Probability for Computer Scientists 36-218 CMU

Created by Lorenzo Xiao

November 8, 2022

1 Sample Space and Probability

1.1 Probablistic Model

1.1.1 Elements of a Probabilitic Model

- Sample Space, the set of all possible outcomes of an experiment.
  - collectively exhaustive, in the sense that no matter what happens in the experiment, we always obtain an outcome that has been included in the sample space .
- The probability law, assigns to a set A of possible coutcome(**event**) a nonnegative $P(A)$that encodes the likelihood of the elements of A.

1.1.2 Sequential Models, Continous Models, and discrete model

- Events like coin tossing is often described in sequenial models such as trees.
- Independent events where the probabilities of single-element is enough to chracterize the probability law uses a Discrete model.
  - If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events

that consist of a single element. In particular, the probability of any event {SI' S2, ... , Sn} is the sum of the probabilities of its elements:

$$P(\{s_1, s_2, s_3 \cdots, s_n\} = P(s_1) + P(s_2) + P(s_3) + \cdots + P(s_m))$$

  – If the sample space consists of n possible outcomes which are equally likely, then the probability of any event A is given by:

$$P(A) = \frac{\text{size of } A}{n}$$

- Events that the probabilities of single-element may not be sufficent to characterize the probability law will use a Countinous model

1.1.3 Properties of Probability Laws

- $A \subset B \to P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$
- $P(A \cup B \cup C) = P(A) + P(A^c \cap B) + P(A^c \cap B^c \cap C)$

1.2 Conditional Model

1.2.1 Properties of Conditional Probability

- The probability of an event A, given an evnet B with $P(B) > 0$ is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Conditional probabilities can also be viewed as a probability law on a new universe B, because all of the conditional probability is concentrated on B.

1.2.2 Using Conditional Probability for Modeling

When constructing probabilistic models for experiments that have a sequential character, it is often natural and convenient to first specify conditional probabilities and then use them to determine unconditional probabilities.

- Muliplicatoin rule :

$$P(\cap_{i=1}^{n} A_i) = P(A_1) P(A_2 \mid A_1) P(A_3 \mid A_1 \cap A_2) \cdots P\left(A_n \mid \cap_{i=1}^{n-1} A_i\right)$$

1.3 Total Probability Theorem and Baye's Rule

1.3.1 Total Probability Theorem

Let $A_1, \cdots, A_n$ be disjoint events that form a partition of the sample space and assume that $P(A_i) > 0$ for all i. Then, for any event B, we have:

$$P(B) = P(A_1 \cap B) + P(A_1 \cap B) + \cdots + P(A_n \cap B)$$

1.3.2 Bayes' Rule

Let $A_1, \cdots, A_n$ be disjoint events that form a partition of the sample space, and assume that $P(A_i) > 0$ for all i:

$$P(A_i \mid B) = \frac{P(A_i) P(B \mid A_i)}{P(B)} = \frac{P(A_i) P(B \mid A_i)}{P(A_1 \cap B) + P(A_1 \cap B) + \cdots + P(A_n \cap B)}$$

1.4 Independece and Counting

1.4.1 Independence

- Two events A and B are said to be independent if:

$$P(A \cap B) = P(A) P(B)$$

- If A and B are independent, so are A and B^C.
- Independence does not imply eonditional independence, and vice versa.

1.4.2 The Counting principle

Consider a process that consists of r stages, suppose that :

- There are n_1 possible results at the first stage
- For every possible result, there are n_2 possible result at the second stage
- The total number of possible results of the r-stage process is:$n_1 n_2 n_3 \cdots n_r$

1.4.3 Counting result

- Permutations of n: $n!$
- k-permutation of n:$\frac{n!}{(n-k)!}$
- Combination of k out of n objects:$\begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{k!(n-k)!}$
- Partition of n object in r group: $\frac{n!}{n_1! n_2! \cdots n_r!}$

2 Discrete Random Variables

2.1 Basic Concepts

2.1.1 Main Concepts Related to Random Variables

- A **random variable** is a real-valued function of the outcome of the experiment.
- A **function of a random** variable defines another random variable.
- We can associate with each random variable certain "averages" of interest, such as the **mean** and the **variance**.
- A random variable can be conditioned on an event or on another random variable.
- There is a notion of **independence** of a random variable from an event or from another random variable.

