Artificial Intelligence 2: Bayes' Theorem Revisit

Shan He

School of Computer Science
University of Birmingham

Outline of Topics

- In today's lecture, we will review the Bayes' theorem we learned from year 1's AI 1 module, which I am sure many of you have already forgot.
- This lecture will help you revisit this theorem by a visual journey of some bias coins
- Then we will discuss some basic concepts of Bayesian inference and how it differs from frequentist' inference.
- Finally we shall use an example to help you put the Bayes' theorem into the knowledge framework we built from our previous weeks' lecture, that is random variables and probability distributions. We shall also revisit some concepts we learned such as marginal Probability Mass Functions, joint Probability Mass Functions, etc.

Tossing biased coins

Example: Suppose we have some biased coins that have two possible biases of $\theta_{0.9} = 0.9$ and $\theta_{0.1} = 0.1$. The probability of a coin have the bias $\theta_{0.9}$ is

$$p(\theta_{0.9}) = c.$$

We observe the probability of that a coin will land heads up is

$$p(x_h) = a,$$

and the likelihood

$$p(x_h|\theta_{0.9}) = d$$

Question: Derive the posterior probability $p(\theta_{0.9}|x_h)$ without using Bayes' theorem.

- Suppose we have some biased coins that have two possible biases of $\theta_{0.9} = 0.9$ and $\theta_{0.1} = 0.1$. The probability of a coin have the bias $\theta_{0.9}$ is

$$p(\theta_{0.9}) = c.$$

- Here we assume that we know the proportion of those biases, which gives us this $p(\theta_{0.9}) = c$.
- We then randomly choose a coin and the toss it, record which side it lands, put it back (called replacement), and then repeat many time to obtian the probability of that a coin will land heads up is

$$p(x_h) = a,$$

- WE then choose a coin with bias 0.9, then toss it many times, record which side it lands, repeat this process many times to obtain the likelihood, or conditional probability of landing heads up given its bias is 0.9, which is b.
- Now given a coin you don't know the bias, can you simply by tossing it many times to infer which bias it has? For example, whether it has the bias 0.9. Or precisely, to infer the posterior probability $p(\theta_{0.9}|x_h)$. You can google Bayes' theorem and use the equation to derive it, but the question is infer the posterior probability without the theorem, or actually without any equations,

Tossing a biased coin

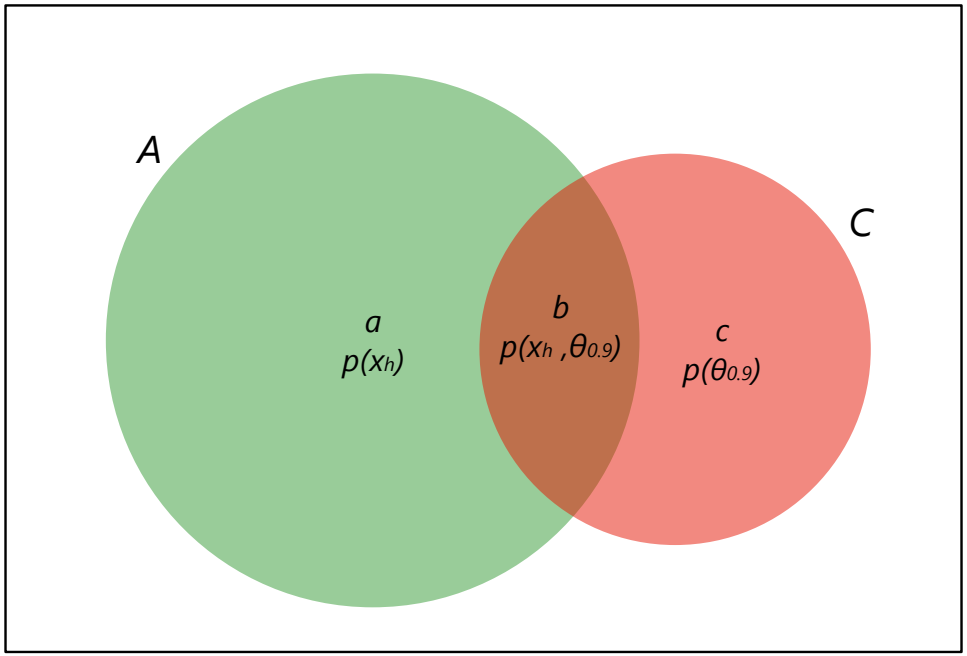


Figure 1: Bayes' theorem from a Venn Diagram. The total area of the whole box is 1. The area a of the green disk A represents the probability $p(x_h)$ that a coin will land with heads, x_h. The area c of the pink disk C represents the probability $p(\theta_{0.9})$ that a coin has bias $\theta = 0.9$. The overlap between A and C is the area b, which represents the joint probability $p(x_h, \theta_{0.9})$ that a coin will land heads up and that a coin has bias $\theta_{0.9}$

