MAST90125: Bayesian Statistical learning

Lecture 1: Bayesian basics

Prepared by Feng Liu and Guoqi Qian

Bayes Theorem

The fundamental basis of Bayesian analysis is Bayes Theorem,

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

in the discrete case, where A and B are events.

Bayes Theorem

The fundamental basis of Bayesian analysis is Bayes Theorem,

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

in the discrete case, where A and B are events. For the continuous case,

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

Bayes Theorem

The fundamental basis of Bayesian analysis is Bayes Theorem,

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

in the discrete case, where A and B are events. For the continuous case,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}, \quad \int_a p(a|B)da = \frac{\int_a p(a,B)da}{p(B)} = 1$$

where $p(\cdot)$ denotes probability density, and A and B are random variables.

Bayes Theorem

The fundamental basis of Bayesian analysis is Bayes Theorem,

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

in the discrete case, where A and B are events. For the continuous case,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}, \quad \int_a p(a|B)da = \frac{\int_a p(a,B)da}{p(B)} = 1$$

where $p(\cdot)$ denotes probability density, and A and B are random variables.

- To understand this, it is necessary to understand probability rules/symbolism. A summary of these can be found on the following slides.

Probability: Symbols

- $\cap$ means **intersection** or equivalently **and**.
  For example $\Pr(A \cap B)$ is the probability of event A and event B occurring together. When thinking of parameters rather than events, we tend to think of joint distributions $\Pr(A, B)$.
- $\cup$ means **union** or equivalently **or**.
  For example $\Pr(A \cup B)$ is the probability of at least one of the events A and B occurring.
- $\bar{\ }$ means **complement** or equivalently **not**.
  For example $\Pr(\bar{A})$ is the probability that event A does not occur.
- $|$ means **conditional on** or equivalently **given**.
  For example $\Pr(A|B)$ is the probability of event A occurring given event B has already happened.

Probability: Rules

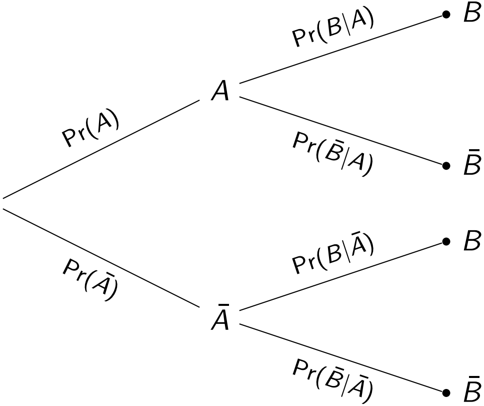
- **A probability of an event must lie between 0 and 1. The probability of the union of all possible events is 1.**

- The Addition Rule: $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- The Multiplication Rule: $\Pr(A \cap B) = \Pr(B|A)\Pr(A)$
  which if A, B are independent reduces to $\Pr(A \cap B) = \Pr(B)\Pr(A)$

- **When considering multiple events, marginal probability of one event corresponds to**
  $\Pr(A) = \sum_i \Pr(A \cap B_i)$, if B_i's are disjoint events constituting the sample space
  and $p(A) = \int p(A, B)dB$, if A, B are continuous.

Tree diagram

In your undergraduate courses, you may have encountered tree diagrams.



$\Pr(A)$

$\Pr(\bar{A})$

A

$\bar{A}$

$\Pr(B|A)$

$\Pr(\bar{B}|A)$

$\Pr(B|\bar{A})$

$\Pr(\bar{B}|\bar{A})$

B

$\bar{B}$

B

$\bar{B}$

$$\Pr(A) \times \Pr(B|A) = \Pr(A \cap B)$$

$$\Pr(A) \times \Pr(\bar{B}|A) = \Pr(A \cap \bar{B})$$

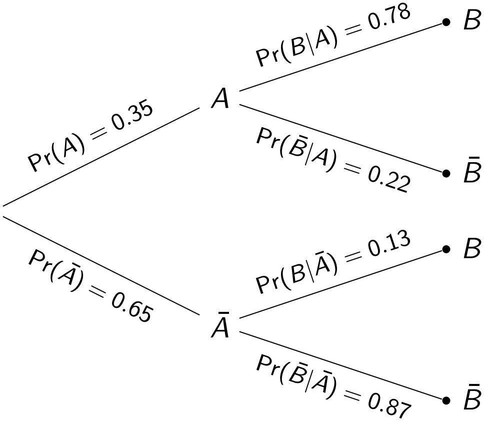
$$\Pr(\bar{A}) \times \Pr(B|\bar{A}) = \Pr(\bar{A} \cap B)$$

$$\Pr(\bar{A}) \times \Pr(\bar{B}|\bar{A}) = \Pr(\bar{A} \cap \bar{B})$$

Example: Diagnostic testing and tree diagrams

- Tree diagrams are often used to explain diagnostic testing.
- In the following example, let A represent disease and B testing positive for the disease.
- You are told the probability of disease in the population is 0.35 and the probability of testing positive for disease is 0.78 if disease is present and 0.13 if disease is not present.
- Based on this information, please fill the tree diagram on the previous slide, marginalise out the event A, and reverse the conditioning.

