# 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 $(\Omega, P)$ 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 $\omega_i$ 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 $\Omega$ 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 $\omega_i$.

[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 $\Omega_n$. 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] $I_{\omega_i}$ is the function that maps $\omega_i$ to 1 and all other $\omega_j$ 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

**Definition 1.3.1.** Two events $A, B$ are independent if $P(A \cap B) = P(A)P(B)$. Except the trivial case with $P(B) = 0$, we can write this as $\frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(\Omega)}$.

Once we have the notion of independence, we can check where does the product probability space come from.

$$P((\omega_1, \omega_2, \ldots, \omega_n)) = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i} \tag{1.9}$$

Let us extend the notion to more than 1 event,

**Definition 1.3.2.** A set of events $A_1, A_2, \ldots, A_n$ are independent if for any subset $I \subset \{1, 2, \ldots, n\}$, we have

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i) \tag{1.10}$$

One very important thing to note is that if the events $A_1, A_2, \ldots, A_n$ satisfy $P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$ it doesn't imply that the events are independent, just take one of the events to be the empty set for example. An interesting example showing that pairwise independence doesn't imply mutual independence is the following. Say we havre a fair coin, define,

$$A_1 = \{HH, HT\}$$
$$A_2 = \{HH, TH\}$$
$$A_3 = \{HT, TH\}$$

Then $A_1, A_2, A_3$ are pairwise independent, but not mutually independent.

### 1.3.1 Independence of Random Variables

Random variables are independent if the events $\{X_i = x\}$ are independent for all $x \in \mathbb{R}$. Alternatively, we can say that the events $(X_1 \in B_1)$ , $(X_2 \in B_2)$, ..., $(X_n \in B_n)$ are independent for all $B_1, B_2, \ldots, B_n \subset \mathbb{R}$. Its not hard to notice the following facts:

- independence of random variables doesn't depend on the order of the random variables.
- random variables $X_1, X_2, \ldots, X_n$ are independent if the events ($X_1 = x_1$ and $X_2 = x_2$ and $\ldots$ and $X_{j-1} = x_{j-1}$)[4] is independent of $(X_j = x_j)$ for all $j \in \{2, 3, \ldots, n\}$, $x_1, x_2, \ldots, x_{j-1}, x_j \in \mathbb{R}$.
- 2 events are independent if and only if the indicator functions are independent.

Now notice the following fact,

**Theorem 1.3.3.** *If $X, Y$ are independent random variables, then*

$$E[XY] = E[X]E[Y] \tag{1.11}$$

[4]Writing ($X_1 = x_1$ and $X_2 = x_2$) is the same as writing $(X_1 = x_1) \cap (X_2 = x_2)$

*Proof.* Let A and B be the range of $X, Y$ respectively,
We have

$$\begin{aligned} E[XY] &= \sum_{x\in A}\sum_{y\in B} xyP(X=x, Y=y) \\ &= \sum_{x\in A}\sum_{y\in B} xyP(X=x)P(Y=y) \\ &= \sum_{x\in A} xP(X=x)\sum_{y\in B} yP(Y=y) && \{\text{By Independence}\} \\ &= E[X]E[Y] \end{aligned}$$

□

**Corollary 1.3.4.** *If $X, Y$ are independent random variables, then $Var[X+Y] = Var[X] + Var[Y]$*

*Proof.*

$$\begin{aligned} Var[X+Y] &= E[(X+Y)^2] - E[X+Y]^2 \\ &= E[X^2] + E[Y^2] + 2E[XY] - E[X]^2 - E[Y]^2 - 2E[X]E[Y] \\ &= Var[X] + Var[Y] \end{aligned}$$

□

## 1.4 Binomial Distribution

Take the product probability space $\Omega_n = \{0,1\}^n$ with $P_n = (q,p)^{\otimes n}$. Let $S_n$ be the number of 1's in the sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. Then $S_n$ is a random variable on $\Omega_n$.

**Theorem 1.4.1.** *The probability mass function of $S_n$ is given by*

$$P(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{1.12}$$

We say the random variable $S_n$ follows a binomial distribution with parameters $n, p$. Note that $q$ is not a parameter, it is just $1-p$.
To prove this, just notice the probability of selecting a specific sequence with $k$ 1's is $p^k q^{n-k}$, and the number of such sequences is $\binom{n}{k}$.

**Definition 1.4.2.** Bernoulli distribution is a special case of binomial distribution with $n = 1$. It is the same as a unbiased coin flip. Few properties-

- $E[X] = p$
- $Var[X] = pq = p(1-p)$

**Theorem 1.4.3.** *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with parameter $p$. Then the sum $S_n = X_1 + X_2 + \ldots + X_n$ follows a binomial distribution with parameters $n, p.s$*

The proof is fairly easy from the results we have already proved.