- Our solution is to visualise this problem by a Venn diagram.
- In this Venn diagram, the whole box represent the sample space and its area is 1. The area a of the green disk A represents the probability $p(x_h)$ that a coin will land with heads up, x_h.
- The area c of the pink disk C represents the probability $p(\theta_{0.9})$ that a coin has bias $\theta = 0.9$.
- The overlap between A and C is the area b, which represents the joint probability $p(x_h, \theta_{0.9})$ that a coin will land heads up and that a coin has bias $\theta_{0.9}$.
- However, from the question, we have not measured this joint probability, so its value b is unknown to us. But as you will see later, this value b is not useful.
- Now, write down something obvious from the Venn diagram. (NEXT SLIDE)

Bayes' theorem from Venn Diagram

From disk A, we can obtain

$$p(\theta_{0.9}|x_h) = \frac{b}{a} = \frac{p(x_h, \theta_{0.9})}{p(x_h)}, \tag{1}$$

and from disk C we have:

$$p(x_h|\theta_{0.9}) = \frac{b}{c} = \frac{p(x_h, \theta_{0.9})}{p(\theta_{0.9})} \tag{2}$$

From the above two equations, we have

$$p(x_h, \theta_{0.9}) = p(\theta_{0.9}|x_h)p(x_h) \tag{3}$$
$$p(x_h, \theta_{0.9}) = p(x_h|\theta_{0.9})p(\theta_{0.9}), \tag{4}$$

which yield

$$p(\theta_{0.9}|x_h)p(x_h) = p(x_h|\theta_{0.9})p(\theta_{0.9}) \Rightarrow p(\theta_{0.9}|x_h) = \frac{p(x_h|\theta_{0.9})p(\theta_{0.9})}{p(x_h)} \tag{5}$$

- From the diagram, or more precisely, disk A, we can see that (PAGE 4), the posterior probability $p(\theta_{0.9}|x_h)$ is essentially the proportion of overlap b to the area of disk A, where the overlap area is the joint probability $p(x_h, \theta_{0.9})$ and the area of disk A is the probability of observing landing heads up $p(x_h)$. This derivation gives us this conditional probability equation 1. But as we just discussed, b is unknown.
- Then let's focus on the pink disk C, we can also derive the conditional probability of observing a coin landing heads up give its has bias of 0.9, which is the proportion of the area b to the area of this pink disk c, which gives us the conditional probability equation 2.
- We then multiply equation 1 with $p(x_h)$ and equation 2 with $p(\theta_{0.9})$, which gives us the following two equations 3 and 4.
- Because two equations 3 and 4 equate to each other, which yield the equation 5
- Finally, by dividing both sides with $p(x_h)$, we obtain the Bayes' theorem or Bayes' rule. You can substitute the values into the equation, which gives us $\frac{b \times c}{a}$

Example: A doctor's patient record

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	Sum
x_1	8	9	9	5	4	1	1	0	0	0	37
x_2	3	5	8	9	14	10	3	3	0	0	55
x_3	0	1	1	10	16	11	12	7	8	5	71
x_4	0	0	1	0	3	5	10	7	7	4	37
Sum	11	15	19	24	37	27	26	17	15	9	200

Table 1: Joint observations of symptoms x_r, i.e., rows $r = \{1, 2, 3, 4\}$ and diseases θ_c, i.e., columns $c = \{1, 2, \ldots, 10\}$. The number $n(x_r, \theta_C)$ in each cell represents the number of patients with the symptoms x_r and disease θ_c.