Filling in the tree diagram



$$\Pr(A) \times \Pr(B|A) = \Pr(A \cap B) = 0.2730$$

$$\Pr(A) \times \Pr(\bar{B}|A) = \Pr(A \cap \bar{B}) = 0.0770$$

$$\Pr(\bar{A}) \times \Pr(B|\bar{A}) = \Pr(\bar{A} \cap B) = 0.0845$$

$$\Pr(\bar{A}) \times \Pr(\bar{B}|\bar{A}) = \Pr(\bar{A} \cap \bar{B}) = 0.5655$$

Additional probabilities

- Some additional probabilities of interest are (to 4 d.p.):
  - $\Pr(B) = \Pr(A \cap B) + \Pr(\bar{A} \cap B) = 0.2730 + 0.0845 = 0.3575$
  - Positive predictive value:
    $\Pr(A|B) = \Pr(A \cap B) / \Pr(B) = 0.2730/0.3575 = 0.7636$
  - Negative predictive value:
    $\Pr(\bar{A}|\bar{B}) = \Pr(\bar{A} \cap \bar{B}) / \Pr(\bar{B}) = \Pr(\bar{A} \cap \bar{B})/(1 - \Pr(B)) = 0.5655/0.6425 = 0.8802.$

Additional probabilities

- Some additional probabilities of interest are (to 4 d.p.):
  - $\Pr(B) = \Pr(A \cap B) + \Pr(\bar{A} \cap B) = 0.2730 + 0.0845 = 0.3575$
  - Positive predictive value:
    $\Pr(A|B) = \Pr(A \cap B) / \Pr(B) = 0.2730/0.3575 = 0.7636$
  - Negative predictive value:
    $\Pr(\bar{A}|\bar{B}) = \Pr(\bar{A} \cap \bar{B}) / \Pr(\bar{B}) = \Pr(\bar{A} \cap \bar{B})/(1 - \Pr(B)) = 0.5655/0.6425 = 0.8802.$
- But why is this of interest for Bayesian analysis?

Returning to Bayes Theorem

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

- You may already be familiar with the following categorisation of the components of Bayes Theorem:
    - $\Pr(A)$ is the prior.
    - $\Pr(B|A)$ is the likelihood.
    - $\Pr(A|B)$ is the posterior.

Returning to Bayes Theorem

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

- You may already be familiar with the following categorisation of the components of Bayes Theorem:
  - $\Pr(A)$ is the prior.
  - $\Pr(B|A)$ is the likelihood.
  - $\Pr(A|B)$ is the posterior.
- So returning to the diagnostic testing example, how would we interpret what we did in the context of a Bayesian analysis?

Rethinking the example

- The probability of disease in the population, $\Pr(A) = 0.35$, is an example of a prior.

Rethinking the example

- The probability of disease in the population, $\Pr(A) = 0.35$, is an example of a prior.

- The probability of testing positive for disease is 0.78 if disease is present and 0.13 if disease is not present correspond to ‘likelihood’ statements.

Rethinking the example

- The probability of disease in the population, $\Pr(A) = 0.35$, is an example of a prior.
- The probability of testing positive for disease is 0.78 if disease is present and 0.13 if disease is not present correspond to ‘likelihood’ statements.
- The positive and negative predictive values are examples of posterior probabilities.

Rethinking the example

- The probability of disease in the population, $\Pr(A) = 0.35$, is an example of a prior.
- The probability of testing positive for disease is 0.78 if disease is present and 0.13 if disease is not present correspond to 'likelihood' statements.
- The positive and negative predictive values are examples of posterior probabilities.
- However what might be odd about the current example?

Rethinking the example

- The probability of disease in the population, $\Pr(A) = 0.35$, is an example of a prior.
- The probability of testing positive for disease is 0.78 if disease is present and 0.13 if disease is not present correspond to 'likelihood' statements.
- The positive and negative predictive values are examples of posterior probabilities.
- However what might be odd about the current example?
  - There is no discussion of data. Now assume the tree diagram was constructed using results of a case-control study.

Process of Bayesian data analysis, Step one

- Set up a full probability model. That is a probability model for both data and parameters.
  - In the tree diagram example, this comprises specifying the probability of disease (A) the parameter of interest.
    - As this is a case control study, the data does not provide information about $\Pr(A)$, so the choice of $\Pr(A)$ is subjective (the prior).