2.1.2 Concepts Related to Discrete Random Variables

- A **discrete random variable** is a real-valued function of the outcome of the experiment that can take a finite or count ably infinite number of values.
- A discrete random variable has an associated **probability mass function (PMF)** which gives the probability of each numerical value that the random variable can take.
- A **function of a discrete random variable** defines another discrete random variable, whose PMF can be obtained from the PMF of the original random variable.

2.2 PMF

2.2.1 Calculation of the PMF of a Random Variable X

For each possible value of x of X:

1. Collect all the possible outcomes that give rise to the event
2. Add their probability to obtain $\rho(x)$

2.2.2 Binomial Distribution

A binomial experiment possesses the following properties:

1. consists of a fixed number of identical trials.
2. Each trial can either be fail or sucess

3. the probaility of success of a single trial is maintined trial to trial as $q = 1 - p$

4. Independent

5. random variable of interest is Y, the number of sucess observed during the n trials.

$$P\left(y\right) = \left(\begin{array}{c} n \\ y \end{array}\right) p^{y} q^{n-y}$$

Let Y be a binomial random variable based on n trials and success probability p. Then:

$$\mu = E\left(Y\right) = np \quad \text{and} \quad \sigma^{2} = V\left(Y\right) = npq$$

2.2.3 Geometric Distribution

A random variable Y is said to have a **geometric probability distribution** if and only if

$$p\left(y\right) = q^{y-1} p$$

Let Y be is a random variable with a geometric distribution:

$$\mu = E\left(Y\right) = \frac{1}{p} \quad \text{and} \quad \sigma^{2} = V\left(Y\right) = \frac{1-p}{p^{2}}$$

2.2.4 Poisson probability distribution

A random variable Y is said to have a Poisson probability distribution if and only if:

$$p\left(y\right) = \frac{\lambda^{y}}{y!} e^{-\lambda}$$

Let Y be is a random variable with aPoisson probability distribution:

$$\mu = E\left(Y\right) = \lambda \quad \text{and} \quad \sigma^{2} = V\left(Y\right) = \lambda$$

2.3 Expectation, Mean, and Variance

2.3.1 Expectation

We define the expected value (also called the expectation or the mean) of a random variable X with PMF ρ_X by:

$$E\left[X\right] = \sum_{x} x \rho_{X}\left(X\right)$$

2.3.2 Expected Value Rule

Let X be a random variable with PMF ρ_X, and let $g(X)$ be a function of X. Then, the expected value of the random variable $g(X)$ is given by :

$$E[g(X)] = \sum_x g(x)\, \rho_X(X)$$

2.3.3 Variance

The variance $\operatorname{var}(X)$ of a random variable X is defined by:

$$\operatorname{var}(X) = E\left[(X - E[X])^2\right] = \sum_x (X - E[X])^2 \rho_X(X)$$

2.4 Joint PMF of Multiple Function

Let X and Y be random variables associated with the same experiment.

- The joint PMF $\rho_{x,y}$ of X and Y is defined by:

$$\rho_{x,y}(x, y) = P(X = x, Y = y)$$

- The marginal PMFs of X and Y can be obtained from the joint PMF, using the formulas:

$$\rho_x(x) = \sum_y \rho_{x,y}(x, y), \qquad \rho_x(y) = \sum_x \rho_{x,y}(x, y)$$

- A function $g(X, Y)$of X and Y defines another random variable, and:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y)\, \rho_{x,y}(x, y)$$

- The above have natural extensions to the case where more than two random variables are involved.

2.5 Conditioning and Independence

2.5.1 Conditional PMFs

- Conditional PMFs are similar to ordinary PMFs, but pertain to a universe where the conditioning event is known to have occurred.

- The conditional PMF of X given an event A with $P(A) > 0$, is defined by:

$$\rho_{x|A}(x) = P(X = x \mid A)$$

  and satisfies$\sum_x \rho_{x|A}(x) = 1$

- If $A_1, \cdots, A_n$ are disjoint events form a partition of the sample space, with with $P(A_i) > 0$, for all i, then

$$\rho_{x|A}(x) = \sum_{i=1}^{n} \rho_{x|A_i}(x)$$

- The conditional PMF of X given $Y = y$ is related to the joint PMF by

$$\rho_{X,Y}(x, y) = \rho_Y(Y)\, \rho_{X|Y}(x \mid y)$$

- The conditional PMF of X given Y can be used to calculate the marginal PMF of X through the formula

$$\rho_X(x) = \sum_y \rho_Y(Y)\, \rho_{X|Y}(x \mid y)$$

- There are natural extensions of the above involving more than two random variables.

