session is focused on **programming the methods of the lecture**.

As we progress, we will shift towards spending the exercise sessions on your projects.

Going to the exercises sessions is **extremely strongly recommended**.

Feel free to program in any language.

Course requirements

The requirements for the course should be low:

- Basics of probability theory and of statistics.
- Basics of analysis (no measure theory).
- Familiarity with Markov Chains (I'll review them briefly).

If there is any point which flies over your head, **please let me know.**

Exam

The course will be graded on an **oral presentation** of a project that you do during the semester.

You will implement multiple Bayesian methods on a **single dataset** and present the results.

You should work on the project during the exercise sessions and **seek feedback often** on your work.

Class interaction

The #ThreeQuestionChallenge:

- Throughout the 14 weeks, I challenge every single one of you to intervene three times.
- Interventions can be big or small.
- **There are no such thing as stupid questions.**

I also ask many questions during the class:

- Most of these are easy !
- Don't worry, there are no wrong answers.

Questions?

Introduction

What is statistics

Statistics is the field that tries to answer the following question:

How can I draw **logical** / **rational conclusions** from random data?

It is thus a non-trivial **extension of logic.**

Both logic and statistics obey the same structure:

1. Start from postulates / premises:
    1. Logic example: definition of addition.
    2. Stats example: My data is Gaussian IID
2. Based on these postulates, draw a conclusion:
    1. Logic: Thus, addition is commutative.[1]
    2. Stats: Thus, $\bar{X}, S^2$ are *great* estimators of $\left(\mu, \sigma^2\right)$.

[1] Cf: Poincaré: *La Science et l'Hypothèse.*

The Frequentist approach

The most well-known statistical framework is the frequentist approach.

Its logic is a little tricky to follow.

1. Premise: my data is distributed according to ***Model***.
2. Furthermore, in ***Model***, the (random) output of ***calculation*** has good properties.
   Eg: estimator $\hat{\theta}$ is almost equal to the truth θ_0.
3. Conclusion: ***calculation*** performed on my data should also have those properties.

Critically, the good properties are a-priori of the data.
Indeed, the data is the only random quantity. Once the data is known, we cannot make probability statements anymore !

The Bayesian approach

The logic of Bayesian inference is quite different.

We assume that the **parameters** of the model **are random.**
We can then **apply Bayes' formula**:
Conditional on the data d, $\boldsymbol{\theta}$ is still random with distribution:

$$f\left(\boldsymbol{\theta}|d\right) = f\left(\boldsymbol{\theta}\right)\frac{f\left(d|\boldsymbol{\theta}\right)}{f\left(d\right)}$$

Any statement from Bayesian inference is **conditional / a posteriori on the data.**

The real reason

This difference is a **very small nitpick**.

In practice, both approaches lead to quite similar conclusions / estimations (**and they should**).

Some strengths of Bayesian inference:

- Interpretability.
- Separation of modeling and inference.
- **Frequentist** optimality results (Cramer-Rao bound).

We'll return to these much later.

A little bit of history

Like the name indicates, Bayes' formula was first proposed by Thomas Bayes (1750's).

Laplace (1800's) was another major influence in the field and can be considered the true father of Bayesian statistics (and all of statistics).

Bayesian inference was originally the dominating paradigm in statistics. This changed around 1850-1900 with the development of the formal axiomatization of probability theory (Kolmogorov) and the development of statistics as an independent field (Pearson K., Fisher, Neyman, Pearson E.)

There have been historically a million attempts to assert that one statistical philosophy is superior to the other. **These are extremely silly and they have exactly zero value.**

Bayes' formula

The historical problem

1. We push a black ball onto a billiard table.
2. We then push n white balls onto the table and record the number X of balls which stop before the black ball.

Assuming the final position of each ball is a **uniform RV**, how can we estimate the position of the black ball θ using X?

The historical problem

X is a binomial distribution:

$$X \sim \mathcal{B}\left(\frac{\theta}{L}, n\right)$$

A first **frequentist** answer to Bayes' problem consists in using the estimator:

$$\theta \approx \hat{\theta} = L\frac{X}{n}$$

The historical problem

This frequentist solution ignores the additional information contained in:

"We push a black ball onto a billiard table."

