ECE 469/ECE 568 - Machine Learning

September 18, 2024

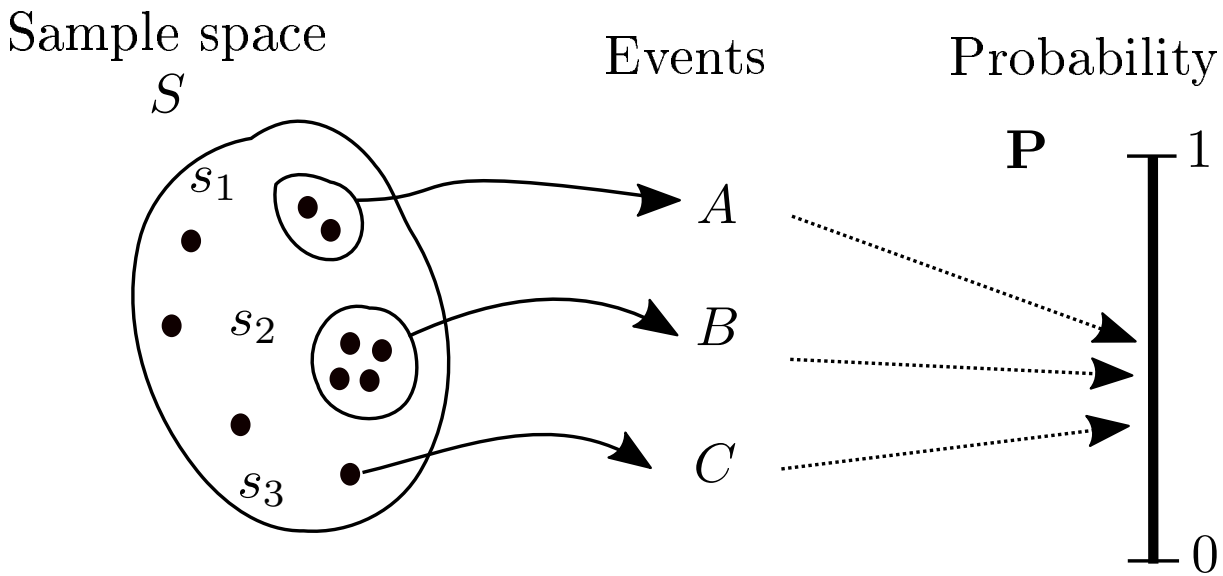
Introduction to probability theory

- Probability can be defined based on the relative frequency of occurrence. In n trials of a random experiment, if we expect an event A to occur m times, then we assign the probability m/n to event A.
- Assume that a random experiment has K possible outcomes, then for the kth possible outcome, we have a point called the "sample point" denoted as s_k. With this, we formulate the following framework for probability.

    1. The set of all possible outcomes of the experiment is called the sample space denoted by S.
    2. An event corresponds to either a single sample point or a set of sample points in the space S.
    3. A single sample point is called an elementary event.
    4. The entire sample space S is called the sure event; and the null set ϕ is called the null/impossible event.
    5. Two events are mutually exclusive if the occurrence of one event precludes the occurrence of the other event.

Probability Axioms

- A probability measure $\mathbf{P}$ is a function that assigns a non-negative number to an event A in the sample space S and satisfies the following three axioms (properties):
  1. $0 \leq \mathbf{P}(A) \leq 1$
  2. $\mathbf{P}(S) = 1$
  3. If A and B are two mutually exclusive events ($A \cap B = \phi$), then $\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B]$
- This abstract definition of a probability system can be illustrated as



Properties of probability measure

- The following properties of the probability measure can be derived from the three main axioms.
  a. $\mathbf{P}[\bar{A}] = 1 - \mathbf{P}[A]$, where $\bar{A}$ is the complement of A.
  b. When events A and B are not mutually exclusive, then $\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B] - \mathbf{P}[A \cap B]$, where $\mathbf{P}[A \cap B]$ is the probability of the joint event A and B.
  c. If $A_1, A_2, \cdots, A_m$ are mutually exclusive events that include all possible outcomes of the random experiment, then $\mathbf{P}[A_1] + \mathbf{P}[A_2] + \cdots + \mathbf{P}[A_m] = 1$.