2.5.2 Conditional Expectation

- The conditional expectation of X given an event A with$P(A) > 0$, is defined by

$$E[X \mid A] = \sum_x x \rho_{X|A}(x)$$

- The conditional expectation of X given a value y of Y is defined by

$$E[X \mid Y = y] = \sum_x x \rho_{X|A}(x \mid y)$$

- If $A_1, \cdots A_n$be be disjoint events that form a partition of the sample space, with $P(A_i) > 0$ for all i, then

$$E[X] = \sum_{i=1}^{n} \rho(A_i)\, E[X \mid A_i]$$

- We have

$$E[X] = \sum_{i=1}^{n} \rho_y(y)\, E[X \mid Y = y]$$

3 General Random Variables

3.1 Continuous Random Variables and PDFs

3.1.1 Countinous Function

A random variable X is called continuous if there is a nonnegative function f x, called the probability density function of X, or PDF for short, such that:

$$P\left(a \le X \le b\right) = \int_a^b f_X\left(x\right) dx$$

can be interpreated as area under the graph of PDF

3.1.2 PDF Properties

- $f_X\left(x\right) dx \ge 0$ for all x
- $\int_{-\infty}^{\infty} f_X\left(x\right) dx = 1$
- if δ is very small, then $P\left(\left[x, x+\delta\right]\right) \approx f_X\left(x\right) \cdot \delta$
- For any subset B of the real line: $\left(X \in B\right) = \int_B f_X\left(x\right) dx$

3.1.3 Expectation

The **expected value** or **expectation** or **mean** of a continuous random variable X is defined by

$$E\left[X\right] = \int_{-\infty}^{\infty} x f_X\left(x\right) dx$$

3.1.4 Properties of Expectations

- The expected value rule for a function g(X) has the form:

$$E\left[g\left(X\right)\right] = \int_{-\infty}^{\infty} g\left(x\right) f_X\left(x\right) dx$$

- The variance of X is defined by

$$\text{var}\left(X\right) = E\left[\left(X - E\left[X\right]\right)^2\right] f_X\left(x\right) dx$$

- If $Y = aX + b$:

$$E\left[Y\right] = aE\left[X\right] + b \qquad \text{var}\left(Y\right) = a^2 \text{var}\left(X\right)$$

- Exponential*

3.2 Cumulative Distribution Functions

3.2.1 Properties of a CDF

The CDF F_X of a random variable X is defined by:

$$F_X\left(x\right) = P\left(X \le x\right), \forall x$$

and has several properties:

- montonically non-decreasing
- $\lim_{x\to\infty} F_X(x) = 1, \lim_{x\to-\infty} F_X(x) = 0$
- If X is discrete, F_X will be a piecewise constant function of x
  - If X is discrete and takes integer values, the PMF and the CDF can be obtained from each other by summing or differencing:

$$F_X(k) = \sum_{i=-\infty}^{k} \rho_X(i)$$

$$\rho_X(k) = P(X \leq k) - P(X \leq k-1) = F_X(x) - F_X(x-1), \forall k \in \mathbb{Z}$$

- If X is continous, F_X will be a piecewise ccontinuous function of x
  - If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation:

$$F_X(k) = \int_{-\infty}^{x} f_X(t)\,dt \qquad f_X(x) = \frac{dF_X}{dx}(x)$$

3.3 Normal Random Variables

3.3.1 Normality Preserved by Linear Transformations

If X is a normal random variable with mean μ and variance σ^2, and if$a \neq 0, b$ are scalars, then the random variable

$$Y = aX + b$$

is also normal with mean and variance:

$$E[Y] = a\mu + b \qquad \text{var}(Y) = a^2\sigma^2$$

3.3.2 CDF Calculation for a Normal Random Variable

For a normal random variable with mean μ and variance, we use a two-step procedure.