Thus, θ is itself random:

$$\theta \sim \mathcal{U}\left([0, L]\right)$$

We thus have a two-step process:

1. Generate θ randomly.
2. Generate X randomly (through generating the random positions of the white balls).

We want to estimate the result of the first step through observing the result of the second.

Bayes' formula

The solution to the problem is given by **Bayes' formula**:

$$f(\theta|x) = f(\theta)\frac{f(x|\theta)}{f(x)}$$

Let us give the correct interpretation of the various elements of this formula:

1. $\theta \to f(\theta)$ is the **prior** of the unobserved quantity θ.
   = what we knew before the data x was observed.
2. $\theta \to f(x|\theta)$ is the likelihood
   = a compatibility measure of θ (variable) with the data x (fixed).
3. $f(x)$ is a **normalization constant**.
   We need to compute it: $f(x) = \int f(\theta) f(x|\theta)$.
   We will refer to:

   $$\tilde{f}(\theta|x) = f(\theta) f(x|\theta)$$

   as the **unnormalized posterior**.

Bayes' formula

Critically, Bayes' formula says that, given our observation of x, θ **is still random** and has density $f(\theta|x)$.

Contrast this with the usual frequentist assumption that our parameters $\boldsymbol{\theta}$ are **fixed properties of the universe**.

As you might have noticed, Bayes' formula isn't statistics. It's an undeniable (basic) law of probability, but it is **not an undeniable law of statistics.**

Bayesian statistics

Model based statistics

What ingredients do we need to do Bayesian statistics:

1. Data $d = (x_1 \dots x_n)$.
2. A parametric model, i.e. a conditional distribution:

$$f\left(\mathcal{D} = d|\boldsymbol{\theta}\right)$$

   where $\boldsymbol{\theta}$ represents parameters of the model. If $\boldsymbol{\theta}$ is finite dimensional, then the model is **parametric**.

3. A prior distribution on the parameters $f\left(\boldsymbol{\theta}\right)$.

Note how similar this is to other **likelihood methods** (e.g. the Maximum Likelihood Estimator): we have only added the prior distribution on top of the conditional model.

The MLE

Let us briefly recall how the MLE approach works.

- The likelihood function:

$$\boldsymbol{\theta} \to f\left(d|\boldsymbol{\theta}\right)$$

  gives a measure of the **compatibility** of each value of $\boldsymbol{\theta}$ with the data.
- The MLE is the value of $\boldsymbol{\theta}$ with maximum compatibility:

$$\hat{\boldsymbol{\theta}}_{ML} = \max_{\boldsymbol{\theta}} \left[f\left(d|\boldsymbol{\theta}\right)\right]$$

- It has various great properties.

The goal of the MLE approach is to answer the most complicated question of classic statistics:

How do I choose my estimator?

A simple variant of the MLE consists in adding a regularization term, eg:

$$\hat{\boldsymbol{\theta}}'_{ML} = \max_{\boldsymbol{\theta}} \left[f\left(d|\boldsymbol{\theta}\right) - \lambda \left\|\boldsymbol{\theta}\right\|_2^2\right]$$

Bayesian statistics

With the addition of the prior $f(\boldsymbol{\theta})$, we have a joint distribution over $(\boldsymbol{\theta}, \mathcal{D})$.
This joint distribution is defined as a two-step process:

- First, pick the value of the parameter RV $\boldsymbol{\theta}$
- Then pick the value of the data, conditional on $\boldsymbol{\theta}$
- Finally, observe one realization of the data d

We can then compute the posterior distribution of $\boldsymbol{\theta}$ given the data we got, d, using Bayes' formula:

$$f(\boldsymbol{\theta}|d) = f(\boldsymbol{\theta}) \frac{f(d|\boldsymbol{\theta})}{f(d)}$$

Under our modelling assumptions, Bayes' formula is undeniably what we should do.
Of course, one could object to our modeling.

Bayesian statistics

Critically, under our modelling assumptions, $\boldsymbol{\theta}$ is still a random variable ! It has density $f\left(\boldsymbol{\theta}|d\right)$.

Once more, contrast this to the usual frequentist assumption that $\boldsymbol{\theta}$ is fixed for all eternity.

The posterior holds **all of the information we have about θ**

The key assumption of Bayesian statistics is that

Any question that we ask should be answered on the basis of the posterior distribution of θ.