Conditional Probability

- Let $\mathbf{P}[B|A]$ denote the probability of event B given that event A has occurred.The probability $\mathbf{P}[B|A]$ is called the conditional probability of B given A.
- Assuming that the event A has non-zero probability, the conditional probability $\mathbf{P}[B|A]$ is defined as

$$\mathbf{P}[B|A] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[A]},$$

  where $\mathbf{P}[A \cap B]$ is the joint probability of "A and B".
- Therefore, the joint probability of A and B can be written as

$$\mathbf{P}[A \cap B] = \mathbf{P}[B|A]\mathbf{P}[A] \qquad \text{and} \qquad \mathbf{P}[A \cap B] = \mathbf{P}[A|B]\mathbf{P}[B]$$

- Bayes' rule:

$$\mathbf{P}[B|A] = \frac{\mathbf{P}[A|B]\mathbf{P}[B]}{\mathbf{P}[A]}$$

Statistical independence

- Suppose that the conditional probability $\mathbf{P}[B|A]$ is simply equal to $P[B]$;

$$\mathbf{P}[B|A] = \mathbf{P}[B]$$

- Under this condition, the probability of the joint even "A and B" is equal to the product of the probabilities of individual events.

$$\mathbf{P}[A \cap B] = \mathbf{P}[A]\mathbf{P}[B]$$

- This follows that $\mathbf{P}[A|B] = \mathbf{P}[A]$.
- In this case, the knowledge of the occurrence of one event tells no more about the probability of the occurrence of the other event than we knew without that knowledge.
- Events that satisfy this condition are said to be statistically independent.

Bayes' theorem and law of total probability

- Let the set $A_1, A_2, \cdots, A_n$ be a partition of the sample space Ω.
- Then they are a set of mutually exclusive events in Ω.

$$A_1 \cup A_2 \cup \cdots \cup A_n = \Omega \quad \text{and} \quad A_i \cap A_j = \emptyset \ \text{ for } i \neq j$$

- For any event B, we have

$$\bigcup_{j=1}^{n} \{A_j \cap B\} = B$$

- Hence, we can write

$$\sum_{j=1}^{n} P(A_j \cap B) = \sum_{j=1}^{n} P(A_j) P(B|A_j) = P(B)$$

- This leads to the law of total probability

$$\boxed{P(B) = \sum_{j=1}^{n} P(A_j) P(B|A_j)}$$

Bayes' theorem and law of total probability

- Let the set $A_1, A_2, \cdots, A_n$ be a partition of the sample space Ω. Let B be an event in a sample space Ω.
- Then, we have

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j)P(B|A_j)}{P(B)}$$

- From the law of total probability, we have

$$P(B) = \sum_{j=1}^{n} P(A_j)P(B|A_j)$$

- Finally, the Bayes' theorem can be written as

$$\boxed{P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{j=1}^{n} P(A_j)P(B|A_j)}}$$

Bayes' theorem and law of total probability

- Recall the Bayes' theorem

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{j=1}^{n} P(A_j)P(B|A_j)}$$

- In the Bayes' theorem, the probability $P[A_j]$ is called the prior probability (or a-priori probability) of event A_j (before event B occurs).
- The conditional probability $P[A_j|B]$ is called the posterior probability (or a-posteriori probability) of event A_j after event B occurs.

Bayesian probabilities

- Bayesian approach takes a subjective view of probability and believe that the notion of probability is applicable to any situation or event for which we attach some uncertainty.
- Bayesian view of probability is useful to "learning theory" from the probabilistic point of view.
- The Bayes' theorem provides the fundamental principle of learning based on data.

Histogram

- A histogram represents a frequency distribution by means of rectangles whose widths represent class intervals and whose areas are proportional to the corresponding frequencies.
- The purpose of a histogram is to roughly assess the probability distribution of a given variable by depicting the frequencies of observations occurring in certain ranges of values.

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Histogram

Sort raw data in ascending order:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- Find range: 58 - 12 = 46
- Select number of classes: 5 (usually between 5 and 20)
- Compute class width: 10 (46/5 then round off)
- Determine class boundaries:10, 20, 30, 40, 50
- Compute class midpoints: 15, 25, 35, 45, 55
- Count observations & assign to classes

Histogram

Data in ordered array:

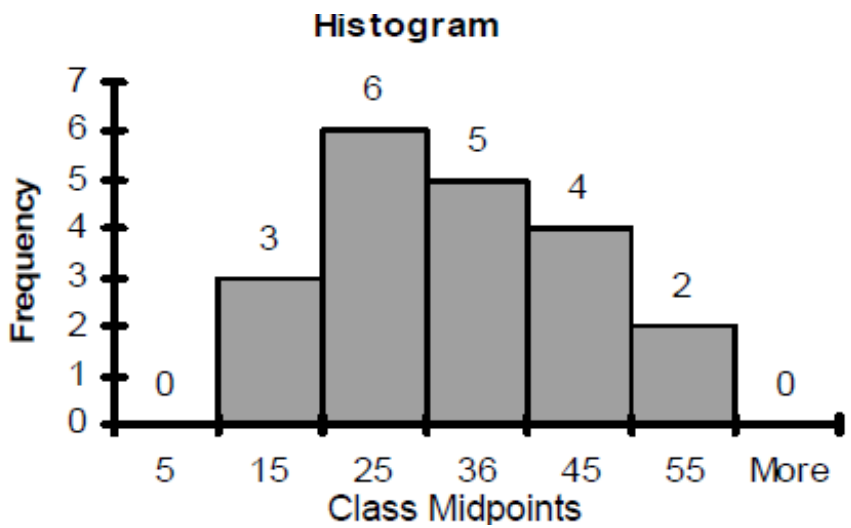
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Frequency Distribution

