Lecture 1. General concepts, formalism, coin-flipping

Introduction to Bayesian Statistical Learning

18.03.2024 Instructors: Alina Bazarova, Oleg Filatov. Technical issues: Alexandre Strube

General concepts

- Bayesian approach to inference is about **preserving uncertainty**

What does that mean? Assume we need to estimate a certain parameter

General concepts

- Bayesian approach to inference is about **preserving uncertainty**

What does that mean? Assume we need to estimate a certain parameter

<u>Classical (frequentist) statistics output</u>: point estimate, we can compute the confidence interval given our assumptions **(reducing uncertainty)**

General concepts

- Bayesian approach to inference is about **preserving uncertainty**

What does that mean? Assume we need to estimate a certain parameter

<u>Classical (frequentist) statistics output</u>: point estimate, we can compute the confidence interval given our assumptions **(reducing uncertainty)**

<u>Bayesian statistics output</u>: a distribution. We do not claim that we have found an exact value of our parameter. However we **quantify the probability** with which our parameter take any value

General concepts

- Bayesian approach to inference is about **preserving uncertainty**

What does that mean? Assume we need to estimate a certain parameter

Classical (frequentist) statistics output: point estimate, we can compute the confidence interval given our assumptions **(reducing uncertainty)**

Bayesian statistics output: a distribution. We do not claim that we have found an exact value of our parameter. However we **quantify the probability** with which our parameter take any value

- Bayesian approach is **based on observed data** and estimates are updated as more data arrive (hence usage of conditional probability)

General concepts

- Bayesian approach to inference is about **preserving uncertainty**

What does that mean? Assume we need to estimate a certain parameter

Classical (frequentist) statistics output: point estimate, we can compute the confidence interval given our assumptions **(reducing uncertainty)**

Bayesian statistics output: a distribution. We do not claim that we have found an exact value of our parameter. However we **quantify the probability** with which our parameter take any value

- Bayesian approach is **based on observed data** and estimates are updated as more data arrive (hence usage of conditional probability)

- Therefore, **more flexibility,** possibly **more information**

General concepts

- Bayesian approach to inference is about **preserving uncertainty**

What does that mean? Assume we need to estimate a certain parameter

Classical (frequentist) statistics output: point estimate, we can compute the confidence interval given our assumptions **(reducing uncertainty)**

Bayesian statistics output: a distribution. We do not claim that we have found an exact value of our parameter. However we **quantify the probability** with which our parameter take any value

- Bayesian approach is **based on observed data** and estimates are updated as more data arrive (hence usage of conditional probability)

- Therefore, **more flexibility,** possibly **more information**

- Does one have to pick a side (Classical or Bayesian)? No! But we will talk about it later…

Typical use-cases of Bayesian statistics

- Situations when new evidence (data) may significantly influence model parameters and thereby require immediate actions.
- Situations where one is interested in the degree of uncertainty of the results (which you get automatically when using Bayesian approach)

Example:

COVID-19 pandemic. Non-pharmaceutical interventions: lockdowns of various degrees, increased testing - all lead to changes in model parameters such as **reproduction number**, **infection rate** etc. Same as vaccine and drug development which came in significantly later.

Such model would be **data-driven** and have **immediate implications** for public health.

Formalism. Conditional probability and Bayes rule (theorem).

Conditional probability $P(A \mid X)$: the probability of event A occurring given that event X **has already occurred**.

Bayes' theorem provides a way to **revise existing predictions** or theories (update probabilities) given **new or additional evidence**

Formalism. Conditional probability and Bayes rule (theorem).

Conditional probability $P(A\,|\,X)$: the probability of event A occurring given that event X **has already occurred**.

Bayes' theorem provides a way to **revise existing predictions** or theories (update probabilities) given **new or additional evidence**

$$P(A\,|\,X) = \frac{P(A,X)}{P(X)} = \frac{P(A)P(X\,|\,A)}{P(X)}$$

Where usually A represents parameters of the model, X represents the data

Formalism. Conditional probability and Bayes rule (theorem).

Conditional probability $P(A \mid X)$: the probability of event A occurring given that event X **has already occurred**.

Bayes' theorem provides a way to **revise existing predictions** or theories (update probabilities) given **new or additional evidence**

$$P(A \mid X) = \frac{P(A, X)}{P(X)} = \frac{P(A)P(X \mid A)}{P(X)}$$

Where usually A represents parameters of the model, X represents the data

$$P(A_j \mid X) = \frac{P(A_j)P(X \mid A_j)}{\sum_{i=1}^{k} P(A_i)P(X \mid A_i)}$$

A full version, where $\{A_1, \ldots, A_k\}$ is a partition of A and $j \in \{1, \ldots, k\}$

Formalism. Conditional probability and Bayes rule (theorem).