- The Bayes' theorem we have just derived from the Venn diagram is for a single probability, which can be use to decide just two possible alternative outcomes, for example, a coin with bias of 0.9 or 0.1, or probably, a person with disease or without disease
- However, practically, to model a real-world random process, or an experiment, we typically need to decide which one of several alternatives, or as we learned in Lecture 1, different possible outcomes in the sample space of the experiment, we need to infer a set of probabilities of occurrence of different possible outcomes.
- This means, practically, we need to extend Bayes's theorem to deal with probability distributions rather than the point probabilities for the quantities in the theorem.
- Let's look at an example first before introducing Bayes' Theorem for probability distributions.
- Here is an record kept by an doctor, or a consultant who specialised in 10 diseases, for example, lung diseases, denoted as θ_1 to θ_10. He also know these diseases typically display 4 symptoms, denoted as x_1 to x_4. In the past 1 year, the doctor saw 200 patients, and kept this record of joint observations of these patients.
- The number $n(x_r, \theta_C)$ in each cell represents the number of patients with the symptoms x_r and disease θ_c.

Doctor's Questions

Questions from the doctor:

- Q1: What is the joint probability $p(x_3, \theta_2)$ that a patient has the symptom x_3 and the disease θ_2?
- Q2: What is the probability $p(x_3)$ that a patient has the symptom x_3?
- Q3: What is the probability $p(\theta_2)$ that a patient has the disease θ_2?
- Q4: What is the conditional probability $p(x_3|\theta_2)$ that a patient has the symptom x_3 given that he has the disease θ_2?
- Q5: What is the conditional probability $p(\theta_2|x_3)$ that a patient has the disease θ_2 give that he has the symptom x_3?

- Suppose this doctor is among you and would like to apply what we learned, including Bayesian theorem to analyse this patient record, he or she might ask the following questions.
- Q1: What is the joint probability $p(x_3, \theta_2)$ that a patient has the symptom x_3 and the disease θ_2?
- Q2: What is the probability $p(x_3)$ that a patient has the symptom x_3?
- Q3: What is the probability $p(\theta_2)$ that a patient has the disease θ_2?
- Q4: What is the conditional probability $p(x_3|\theta_2)$ that a patient has the symptom x_3 given that he has the disease θ_2?
- Q5: What is the conditional probability $p(\theta_2|x_3)$ that a patient has the disease θ_2 give that he has the symptom x_3?
- Please pause this video for 10 mins and try to work out the solutions as many as possible. You might need to use the knowledge we learned in Lecture 5, joint probability distributions. I shall give you the solutions in my next video.