**Theorem 1.4.4.** *$E[S_n] = np$ and $Var[S_n] = npq$*

*Proof.*

$$\begin{aligned} E[S_n] &= E[X_1 + X_2 + \ldots + X_n] \\ &= E[X_1] + E[X_2] + \ldots + E[X_n] \\ &= np \end{aligned}$$

For variance, we have

$$\begin{aligned} Var[S_n] &= Var[X_1 + X_2 + \ldots + X_n] \\ &= Var[X_1] + Var[X_2] + \ldots + Var[X_n] \\ &= npq \end{aligned}$$

□

## 1.5 Weak Law of Large Numbers

As in the last section, let $S_n$ be the number of successful trials in $n$ independent Bernoulli trials with parameter $p$. Then $S_n$ follows a binomial distribution with parameters $n, p$. $S_n$ is a random variable over the product probability space $(\Omega_n, P_n)$
The experimental probability of success is $\frac{S_n}{n}$, intutively we expect this to be close to $p$ as $n$ increases. This is the weak law of large numbers. [5]

**Theorem 1.5.1.** *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with parameter $p$. Let $S_n = X_1 + X_2 + \ldots + X_n$. Then for any $\epsilon > 0$, we have*

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \to 0 \text{ as } n \to \infty \tag{1.13}$$

*Proof.* Use Chebyshev's inequality, we have

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) &\leq \frac{Var[S_n/n]}{\epsilon^2} \\ &= \frac{Var[S_n]}{n^2\epsilon^2} \\ &= \frac{p(1-p)}{n\epsilon^2} \qquad \text{(By Independence)} \\ &\to 0 \text{ as } n \to \infty \end{aligned}$$

□

This has a nice application in analysis, the problem of uniformly approximating a continuous function by a polynomial. This is known as Weierstrass Approximation Theorem. To state it rigorously, let $f : [0,1]^6 \to \mathbb{R}$ be a continuous function. Then for any $\epsilon > 0$, there exists a polynomial $P(x)$ such that $|f(x) - P(x)| < \epsilon$ for all $x \in [0,1]$.
Serge Bernstein gave a probabilistic proof of this theorem.

**Lemma 1.5.2.** *Let $f$ be a continuous function on $[0,1]$. Then,*

$$sup_{x\in[0,1]}|f(x) - P_n(x)| \to 0 \text{ as } n \to \infty \text{ ,where } P_n(x) = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{n}{k}\right) (x^k)(1-x)^k \tag{1.14}$$

*Proof.* Fix $\epsilon > 0$, $\exists\ \eta$ such that

$$|x - y| < \eta \implies |f(x) - f(y)| < \epsilon \text{ where } 0 \leq x, y \leq 1 \tag{1.15}$$

Consider the space $(\Omega_n, P_n)$, and the random variable $f\left(\frac{S_n}{n}\right)$ where $S_n$ is as above. We have

$$\begin{aligned} E\left[f\left(\frac{S_n}{n}\right)\right] &= \sum_{k=0}^{n} f\left(\frac{k}{n}\right) P(S_n = k) \\ &= \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

By WLLN, we have

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \eta\right) \to 0 \text{ as } n \to \infty \tag{1.16}$$

[5] The proof presented is not true for all distributions, it is true only for Bernoulli here, and more generally for any distribution with finite variance, the standard statement of weak law of large numbers is for any distribution with finite expectation, and the existence of variance is not necessary, and that proof is much more complicated.

[6] This can be generalised to any closed interval [a,b]

Hence, $\exists\ N_0$ independent of $p$, such that for every $n \geq N_0$

$$P\left(\left|\frac{S_n}{n} - p\right| > \eta\right) < \epsilon \tag{1.17}$$

Also we have

$$\left|E_n\left[f\left(\frac{S_n}{n}\right)\right] - f(p)\right| = \left|\sum_{k=0}^{n} f\left(\frac{k}{n} - f(p)\right) P_n(S_n = k)\right|$$

breaking the sum into 2 parts, and applying the triangle inequality, we get the following upper bound,

$$\begin{aligned}
&\sum_{\left|\frac{k}{n}-p\right|\leq\eta} \left|f\left(\frac{k}{n}\right) - f(p)\right| P_n(S_n = k) + \sum_{\left|\frac{k}{n}-p\right|>\eta} \left(\left|f\left(\frac{k}{n}\right)\right| + |f(p)|\right) P_n(S_n = k) \\
\leq &\sum_{\left|\frac{k}{n}-p\right|\leq\eta} \epsilon P_n(S_n = k) + \sum_{\left|\frac{k}{n}-p\right|>\eta} 2 \sup_{0\leq x\leq 1} |f(x)| P_n(S_n = k) \\
= &\epsilon + 2 \sup_{0\leq x\leq 1} |f(x)| P_n\left(\left|\frac{S_n}{n} - p\right| > \eta\right) = \epsilon + 2 \sup_{0\leq x\leq 1} |f(x)|\epsilon
\end{aligned}$$

Which shows the upper bound can be made arbitrarily small, hence the proof. □

# 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.