1. Standardize X
2. Read the cdf value from the table

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

3.4 Joint PDFs of Multiple Random Variables

- **Joint PDF** is used to calculate probabilities:

$$\iint_{(x,y)\in b} f_{X,Y}(x,y)\,dx\ dy$$

- **Marginal PDF** of X and Y can be obtained by the joint PDF as:

$$f_{X,Y}(x) = \iint_{(x,y)\in b} f_{X,Y}(x,y)\,dx \qquad f_{X,Y}(y) = \iint_{(x,y)\in b} f_{X,Y}(x,y)\,dy$$

- **Joint CDF** is defined by $f_{X,Y}(x,y) = P(X \leq x, Y \leq y)$,and determines the joint PDF through:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y)$$

  for each $(x,y)$where the joint cdf is continous.

- A function $g(x,y)$ of X and Y defines a new random variable, and

$$E[g(X)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\,dx\ dy$$

- The above have natural extensions to the case where more than two random variables are involved.

3.5 Conditioning

3.5.1 Conditional PDF Given an Event

- The conditional PDF $f_{X|A}$ of a continuous random variable X, given an event A with $P(A) > 0$, satisfies

$$P(X \in B \mid A) = \int_B f_{X|A}(x)\ dx$$

- If A is a subset of the real line with $P(X \in A) > 0$, then

$$f_{X|\{X\in A\}}(x) = \begin{cases} \frac{f_x(X)}{P(X\in A)} & x \in A \\ 0 & \text{otherwise} \end{cases}$$

- Let $A_1, A_2, \cdots, A_n$ be disjoint events that form a partition of the smaple space and assume that $P(A_i) > 0$:

$$f_X(x) = \sum_{i=1}^{n} P(A_i)\, f_{X|A_i}(x)$$

3.5.2 Conditional PDF Given a Random Variable

- The joint, marginal, and conditional PDFs are related to each other by the formulas:

$$f_{X,Y}\left(x,y\right)=f_{Y}\left(y\right)f_{X|Y}\left(x\mid y\right)$$

$$f_{X}\left(x\right)=\int_{-\infty}^{\infty}f_{Y}\left(y\right)f_{X|Y}\left(x\mid y\right)dy$$

3.5.3 Conditional Expectations

- Definitions: The conditional expectation of X given the event A is defined by:

$$E\left[X\mid A\right]=\int_{-\infty}^{\infty}xf_{X|A}\left(x\right)dx$$

  The conditional expectation of X given that Y = y is defined by:

$$E\left[X\mid Y=y\right]=\int_{-\infty}^{\infty}xf_{X|Y}\left(x\mid y\right)dx$$

- The expected value rule: For a function g(X), we have

$$E\left[g\left(X\right)\mid A\right]=E\left[X\mid A\right]=\int_{-\infty}^{\infty}g\left(x\right)f_{X|A}\left(x\right)dx$$

  The conditional expectation of X given that Y = y is defined by:

$$E\left[g\left(X\right)\mid Y=y\right]=\int_{-\infty}^{\infty}g\left(x\right)f_{X|Y}\left(x\mid y\right)dx$$

- Total expectation theorem: Let $A_1, A_2, \cdots, A_n$ be disjoint events that form a partition of the smaple space and assume that $P\left(A_i\right)>0, \forall i$:

$$E\left[X\right]=\sum_{i=1}^{n}P\left(A_i\right)E\left[X\mid A_i\right]$$

  Similarly,

$$E\left[X\right]=\int_{-\infty}^{\infty}E\left[X\mid Y=y\right]f_{Y}\left(y\right)dy$$

3.5.4 Independence of Continuous Random Variables

- X and Y are independent if

$$f_{x,y}\left(x,y\right)=f_{X}\left(x\right)f_{Y}\left(y\right),\forall x,y$$

- If X and Y are independent, then:

$$E\left[XY\right]=E\left[X\right]E\left[Y\right]$$

$$\operatorname{var}\left(X+Y\right)=\operatorname{var}\left(X\right)+\operatorname{var}\left(Y\right)$$