Conditional probability $P(A \mid X)$: the probability of event A occurring given that event X **has already occurred**.

Bayes' theorem provides a way to **revise existing predictions** or theories (update probabilities) given **new or additional evidence**

$$P(A \mid X) = \frac{P(A, X)}{P(X)} = \frac{P(A)P(X \mid A)}{P(X)}$$

Where usually A represents parameters of the model, X represents the data

$$P(A_j \mid X) = \frac{P(A_j)P(X \mid A_j)}{\sum_{i=1}^{k} P(A_i)P(X \mid A_i)}$$

A full version, where $\{A_1, \ldots, A_k\}$ is a partition of A and $j \in \{1, \ldots, k\}$

$$posterior = \frac{prior \times likelihood}{evidence}$$

Reformulated in Bayesian language

Continuous space

Let $X, Y \in \mathbb{R}$, $p_X(x)$ is probability density of X (and respective of Y),

Continuous space

Let $X, Y \in \mathbb{R}$, $p_X(x)$ is probability density of X (and respective of Y),

namely $p_X(x) > 0$ and $\int_{\mathbb{R}} p_X(x)dx = 1$ $\qquad P(X \in A) = \int_A p_X(x)dx$

Continuous space

Let $X, Y \in \mathbb{R}$, $p_X(x)$ is probability density of X (and respective of Y),

namely $p_X(x) > 0$ and $\displaystyle\int_{\mathbb{R}} p_X(x)dx = 1$ $\qquad P(X \in A) = \displaystyle\int_A p_X(x)dx$

Then, $p_{X|Y}(x\,|\,y) = \dfrac{p_{X,Y}(x, y)}{p_Y(y)}$, $\qquad p_Y(y) = \displaystyle\int_{\mathbb{R}} p_{X,Y}(x, y)dx$

Continuous space

Let $X, Y \in \mathbb{R}$, $p_X(x)$ is probability density of X (and respective of Y),

namely $p_X(x) > 0$ and $\int_{\mathbb{R}} p_X(x)dx = 1$ $\qquad P(X \in A) = \int_A p_X(x)dx$

Then, $p_{X|Y}(x\,|\,y) = \dfrac{p_{X,Y}(x, y)}{p_Y(y)}$, $\qquad p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(x, y)dx$

$$p_{Y|X}(y\,|\,x) = \frac{p_Y(y)p_{X|Y}(x\,|\,y)}{p_X(x)} = \frac{p_Y(y)p_{X|Y}(x\,|\,y)}{\int_{\mathbb{R}} p_X(x)p_{Y|X}(y\,|\,x)dy}$$

continuous Bayes rule

Continuous space

Let $X, Y \in \mathbb{R}$, $p_X(x)$ is probability density of X (and respective of Y),

namely $p_X(x) > 0$ and $\int_{\mathbb{R}} p_X(x)dx = 1$ $\quad P(X \in A) = \int_A p_X(x)dx$

Then, $p_{X|Y}(x\,|\,y) = \dfrac{p_{X,Y}(x, y)}{p_Y(y)}$, $\quad p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(x, y)dx$

$$p_{Y|X}(y\,|\,x) = \frac{p_Y(y)p_{X|Y}(x\,|\,y)}{p_X(x)} = \frac{p_Y(y)p_{X|Y}(x\,|\,y)}{\int_{\mathbb{R}} p_Y(y)p_{X|Y}(x\,|\,y)dy}, \quad p(y\,|\,x) \propto p(x\,|\,y)p(y)$$

continuous Bayes rule

Possible issues with $\frac{p_Y(y)p_{X|Y}(x|y)}{\int_{\mathbb{R}} p_Y(y)p_{X|Y}(x|y)dy}$

- Likelihood $p(x|y)$ might be very complicated
- The integral in the denominator is often intractable, hence computational methods (MCMC, Variational Bayes etc.)

Note:

- $p(x|y)$ is our model of the data: data-generating distribution
- $p(y)$ is what we think about the parameters of the model $a\ priori$ (prior)

Example: Bayesian vs Frequentist murder trial

Assume you are (hopefully falsely) accused of a murder and have to face a jury in a misfortunate country where the guilt presumed over innocence (null hypothesis is that one is guilty).

The CCTV footage indicates that you were in the same house as the victim on the night of a murder. There are two types of trial:

1. **Frequentist trial.** The jurors specify a model based on the previous trials: if you commit the murder, 30% of the time you would have been seen by the CCTV. Since the probability $P(security\ camera\ footage \mid guilt) > 0.05$, you are declared guilty.