    It also comprises specifying a distribution for observed data (the generative model).
    - In the tree diagram example, this would be $y_{B|A} \sim \text{Bin}(n_A, \Pr(B|A))$ and $y_{B|\bar{A}} \sim \text{Bin}(n_{\bar{A}}, \Pr(B|\bar{A}))$, where $y_{B|A}$ and $y_{B|\bar{A}}$ are counts of positive cases and $n_A, n_{\bar{A}}$ are the number tested for each group.

Process of Bayesian data analysis, Step two

- Regarding the observed data.
  - In the tree diagram example, this comprises calculating the posterior probabilities.
    - Note that in this example, we have treated the observed data purely as point estimates, and ignored the uncertainty in the count.
    - Taking into account this uncertainty in this example (the likelihood).

Process of Bayesian data analysis, Step three

- Evaluate the fit of the model and the implications of the posterior distribution.
  - In the tree diagram example, there is relatively little to consider with regard to model fit. As for implications, would you consider $\Pr(A|B)$ and $\Pr(\bar{A}|\bar{B})$ to be sufficiently high to minimise incorrect test results?

Additional points before moving on

- Before delving further, we should clarify some things that are important or underpin this course theoretically.
  - The distinction between parameter and data.
  - Exchangeability.
  - Hierarchical modelling.

Parameter vs. Data

- Parameters are unobserved numerical quantities that define the mechanism that generate the observable data.

- Data is an observed quantity that is the outcome of a process that includes a random component. That is, data, y, are realisations of the random variable Y.

- The difference in Bayesian, compared to frequentist analysis, is that parameters are also treated as random variables.

Exchangeability

- Exchangeability states that the joint distribution of the observed data is invariant to permutation (π) of the indices,

$$\Pr(y_1, \ldots, y_n) = \Pr(y_{\pi_1}, \ldots, y_{\pi_n}),$$

- This is a common assumption in statistical analyses and is very similar to the assumption that observations y_i are identically and independently distributed.
- In fact, if y_i are *i.i.d.*, then y_i must be exchangeable as
$$\Pr(y_1, \ldots, y_n) = \prod_{i=1}^n \Pr(y_i) = \prod_{i=1}^n \Pr(y_{\pi_i}) = \Pr(y_{\pi_1}, \ldots, y_{\pi_n})$$
- However the converse is not true, can you think of an example?

Exchangeability

- Consider sampling without replacement.
- For example, lets say we select three marbles, two green and one white from a bag of 10 white and 15 green marbles. The joint probability is

Exchangeability

- Consider sampling without replacement.
- For example, lets say we select three marbles, two green and one white from a bag of 10 white and 15 green marbles. The joint probability is

$$\Pr(GGW) = \Pr(GWG) = \Pr(WGG) = \frac{15 \times 14 \times 10}{25 \times 24 \times 23} = 21/138 = 0.1522\ (4d.p.),$$

regardless of order, but as the order of events matters for the conditional probability,

$$\Pr(W = 1^{\text{st}}) = 10/25 \neq \Pr(W = 2^{\text{nd}} | G = 1^{\text{st}}) = 10/24,$$

the events are not independent.

Exchangeability

- Why is this of interest in Bayesian statistics?

  By the de Finetti Theorem, if y_i are drawn from an infinite sequence of exchangeable random variables, then

$$p(y_1, \ldots, y_n) = \int \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

  that is conditional on $\boldsymbol{\theta}$, $y_i; i = 1, \ldots, n$ are $i.i.d$.

Exchangeability

- Why is this of interest in Bayesian statistics?

  By the de Finetti Theorem, if y_i are drawn from an infinite sequence of exchangeable random variables, then

  $$p(y_1, \ldots, y_n) = \int \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta},$$

  that is conditional on $\boldsymbol{\theta}$, $y_i; i = 1, \ldots, n$ are $i.i.d$.

- Hence if observations are exchangeable, then the observation must represent a random sample from some model and there must exist a prior distribution for the parameters $\boldsymbol{\theta}$.

Hierarchies

- During this subject, we will look at, among other things, hierarchical models.
- Before we do, it would be useful to show how hierarchies work in probability.
  - How would you re-write $\Pr(A, B, C)$?

$$\Pr(A, B, C) = \Pr(A|B, C) \Pr(B|C) Pr(C)$$

- How could this be useful in a Bayesian analysis?
  - If C represents data and A, B are parameters then the joint posterior can be decomposed as follows

$$\Pr(A, B|C) = \Pr(A|B, C) \Pr(B|C)$$

    which may simplify computation.