Random Variables for doctors

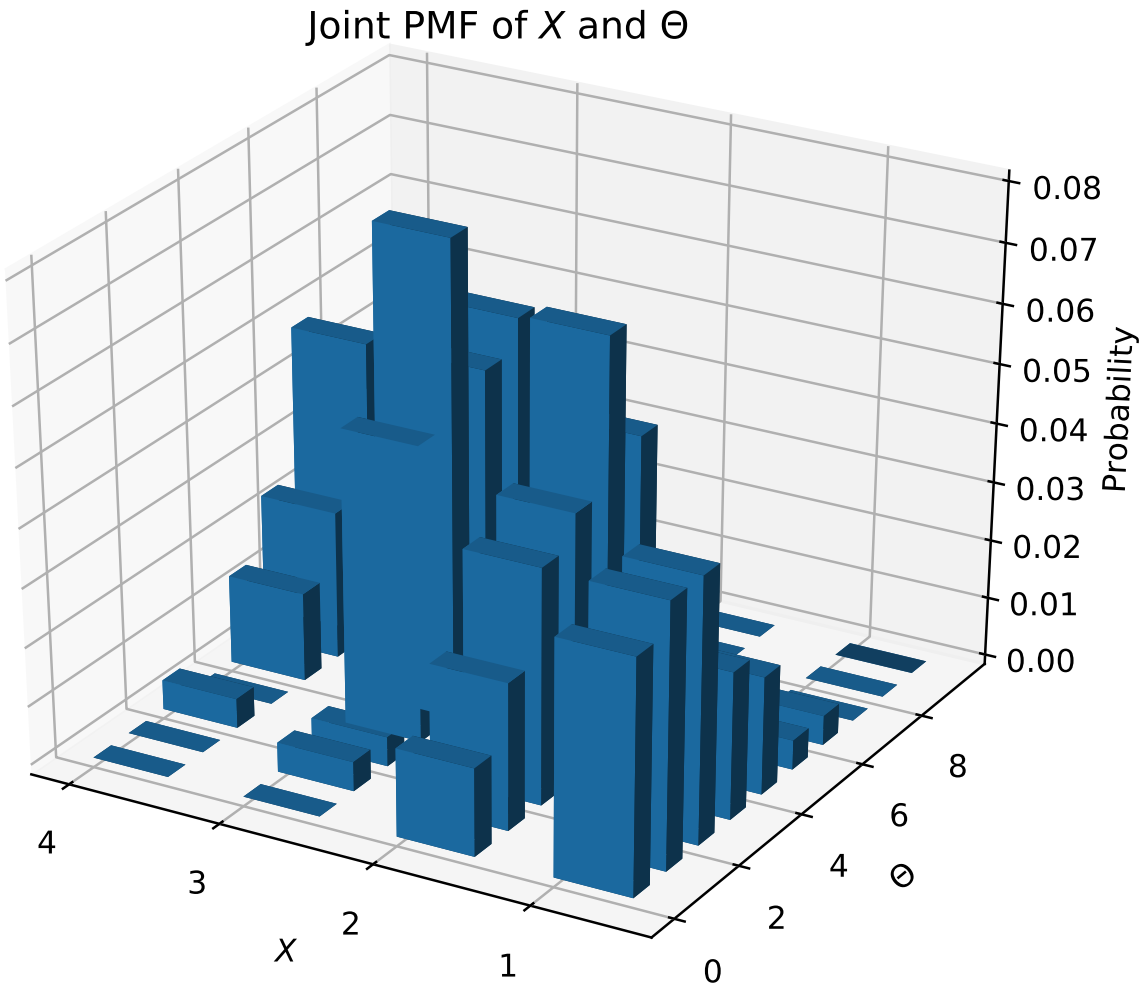
We first use random variables to model the patient record: let the discrete random variable X to represent the row number, i.e., one of the four symptoms, of which the range $R_X = \{x_1, x_2, x_3, x_4\}$ and Θ to represent the column number, i.e., one of the 10 diseases, of which the range $R_\Theta = \{\theta_1, \theta_2, \ldots, \theta_{10}\}$. We can now calculate the joint probability and obtain the joint PMF of X and Θ as the following table

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	$P(X)$
x_1	0.04	0.045	0.045	0.025	0.02	0.005	0.005	0	0	0	0.185
x_2	0.015	0.025	0.04	0.045	0.07	0.05	0.015	0.015	0	0	0.275
x_3	0	0.005	0.005	0.05	0.08	0.055	0.06	0.035	0.04	0.025	0.355
x_4	0	0	0.005	0	0.015	0.025	0.05	0.035	0.035	0.02	0.185
$P(\Theta)$	0.055	0.075	0.095	0.12	0.185	0.135	0.13	0.085	0.075	0.045	1

Table 2: Joint PMF of two discrete random variables X and Θ and their marginal PMFs. The marginal PMFs $P(X)$ and $P(\Theta)$ are also know as **the marginal likelihood** (of symptoms) and **the prior distribution** (of disease), respectively.

- As always in my lectures, we use random variables to represent our problems, which give us a mathematical framework to model this patient record.
- We use a discrete random variable X to represent the symptoms, or more accurately, the row number, i.e., one of the four symptoms. That means, the range of X $R_X = \{x_1, x_2, x_3, x_4\}$
- We then use Θ to represent the disease, or more precisely, the column number, i.e., one of the 10 diseases, of which the range $R_\Theta = \{\theta_1, \theta_2, \ldots, \theta_{10}\}$.
- We can now calculate the joint probability by dividing the number of each cell by the total number of patients (TURN TO PAGE 6), that is 200.
- This simple calculation can be easily done even by the all mighty Microsoft Excel. Now, we obtain the joint PMF of X and Θ as the following table.
- Just remind you, we also obtain the marginal distributions, or marginal PMFs, that is $P(X)$, the last column, and $P(\Theta)$, the last row. $P(X)$ is also know as the marginal likelihood (of symptoms) and $P(\Theta)$ is known as the prior distribution (of disease), respectively.