2. **Bayesian trial.** The jury first are looking at the evidence, such as absence of previous violent conduct etc. and based on that assign a prior probability of $\frac{1}{1000}$. They compute probability according to Bayes rule

$$P(guilt \mid security\ camera\ footage) = \frac{P(security\ camera\ footage \mid guilt) P(guilt)}{P(security\ camera\ footage)} = \frac{0.3 \cdot 0.001}{0.3 \cdot 0.001 + 0.3 \cdot 0.999} = 0.001 < 0.05$$

And therefore you are declared innocent.

Coin-flipping example

Suppose, that you are unsure about the probability of heads in a coin flip (spoiler alert: usually it's 50%).

You believe there is some true underlying ratio, call it p , but have no prior opinion on what p might be.

We begin to flip a coin, and record the observations: either H or T. This is our observed data.

Question to ask: how will our inference change as we observe more and more data?

$P(H = s) = \binom{n}{s} p^s (1-p)^{n-s}$**,** prior is uniform (constant density function $= 1$), s and n are our data, p is the parameter

Coin-flipping example

Suppose, that you are unsure about the probability of heads in a coin flip (spoiler alert: usually it's 50%).

You believe there is some true underlying ratio, call it p , but have no prior opinion on what p might be.

We begin to flip a coin, and record the observations: either H or T. This is our observed data.

Question to ask: how will our inference change as we observe more and more data?

$P(H = s) = \binom{n}{s} p^s (1-p)^{n-s}$, prior is uniform (constant density function $= 1$), s and n are our data, p is the parameter

$$P(p = x \mid s, n) = \frac{P(s, n \mid x)P(x)}{\int P(s, n \mid y)P(y)dy}$$

Coin-flipping example

Suppose, that you are unsure about the probability of heads in a coin flip (spoiler alert: it's 50%). You believe there is some true underlying ratio, call it p , but have no prior opinion on what p might be.

We begin to flip a coin, and record the observations: either H or T. This is our observed data.

Question to ask: how will our inference change as we observe more and more data?

$P(H = s) = \binom{n}{s} p^s (1-p)^{n-s}$**,** prior is uniform (constant density function $= 1$), s and n are our data, p is the parameter

$$P(p = x \mid s, n) = \frac{P(s, n \mid x) P(x)}{\int P(s, n \mid y) P(y) dy} = \frac{\binom{n}{s} x^s (1-x)^{n-s}}{\binom{n}{s} \int y^s (1-y)^{n-s} dy}$$

Coin-flipping example

Suppose, that you are unsure about the probability of heads in a coin flip (spoiler alert: it's 50%). You believe there is some true underlying ratio, call it p , but have no prior opinion on what p might be.

We begin to flip a coin, and record the observations: either H or T. This is our observed data.

Question to ask: how will our inference change as we observe more and more data?

$P(H = s) = \binom{n}{s} p^s (1-p)^{n-s}$**,** prior is uniform (constant density function $=1$), s and n are our data, p is the parameter

$$P(p = x \mid s, n) = \frac{P(s, n \mid x)P(x)}{\int P(s, n \mid y)P(y)dy} = \frac{\binom{n}{s} x^s (1-x)^{n-s}}{\binom{n}{s} \int y^s (1-y)^{n-s} dy} = \frac{x^s (1-x)^{n-s}}{B(s, n-s)}$$

Coin-flipping example

Suppose, that you are unsure about the probability of heads in a coin flip (spoiler alert: it's 50%). You believe there is some true underlying ratio, call it p , but have no prior opinion on what p might be.

We begin to flip a coin, and record the observations: either H or T. This is our observed data.

Question to ask: how will our inference change as we observe more and more data?

$P(H = s) = \binom{n}{s} p^s (1-p)^{n-s}$**,** prior is uniform (constant density function $= 1$), s and n are our data, p is the parameter

$$P(p = x \mid s, n) = \frac{P(s, n \mid x)P(x)}{\int P(s, n \mid y)P(y)dy} = \frac{\binom{n}{s} x^s (1-x)^{n-s}}{\binom{n}{s} \int y^s (1-y)^{n-s} dy} = \frac{x^s (1-x)^{n-s}}{B(s, n-s)} \sim Beta(s+1, n-s+1)$$

Jupyter notebook Lecture_1_examples: coin flipping example

Some implications I

If $p \mid s, n \sim Beta(s+1, n-s+1)$, which is $Ep = \dfrac{s+1}{n+2} \approx \dfrac{s}{n}$ for large n, similarly

$$Var(p) = \frac{(s+1)(n-s+1)}{(n+3)(n+2)^2} \approx \frac{s(n-s)}{n^3}$$

Some implications I

If $p \mid s, n \sim Beta(s+1, n-s+1)$, which is $Ep = \dfrac{s+1}{n+2} \approx \dfrac{s}{n}$ for large n, similarly

$$Var(p) = \frac{(s+1)(n-s+1)}{(n+3)(n+2)^2} \approx \frac{s(n-s)}{n^3}$$

In classical statistics one often estimates $p = \dfrac{\#successes}{\#experiments} = \dfrac{s}{n}$, the variance estimator then would be $\dfrac{p(1-p)}{n}$, however in case p is unknown $\dfrac{s(n-s)}{n^3}$!

