Advanced probabilistic machine learning

Linus Falk

August 28, 2023

Contents

1 Introduction

This course bla bla bla

2 Probability review

To begin with, lets start with a review of what we already know about probability and a look at the *Sample space*:

- Dice: Ω = {1,2,3,4,5,6}
- Coin flip: Ω = {heads, tails}

There are also *Events* which is a subset of the sample space:

- Dice: A = { 1,2}
- Coin flip = {tails}

Three important axioms to remember are.

1. P(A) $\geq$ 0
2. P(Ω) = 1
3. For disjoint sets, A_1,A_2, ...
   $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$

To sum it up here are some consequences of the these axioms:

1. $A \subseteq B \Rightarrow P(A) \leq P(B)$
2. $P(A^c) = 1 - P(B)$
3. $P(\varnothing) = 0$
4. $0 \leq P(A) \leq 1$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2.1 Conditional probability

The definition of conditional probability for two events are the following:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1}$$

Provided that the probability of event A is greater than 0, $P(A) > 0$. From this we can derive the **product rule:** $P(A \cap B) = P(B|A)P(A)$

2.2 Bayes theorem

Given two events called A and B:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2}$$

Example: medical test paradox

Let's characterize a medical test:

- Sensitivity(True positive rate) $= P(+|\text{Disease})$
- Specificity (True negative rate) $= P(-|\text{No Disease})$

Then characterize a medical condition in a population:

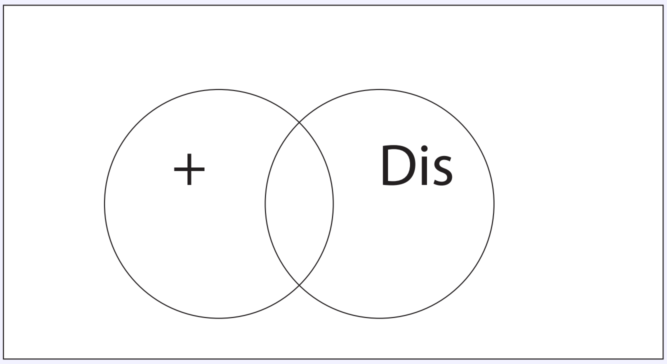
- Prevalence = "*proportion of a particular population found to be affected by a medical condition.*"

Question: Assume that for a certain disease the Sensitivity is 0.99, Specificity is 0.99 and Prevalence $= 0.01$ If you pick someone randomly from the population and test the person for this disease, obtaining a positive result. What is the probability of the person actually having the disease?

Using Bayes theorem:

$$P(\text{Disease}|+) = \frac{P(+|\text{Disease})P(\text{Disease})}{P(+)} = \frac{\text{tpr} \cdot p}{P(+)} \tag{3}$$

and by the law of total probability:
$P(+) = P(+ \cap \text{Disease}) + P(+ \cap \text{No disease})$
$P(+) = P(+|\text{disease})P(\text{disease}) + (1 - P(-|\text{No disease}))(1 - P(disease))$
$P(+) = \text{tpr} \cdot p + (1 - \text{tnr})(1 - p)$



Hence we have:

$$P(\text{disease}|+) = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.01 + 0.001 \cdot 0.99} \approx \frac{1}{11} \tag{4}$$

2.3 Random variables

The definition of a random variable $\mathcal{X}$ is a quantity that depends on a random event. Where a discrete random variable is a countable number of values. Described by the probability mass function $p(x) = P(X = x)$. An **example:** 6-sided dice with $p(x) = 1/6, x = 1, \ldots, 6$. A continuous random variable is an uncountable number of values. Ex: any value in $\mathbb{R}$. **Example:** $\mathcal{X}$ is the height of an adult randomly sampled from the population. It is described by the probability density function: $p(x)$.
A probability density function $p(x)$ describes the probability for a *continuous* random variable x falling into a given interval:

$$P(a < X < b) = \int_a^b p(x)\,\mathrm{d}x \tag{5}$$

2.4 Joint probability

Given two random variables: X and Y:

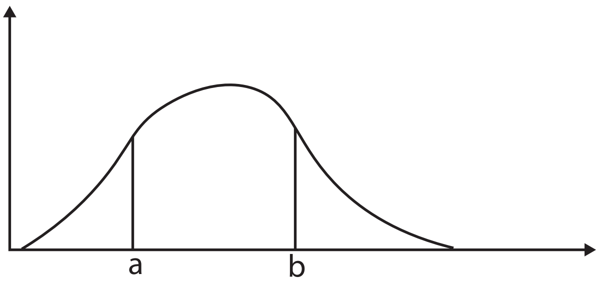


Figure 1: Example of caption

- Discrete: probability mass function
  $p(x,y) = P(X = x, Y = y)$
- Continuous: (probability density function)
  $P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d p(x,y)\,\mathrm{d}y\mathrm{d}x$