The three questions of statistics

Let's see how we can answer the three questions of statistics using the posterior:

1. Estimation:
   1. Point estimation.
   2. Interval estimation.
2. Hypothesis testing.
3. Prediction.

What is the difference?

1. Estimation / Testing
   1. Trying to understand part of the distribution of X
2. Prediction:
   1. Understand all / as much as we can of the distribution of X

Estimation

Classical estimation problem: guess the value of $\boldsymbol{\theta}$.

In Bayesian statistics, this requires:

- Summarizing the posterior distribution $f\left(\boldsymbol{\theta}|\mathcal{D}=d\right)$.
- Into **a single scalar that represents the whole distribution**.

There are many possibilities for Bayesian estimation.

Estimation

Most important ones:

- Posterior mean: $\mathbb{E}\left(\boldsymbol{\theta} | \mathcal{D} = d\right)$
- Posterior median.
- Posterior mode (or Maximum A Posteriori value; MAP).

A general recipe:

1. Choose a **loss function** $L\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\right)$.
2. Minimize the **expected loss under the posterior:**

$$\hat{\boldsymbol{\theta}}_L = \mathbf{argmin}_{\hat{\boldsymbol{\theta}}} \left[\mathbb{E}\left(L\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\right) | \mathcal{D} = d\right)\right]$$

Interval estimation

Point estimates are a **very poor way** to solve the estimation question because there is no notion of precision.

Frequentist answer: Confidence Interval

Bayesian answer: **Credible** Interval:

$$I \text{ such that } \mathbb{P}\left(\boldsymbol{\theta} \in I | \mathcal{D} = d\right) = 1 - \alpha$$

There are infinitely many possible choices. Usually, we either:

- Take the interval with minimal length.
- Or one centered around an interesting estimator

Interval estimation

For example, if the posterior is a p-dimensional Gaussian or almost Gaussian:

$$\boldsymbol{\theta}|\mathcal{D} = d \sim \mathcal{N}\left(\boldsymbol{\mu}_{post}, \Sigma_{post}\right)$$

then, we can construct a confidence zone that is egg-shaped:

$$\left(\boldsymbol{\theta} - \boldsymbol{\mu}_{post}\right)^T \left[\Sigma_{post}\right]^{-1} \left(\boldsymbol{\theta} - \boldsymbol{\mu}_{post}\right) \leq q_\alpha\left(\chi^2_p\right)$$

Examples

Medecine: is Alice sick?

As a simple example, consider the following problem:

- Alice is at the doctor showing possible signs of the Cold.
- The doctor does a simple blood test which comes back positive.
- Is Alice sick?

Bayesian answer:

- Prior probability of a patient having a cold: 0.10.
- Probability of false positive (confidence): 0.20.
- Probability of true positive (power): 1.
- Posterior probability of Alice being sick?

Medecine: is Alice sick?

Key lessons:

- Confidence statements do not translate to credible statements:
  - Confidence and Power of test can't be interpreted as posterior probabilities.
- Posterior probability gives us naturally a notion of certainty.
- Posterior probability can be recombined with new information.

Limits:

- Where did prior probability come from?
  - It seems legitimate to have some prior information.
    - In summer: no Cold epidemic. $p \approx 0$.
    - In winter: Cold epidemic. $p \gg 0$.
  - But how do we get a value?
    - Fraction of people that are sick at a given point in time?
    - Additional information: Country, Town, Demographic information?

Grade difference after tutoring

We want to test the effect of tutoring on student grades (in $[0, 100]$).

We construct two groups of n students. Group 1 receives 1-on-1 tutoring while Group 2 does normal supervised exercises sessions.
Is the expense worth it?

Data: $X_1 \dots X_n$ and $Y_1 \dots Y_n$.

Conditional model:

$$X_i \sim \mathcal{N}\left(\mu_X, (10)^2\right)$$
$$Y_i \sim \mathcal{N}\left(\mu_Y, (10)^2\right)$$

Prior model:

$$\mu \overset{IID}{\sim} \mathcal{N}\left(60, (10)^2\right)$$

encoding the fact that I expect the average grade of students to not deviate too much from the typical value 60.