Some implications I

If $p\,|\,s, n \sim Beta(s+1, n-s+1)$, which is $Ep = \frac{s+1}{n+2} \approx \frac{s}{n}$ for large n, similarly

$$Var(p) = \frac{(s+1)(n-s+1)}{(n+3)(n+2)^2} \approx \frac{s(n-s)}{n^3}$$

In classical statistics one often estimates $p = \frac{\#successes}{\#experiments} = \frac{s}{n}$, the variance estimator then would be $\frac{p(1-p)}{n}$, however in case p is unknown $\frac{s(n-s)}{n}^3$!

Punchline: if sample is large enough there is no difference whether to use Bayesian or frequentist approach!

Some implications II. Conjugate priors

In the coin-flipping example the posterior **matched the well-known distribution** - that was nice!

Some implications II. Conjugate priors

In the coin-flipping example the posterior **matched the well-known distribution** - that was nice!

Our prior was uniform. But what if now we use a $Beta$ distributed prior?

Some implications II. Conjugate priors

In the coin-flipping example the posterior **matched the well-known distribution** - that was nice!

Our prior was uniform. But what if now we use a $Beta$ distributed prior?

$$P(p = x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

Some implications II. Conjugate priors

In the coin-flipping example the posterior **matched the well-known distribution** - that was nice!

Our prior was uniform. But what if now we use a $Beta$ distributed prior?

$P(p = x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ and hence the posterior $P(p = x \,|\, n, s) \propto P(p = x)\binom{n}{s}x^s(1-x)^{n-s}$

Some implications II. Conjugate priors

In the coin-flipping example the posterior **matched the well-known distribution** - that was nice!

Our prior was uniform. But what if now we use a $Beta$ distributed prior?

$P(p = x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ and hence the posterior $P(p = x \mid n, s) \propto P(p = x)\binom{n}{s}x^s(1-x)^{n-s}$

$P(p = x \mid n, s) \propto \binom{n}{s}x^{s+\alpha-1}(1-x)^{n-s+\beta-1}$ $\propto$ pdf of $Beta(\alpha + s, \beta + n - s)$

Some implications II. Conjugate priors

In the coin-flipping example the posterior **matched the well-known distribution** - that was nice!

Our prior was uniform. But what if now we use a $Beta$ distributed prior?

$P(p = x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ and hence the posterior $P(p = x \mid n, s) \propto P(p = x)\binom{n}{s}x^s(1-x)^{n-s}$

$$P(p = x \mid n, s) \propto \binom{n}{s}x^{s+\alpha-1}(1-x)^{n-s+\beta-1} \propto \text{pdf of } Beta(\alpha + s, \beta + n - s)$$

The prior coming from the same distribution family as prior is called **conjugate prior.**

Some implications II. Conjugate priors

In the coin-flipping example the posterior **matched the well-known distribution** - that was nice!

Our prior was uniform. But what if now we use a $Beta$ distributed prior?

$P(p = x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$ and hence the posterior $P(p = x \mid n, s) \propto P(p = x)\binom{n}{s}x^s(1-x)^{n-s}$

$$P(p = x \mid n, s) \propto \binom{n}{s}x^{s+\alpha-1}(1-x)^{n-s+\beta-1} \propto \text{pdf of } Beta(\alpha + s, \beta + n - s)$$

The prior coming from the same distribution family as prior is called **conjugate prior.**

This is very useful both for numerical and analytical methods.

A comprehensive list of **pairs likelihood - conjugate prior** https://en.wikipedia.org/wiki/Conjugate_prior

Jupyter notebook 1 - play around with a prior in a coin-flipping example, look how posterior changes

Continuous distributions

A typical (and somewhat simplified) question: what is the parameter of the distribution based on the data?

Example: exponential distribution with pdf $p_X(x \mid \lambda) = \lambda e^{-\lambda x}$, where X is our r.v.

What can we say about λ if we can only observe values of X?

Continuous distributions

A typical (and somewhat simplified) question: what is the parameter of the distribution based on the data?

Example: exponential distribution with pdf $p_X(x \mid \lambda) = \lambda e^{-\lambda x}$, where X is our r.v.

What can we say about λ if we can only observe values of X?

Bayesian inference: rather than guessing λ exactly we try assigning a probability distribution to it, hence our **inference provides confidence intervals automatically**.

Jupyter notebook Lecture_1_examples: example with text message data