3.6 The Continuous Bayes' Rule

- If X is a continuous random variable, we have

$$f_Y(y) f_{X|Y}(x \mid y) = f_X(x) f_{Y|X}(y \mid x)$$

and

$$f_{X|Y}(x \mid y) = \frac{f_X(x) f_{Y|X}(y \mid x)}{f_Y(y)} = \frac{f_X(x) f_{Y|X}(y \mid x)}{\int_{-\infty}^{\infty} f_X(t) f_{Y|X}(y \mid t)\, dt}$$

- If N is a discrete random variable, we have

$$f_Y(y) P(N = n \mid Y = y) = \rho_N(n) f_{Y|N}(y \mid n)$$

resulting in:

$$P(N = n \mid Y = y) = \frac{\rho_N(n) f_{Y|N}(y \mid n)}{f_Y(y)} = \frac{\rho_N(n) f_{Y|N}(y \mid n)}{\sum_i \rho_N(i) f_{Y|N}(y \mid i)}$$

- There are similar formulas for $P(A \mid Y = y)$ and $f_{Y|A}(y)$.

3.7 Covariance and Correlation

3.7.1 Covariance

The covariance of two random variable X and Y, denoted by $\operatorname{cov}(X, Y)$,is defined by:

$$\operatorname{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

When $\operatorname{cov}(X, Y) = 0$, two functions are **uncorrelated**. If X and Y are independent, they are uncorrelated. We also have:

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y)$$

3.7.2 Correlation coefficent

The correlation coefficent $\rho(X, Y)$ of two random variables X and Y that have nonzero vairances is defiend as:

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}, \qquad -1 \leq \rho(X, Y) \leq 1$$

4 Limit Theroem

4.1 Markov and Chebyshev Inequalities

4.1.1 Markov Inequality

$$P(X \geq a) \leq \frac{E[X]}{a}, \forall a > 0$$

4.1.2 Chebyshev Inequality

$$P\left(|X-\mu|\geq c\right)\leq\frac{\sigma^2}{c^2},\forall c>0$$

4.2 Central Limit Theorem

4.2.1 Central Limit Theorem

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent identically distributed random variables with common mean μ and variance σ^2:

$$Z_n=\frac{X_1,X_2,\cdots,X_n-n\mu}{\sigma\sqrt{n}}$$

Then, the CDF of $Z_n$converges to standard normal CDF:

$$\Phi\left(z\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{x^2}{2}}\ dx$$

in the sense that:

$$\lim_{n\to\infty}P\left(Z_n\leq z\right)=\Phi\left(z\right),\forall z$$

4.2.2 Normal Approximation Based on the Central Limit Theorem

Let $S_n = X_1+\cdots+X_n$, where the $X_i$are independent identically distributed random variables with mean μ and variance σ^2. If n is large, the probability$P\left(S_n\leq c\right)$ can be approximated by treating $S_n$as normal through:

1. calculate the mean $n\mu$ and the variance $n\sigma^2$ of S_n.
2. calculate the normalized z-value $z=\frac{(c-n\mu)}{\sigma\sqrt{n}}$
3. Use approximation
$$P\left(S_n\leq c\right)\approx\Phi\left(z\right)$$

4.2.3 De Moivre-Laplace Approximation to the Binomial

If S_n is a binomial random variable with parameters n and p, n is large and k, l are nonnegative integers, then:

$$P\left(k\leq S_n\leq l\right)=\Phi\left(\frac{l+\frac{1}{2}-np}{\sqrt{np\left(1-p\right)}}\right)-\Phi\left(\frac{k-\frac{1}{2}-np}{\sqrt{np\left(1-p\right)}}\right)$$

4.3 Laws of Large Numbers

4.3.1 Weak Laws of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be independent identically distributed random variables with common mean μ. For every $\epsilon > 0$:

$$P\left(|M_n - \mu| \geq \epsilon\right) \to 0, \text{ as } n \to \infty$$

4.3.2 Strong Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent identically distributed random variables with common mean μ. Then, the sequence of sample means $M_n = \frac{(X_1 + \cdots + X_n)}{n}$ converges to μ:

$$P\left(\lim_{n\to\infty} \frac{(X_1 + \cdots + X_n)}{n} = \mu\right) = 1$$

4.3.3 Convergence with Probability

Let $X_1, X_2, \cdots, X_n$ be a sequence of random variables. Let c be a real number. We say that Y_n converges to c with a probability 1 if:

$$P\left(\lim_{n\to\infty} Y_n = c\right) = 1$$