Joint PMF of X and Θ



Doctor's Questions

- Q1: What is the joint probability $p(x_3, \theta_2)$ that a patient has the symptom x_3 and the disease θ_2? A1: $p(x_3, \theta_2) = 0.005$
- Q2: What is the probability $p(x_3)$ that a patient has the symptom x_3? $p(x_3) = 0.355$
- Q3: What is the probability $p(\theta_2)$ that a patient has the disease θ_2? $p(\theta_2) = 0.075$
- Q4: What is the conditional probability $p(x_3|\theta_2)$ that a patient has the symptom x_3 given that he has the disease θ_2?
  A4:
$$p(x_3|\theta_2) = \frac{p(x_3, \theta_2)}{p(\theta_2)} = \frac{0.005}{0.075} = \frac{1}{15},$$
- Q5: What is the conditional probability $p(\theta_2|x_3)$ that a patient has the disease θ_2 give that he has the symptom x_3? A5:
$$p(\theta_2|x_3) = \frac{p(x_3|\theta_2)p(\theta_2)}{p(x_3)} = \frac{\frac{1}{15} \times 0.075}{0.355} = \frac{1}{71}$$

- Now, with this joint PMFs and marginal PMFs, we can easily solve the 5 questions.
- For the first question, that is, What is the joint probability $p(x_3, \theta_2)$ that a patient has the symptom x_3 and the disease θ_2, we can simply check the table, to find the cell with $X = x_3$ and Θ equals to the value θ_3. The answer is 0.005.
- Question 2, What is the probability $p(x_3)$ that a patient has the symptom x_3? $p(x_3) = 0.355$. That is essentially the marginal PMF $P(X)$ evaluated at x_3, and by the table, we can easily obtain it is 0.355. This is also the marginal likelihood of symptoms x_3, which means, regardless the disease, what is the likelihood a patient will have symptom x_3.
- Similarly, we can work out the answer to Q4, that is the marginal PMF $P(\Theta)$ evaluated at θ_2. As we just discussed, it is also the prior of disease θ_3.
- For question 4, **READ SLIDES**, we can use the conditional probability equation and use what we have obtained to calculate, which gives us 1 over 15.

More questions

- **Additional Question 1**: What is the probability of observing x_3 (symptom) given different values of Θ (disease)
- **Additional Question 2**: What are the conditional probabilities of a patient has all these disease given that he has the symptom x_3

- Now these answers brings more questions regarding how to generalise our answers to questions 4 and 5
- The first question is, what is the probability of observing x_3 given different values of Θ.
- The second question is, what are the conditional probabilities of a patient has all these disease given that he has the symptom x_3
- The second question is more practical, which essentially give us the ability to make decision. That is based on the all the conditional probabilities, we know that, given the symptom x_3, which disease the patient is more likely to have.
- But from our solution using the Bayes' rule, we need to solve the first question first.
- An it turns out, the answer is essentially what we learned before **(NEXT SLIDE)**

The Likelihood function

Answer to Additional Question 1: For this example, we use the likelihood function of observing x_3 given different values of Θ, defined as

$$\begin{aligned} p(x_3|\Theta) &= \{p(x_3|\theta_1), p(x_3|\theta_2), \ldots, p(x_3|\theta_{10})\} \\ &= \{0, \frac{1}{15}, \frac{1}{19}, \ldots, \frac{5}{9}\}, \end{aligned}$$

which can be written as

$$P(X = x|\Theta) = p(x|\Theta) \tag{6}$$

More Question: is this likelihood function the same as we learned in maximum likelihood method?

- Let's solve this additional question first. This is straightforward. We simply apply the same logic, or the same equation to calculate the conditional probabilities of $p(x_3)$ given different values of Θ, which gives us essentially a function, of which the range is $\{0, \frac{1}{15}, \frac{1}{19}, \ldots, \frac{5}{9}\}$. We call this the likelihood function.
- More generally, for this example which has a one dimensional X and one dimensional Θ, we can the likelihood function as equation 6.
- Now, some of you might ask, likelihood function, is the same as we learned in Lecture 7, in the Maximum Likelihood Estimation method?

The Likelihood function

Question: is this likelihood function the same as we learned in the Maximum Likelihood Estimation (MLE)?