Grade difference after tutoring

Posterior is independent:

$$\mu_X | x_1 \dots x_n \sim \mathcal{N}\left(\frac{n\bar{x}+60}{n+1}, \frac{(10)^2}{n+1}\right)$$

$$\mu_Y | y_1 \dots y_n \sim \mathcal{N}\left(\frac{n\bar{y}+60}{n+1}, \frac{(10)^2}{n+1}\right)$$

Posterior difference of $\mu_X - \mu_Y$ is marginally Gaussian:

$$\mu_X - \mu_Y | d \sim \mathcal{N}\left(\frac{n}{n+1}\left(\bar{x}-\bar{y}\right), \frac{2}{n+1}\left(10\right)^2\right)$$

Estimation of the difference of means?

Grade difference after tutoring

Estimation:

- Posterior mean, median and mode are identical:

$$\widehat{\mu_X - \mu_Y} = \frac{n}{n+1}\left(\bar{x} - \bar{y}\right)$$

- 0.95 credible region:

$$I = \frac{n}{n+1}\left(\bar{x} - \bar{y}\right) \pm 2 * 10\sqrt{\frac{2}{n+1}}$$

Grade difference after tutoring

Limits:

- Model is extremely unrealistic:
    - Variance assumed known.
    - Gaussian conditional model when grade is bounded.
    - Gaussian prior model when mean-grade is bounded too.

  (it makes my life simple: explicit posterior)
- Choosing the right model?
- Justify prior information??

Conclusion

The story so far

Bayesian inference is very close to MLE methods.
It requires one additional ingredient: the prior distribution $f(\boldsymbol{\theta})$.

The posterior distribution is sometimes analytically tractable, most often not.
Most of the course will focus on the computational tools which can deal with this issue.

So far, we have given **no good argument** in favor of Bayesian inference.
These will come later in the course.

Bayes formula

If $\boldsymbol{\theta}$ and $\mathcal{D}$ are two random variables defined on the same probability space Then, observing $\mathcal{D} = d$ modifies the probability of $\boldsymbol{\theta}$.

- Event based formulation:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)}{\mathbb{P}(B)}\mathbb{P}(A)$$

- If $f(\boldsymbol{\theta})$ is a density (with respect to some base measure $d\nu_\theta$) and, for every $\boldsymbol{\theta}$, $f(\mathcal{D} = d|\boldsymbol{\theta})$ is a density (with respect to some base measure $d\nu_{\mathcal{D}}$),
  Then the posterior is a density (with respect to the base measure $d\nu_\theta$):

$$f(\boldsymbol{\theta}|\mathcal{D} = d) = f(\boldsymbol{\theta})\frac{f(\mathcal{D} = d|\boldsymbol{\theta})}{f(\mathcal{D} = d)}$$

Define the unnormalized posterior:

$$\tilde{f}(\boldsymbol{\theta}|\mathcal{D} = d) = f(\boldsymbol{\theta})f(\mathcal{D} = d|\boldsymbol{\theta}) = f(\mathcal{D} = d \;\&\; \boldsymbol{\theta})$$

Bayesian inference

Bayesian statistical inference:

- Model a dataset $\mathcal{D} = d$ with:
  - A conditional model: $f\left(\mathcal{D} = d | \boldsymbol{\theta}\right)$
  - A prior model over all parameters: $f\left(\boldsymbol{\theta}\right)$
- Perform inference:
  - Compute the posterior distribution:

$$f\left(\boldsymbol{\theta} | \mathcal{D} = d\right) = f\left(\boldsymbol{\theta}\right) \frac{f\left(\mathcal{D} = d | \boldsymbol{\theta}\right)}{f\left(\mathcal{D} = d\right)}$$

  - Squeeze out the answer from the posterior: depends on what we want.

Bayesian estimation

If we want to estimate the unknown parameters $\boldsymbol{\theta}$ given a posterior:

- Point estimates:
  - Posterior mode (MAP)
  - Posterior Mean
  - Posterior median
  - General: minimize loss:

$$\hat{\boldsymbol{\theta}}_L = \mathrm{argmin}_{\hat{\boldsymbol{\theta}}} \left[\mathbb{E} \left(L \left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \right) | \mathcal{D} = d \right) \right]$$

- Interval estimates:
  - Credible interval: I_α such that:

$$\mathbb{P} \left(\boldsymbol{\theta} \in I_\alpha | \mathcal{D} = d \right) = 1 - \alpha$$

  - General: minimize loss??