2.5 Marginalization and conditioning

Marginalization (also called the **sum rule**) is defined as

$$\begin{aligned} p(x) &= \int_y p(z,y)\,\mathrm{d}y \quad \text{if y is continuous} \\ p(x) &= \sum_y p(x,y) \quad \text{if y is discrete} \end{aligned} \tag{6}$$

Conditional probability also called the **product rule** is defined as:

$$\begin{aligned} p(x,y) &= \frac{p(x,y)}{p(y)} \quad \text{where} p(y) \neq 0 \\ &\Rightarrow p(x,y) = p(x|y)p(y) \end{aligned} \tag{7}$$

2.6 Bayes theorem: random variables

Both for probability mass functions and probability density functions:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{8}$$

2.7 Probabilistic modelling

In probabilistic modelling we consider two types of variables:

$$\begin{aligned} \mathcal{D} = \{x_1, x_2, \ldots, X_N\}: &\quad \text{observed variables} \\ \Theta = \{x_1, x_2, \ldots, X_N\}: &\quad \text{latent variables} \end{aligned} \tag{9}$$

In **probabilistic modelling** we treat *both* the observed variables and the latent variables as **random variables**

We model this relationship between $\mathcal{D}$ and Θ with its **Joint distribution**

$$p(\mathcal{D}, \Theta) = p(x_1, \ldots, X_N, z_1, \ldots, z_M) \tag{10}$$

2.8 Bayes theorem in Probabilistic modelling

The joint distribution factorizes into **likelihood** and a **prior**

$$p(\mathcal{D}, \Theta) = p(\mathcal{D}|\Theta)p(\Theta) \tag{11}$$

We can now find $p(\Theta|\mathcal{D})$ using **Bayes' theorem**

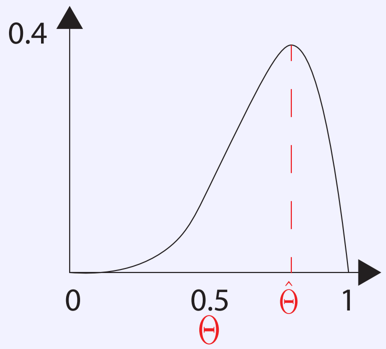
$$p(\Theta|\mathcal{D}) = \frac{p(\mathcal{D}|p(\Theta))}{p(\mathcal{D})} \tag{12}$$

- $\mathcal{D}$: observed data
- Θ : latent variables explaining the data
- $p(\Theta)$: **prior** belief of latent variables before seeing data
- $p(\mathcal{D}|\Theta)$: **likelihood** of the data in view of the latent variables
- $p(\Theta|\mathcal{D})$: **posterior** belief of latent variables in view of data
- $p(\mathcal{D})$: The **marginal likelihood** (presented in lecture 3)

Example: Coin flip (>Frequentist viewpoint)

- $x \in \{0, 1\}$ represent the outcome of flipping a damaged coin
- $p(x = 1|\Theta) = \Theta$, is a deterministic parameter
- Nor prior belief encoded

Question: After we observe N coin flips $\mathcal{D} = \{x_1, \ldots, x_N\}$, which $\hat{\Theta}$ makes the observed data most likely?
Solution: Maximize the likelihood function $p(\mathcal{D}|\Theta)$



Example: Coin flip (Bayesian viewpoint)

- $x \in \{0, 1\}$ represent the outcome of flipping a damaged coin
- $p(x = 1|\mu) = \mu$, which is also a random variable.
- Probability distribution $p(\mu)$: our *prior belief*

Question: After we observe N coin flips $\{x_1, \ldots, x_N\}$, what is our belief $p(\mu|x_1, \ldots, x_N)$?
Solution: Bayes theorem states that:

$$p(\mu|x_1, \ldots, x_N) \propto p(x_1, \ldots, x_N|\mu)p(\mu) \tag{13}$$

We will continue with this example next lecture...

2.9 Condluding remarks

Probabilistic/Bay inference is a flexible way of dealing the machine learning problems. Some properties to remember:

- Treat not only the data, but also the model and its parameters (if they are parametric) as random variables
- After learning you not only get a single model, you get a distribution of likely models.
- You can also encode prior knowledge you might have about the model and its parameters.

2.10 Summary

Probability distribution is a function that describes the likelihood of obtaining the possible values that a random variable can assume. **Conditional and marginalization** are two basic rules for manipulating probability distributions. **Frequentist vs Bayesian**: the first assume true values underlying some experiment and the second require some initial belief to be set on possible values.

- **Bayes theorem**, $p(x|y) = p(y|x)p(x)/p(y)$
- **Prior**, belief of parameters before we have seen any data
- **Likelihood**, belief of data in view of the parameters
- **Posterior**, belief of parameters after inferring data