Answer: The same but with different notations and different interpretations. The likelihood function in MLE:

$$L(\theta|x) = P_X(x;\theta),$$

where $P_X(x;\theta)$ is the PMF of X parametrised by parameter θ

- The short answer to this question is "Yes, they are the same". The long answer is The same but with different notations and different interpretations.
- Let's take a look at the notation first. Recalled what we learn in Lecture 7 (Maximum Likelihood Estimation), the likelihood function in Maximum Likelihood Estimation is written in the following equation, which essentially represents the PMF of X parametrised by parameter θ.
- However, look at equation 6 on page 10, it looks like a conditional joint probability distribution. Why is the difference?
- The reason is the different interpretations from two schools of statistics, that is Frequentist and Bayesian.

The Likelihood function: Frequentist vs Bayesian

Two interpretations from Frequentist and Bayesian of the Likelihood function, based on different interpretations of θ and Θ (although they are the same):

- Frequentist:
  - θ is the parameter of a predefined statistical model
  - The likelihood function is a statistical model that summarises a single random sample from a population [1], whose output value depends on a choice of parameters θ
- Bayesian:
  - Θ is a random variable
  - A function whose value is proportional to the probability of the conditional probability distribution X given Θ.

[1] Here "a single sample" means a specific subset of the population. For example, if we define the population as all patients admitted to QE hospital, a (random) sample could be defined as the patients aged 20-44 with trauma admitted in the past 1 months.

- This two interpretations from Frequentist and Bayesian are essentially the different interpretations of the same random variable Θ or the parameter θ
- For Frequentist, θ is regarded as the parameters of a predefined statistical model. Because of this interpretation, the likelihood function is regarded as a statistic that summarises a single sample from a population, whose calculated value depends on the model and the value of parameter θ
- Here, please do not be confused by "a single sample", it actually mean a specific subset of the population. For example, if we define the population as all patients admitted to QE hospital, a (random) sample could be defined as the patients aged 20-44 with trauma admitted in the past 1 months.
- For Bayesian, as we just discussed, Θ is a random variable. As we can see, it is a function whose value is proportional to the probability of the conditional probability distribution X given Θ.

Bayes' Theorem for Discrete Distributions

Bayes' theorem for Probability Distributions: Given two random variables, X and Θ

$$p(\Theta|x) = \frac{p(x|\Theta)p(\Theta)}{p(x)} \tag{7}$$

where $p(\Theta|x)$ is called posterior distribution, $p(x|\Theta)$ is the likelihood function, $p(\Theta)$ is called the prior, x is a value of X and $x \in R_X$, and $p(x)$ is a scalar, called marginal likelihood. In plain English:

$$\text{posterior distribution} = \frac{\text{likelihood function} \times \text{prior distribution}}{\text{marginal likelihood}}$$

- We can extend the above Bayes' theorem for Probability Distribution. We consider two random variables, X and Θ, then the Bayes' theorem is defined by this equation 5, where $p(\Theta|x)$ is called posterior distribution, $p(x|\Theta)$ is the likelihood function, $p(\Theta)$ is called the prior, x is a value of X and $x \in R_X$, and $p(x)$ is a scalar, which is called marginal likelihood.
- In plain English, it is, the posterior distribution can be obtained by the likelihood function times the prior distribution divided by the marginal likelihood.

Bayes' Theorem for Discrete Distributions

Answer to Additional Question 2 (pp.10):

$$p(\Theta|x_3) = \frac{p(x_3|\Theta)p(\Theta)}{p(x_3)}$$
$$= \{0, 0.01, 0.01, 0.14, 0.22, 0.15, 0.17, 0.01, 0.11, 0.07\}$$

Maximum a posteriori probability (MAP) estimate: the value of Θ that correspond to the maximum value of $p(\Theta|x_3)$, i.e.,

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\, p(\Theta|x_3) = \theta_5$$

- Now after a long journey, let's answer the question we let in page 10, that is, What is the posterior distribution of diseases given a patient has the symptom x_3.
- By some simple calculation using equation 7, we can obtain the following conditional probability distribution
- This posterior distribution allows us to identify the value of Θ that correspond to the maximum value of $p(\Theta|x_3)$, which defines the **maximum a posteriori (MAP)** estimate of the true value of Θ
- For this particular example, we know that the MAP estimate is disease number 5, which corresponds to the maximum posterior of 0.22.

Summary and Further reading

- Probability and Statistics: Chapter 7. Bayesian Inference