Supervised Learning Program
Probability

Prepared by: Aditya Khambete
Supervisor : Prof. Ayan Bhattacharya

Spring 2024-25

Contents

Abstract

This is the running document for the supervised learning program I am doing this semester under the guidance of Prof. Ayan Bhattacharya. The aim of this document is to provide a comprehensive understanding of the topics I cover in the program. The document is divided into chapters, each chapter covering what I have learned in a week. The document is a work in progress and will be updated regularly. The references I am using for this program are mentioned in the beginning of each chapter.

Chapter 1

Week 1

Abstract

Covered Topics:

- Chapter 1-5 from the book

References: Lecture notes of Prof. Ayan Bhattacharya [1] and the book by Lesigne [2]

1.1 Basic Probability

Let (Ω, P) be a finite probability space. [1] Write $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ and $P(\omega_i) = p_i$. We have the most basic formula of probability i.e.

$$1 = \sum_i p_i, \text{ where each } 0 \leq p_i \leq 1 \tag{1.1}$$

Definition 1.1.1. The probability of an event A is defined as

$$P(A) = \sum_{\omega_i \in A} p_i = \sum_{i=1}^{n} p_i \mathbb{I}_A(\omega_i) \tag{1.2}$$

where $\mathbb{I}_A$ is the indicator function of A. This function maps ω_i to 1 if $\omega_i \in A$ and 0 otherwise.

Some more basic properties of probability are:

- $P(\Omega) = 1$
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

Such set Ω is called a sample space, and P is called a probability function. It is easy to see from above properties that

- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Definition 1.1.2. The Probability Space is called uniform if p_i is the same for all ω_i.

[1] The book doesn't mention the sigma field $\mathscr{F}$.

1.1.1 Sequential Experiments

Let us demonstrate this through an elementary example. Assume a binary experiment, taking outcomes 0,1 with q,p respectively. Easy to see that $p + q = 1$. Now, consider we repeat this experiment n times. The sample space in this case is $\Omega_n = \{0, 1\}^n$. The probability of a sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ is

$$P((\omega_1, \omega_2, \ldots, \omega_n)) = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i} \tag{1.3}$$

This is the probability function defined on the sample space Ω_n. This is a simple example of a product probability space. We say the space $\Omega_n = \{0, 1\}^n$ is equipped with the probability function $P_n = (q, p)^{\otimes n}$, where q,p are the probabilities of 0,1 respectively.
More details on why we did this product come from the notion of independence.

1.2 Random Variables

Definition 1.2.1. A random variable is a function $X : \Omega \to \mathbb{R}$.

We use the denotion $(X = x)$ for the set $\{\omega \in \Omega : X(\omega) = x\}$. The probability of this event is $P(X = x)$, this is also known as the probability mass function of X. Similarly cumulative distribution function is defined as $F(x) = P(X \leq x)$.

A very underrated fact is the space of random variables is a vector space. This is because the sum of two random variables is also a random variable, so is the product of a random variable with a scalar. The basis of this vector space is the indicator functions of the form $\mathbb{I}_{\omega_i}$ where $\omega_i \in \Omega$. [2]

Definition 1.2.2. Expectation of a random variable X is defined as (if the sum converges)

$$E[X] = \sum_{i=1}^{k} x_i P(X = x_i), \text{ where } k \text{ is the number of distinct values of X} \tag{1.4}$$

Some easy to see properties[3] of expectation are:

- $|E[X]| \leq E[|X|]$
- $E[X] \geq 0$ if $X \geq 0$
- $E[c] = c$ for any constant c, particularly $E[E[X]] = E[X]$
- $E[aX] = aE[X]$
- $E[X + Y] = E[X] + E[Y]$

Say we write $X = \sum_{i=1}^{k} x_i \mathbb{I}_{A_i}$, where $A_i = (X = x_i)$. Then $E[X] = \sum_{i=1}^{k} x_i P(A_i)$, say we take some function $g : \mathbb{R} \to \mathbb{R}$, then write $Y = g(X) = \sum_{i=1}^{k} g(x_i) \mathbb{I}_{A_i}$, hence applying E on both sides, we get

$$E[Y] = E[g(X)] = \sum_{i=1}^{k} g(x_i) P(A_i) \tag{1.5}$$

1.2.1 Some nice inequalities

Theorem 1.2.3. ***Markov's Inequality:*** *Let X be a non-negative random variable, then for any $a > 0$, we have*

$$P(X \geq a) \leq \frac{E[X]}{a} \tag{1.6}$$

This inequality right above is in some sense the mother of all inequalities. The proof is fairly easy, just use the fact that $P(X \geq a)$ can be written as a summation of $P(X = x_i)$ for $x_i \geq a$, multiply by $\frac{x_i}{a}$ and sum over all x_i.

[2] I_{ω_i} is the function that maps ω_i to 1 and all other ω_j to 0. As one might expect, this is a random variable as well.
[3] The last 2 facts imply E is a linear functional on the vector space of random variables.

Theorem 1.2.4. ***Chebyshev's Inequality:*** *Let X be a random variable with finite expectation and variance, then for any $a > 0$, we have*

$$P(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2} \tag{1.7}$$

Follows from Markov's inequality, just use the fact that $Var[X] = E[(X-E[X])^2]$, and apply Markov's inequality on $Y = (X - E[X])^2$.

Definition 1.2.5. The variance of a random variable X is defined as

$$Var[X] = E[(X - E[X])^2], \text{ it simplifies to } E[X^2] - E[X]^2 \tag{1.8}$$

1.3 Independence

Bibliography

[1] Ayan Bhattacharya. Lecture notes on probability ii. SI 537 IIT Bombay, Autumn 2024-25.

[2] E. Lesigne. *Heads or Tails: An Introduction to Limit Theorems in Probability*. Student mathematical library. American Mathematical Society, 2005.