Class	Frequency	Relative Frequency
10 but under 20	3	.15
20 but under 30	6	.30
30 but under 40	5	.25
40 but under 50	4	.20
50 but under 60	2	.10
Total	20	1.00

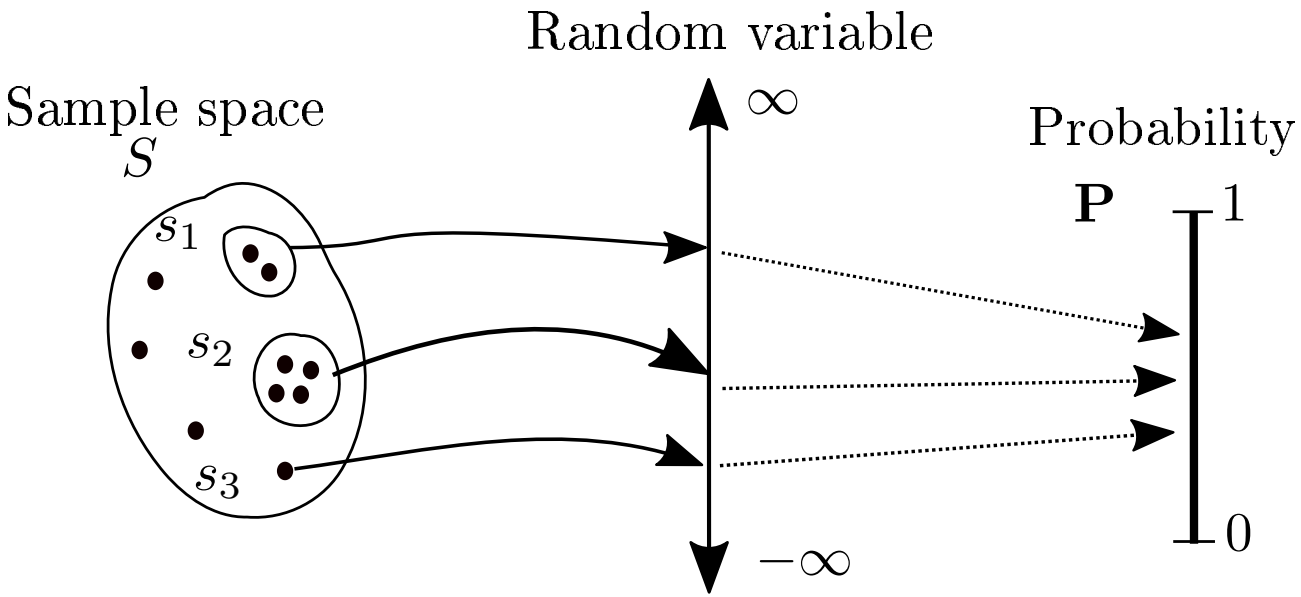
Histogram

- The **classes** or **intervals** are shown on the horizontal axis
- **frequency** is measured on the vertical axis
- Bars of the appropriate heights can be used to represent the number of observations within each class
- Such a graph is called a **histogram**



Random variables

- We can assign a number or a range of values to the outcomes of a random experiment. For example, in a coin-flip, a head could corresponds to 1 and a tail to 0. We use the expression random variable to describe this process of assigning a number to the outcome of a random experiment.



Random variables Continued...

Consider the random variable X and the probability of the event $X \leq x$ where x is dummy-variable. The probability of the event $X \leq x$ is called the cumulative distribution function (CDF) of X.

Cumulative distribution function (CDF): $F_X(x) = \mathbf{P}[X \leq x]$

1. The CDF $F_X(x)$ is bounded between zero and one.
2. The CDF $F_X(x)$ is a monotone-nondecreasing function of x.

$$F_X(x_1) \leq F_X(x_2) \;\; if \;\; x_1 < x_2.$$

Random variables Continued...

3. If the random variable is continuous, the derivative of the CDF is called the probability density function (PDF).

$$f_X(x) = \frac{d}{dx} F_X(x) \qquad \Longrightarrow \qquad F_X(x) = \mathbf{P}[X \leq x] = \int_{-\infty}^{x} f_X(y) dy$$

4. The name density function arises due to the fact that the probability of the event $x_1 \leq X \leq x_2$ equals to

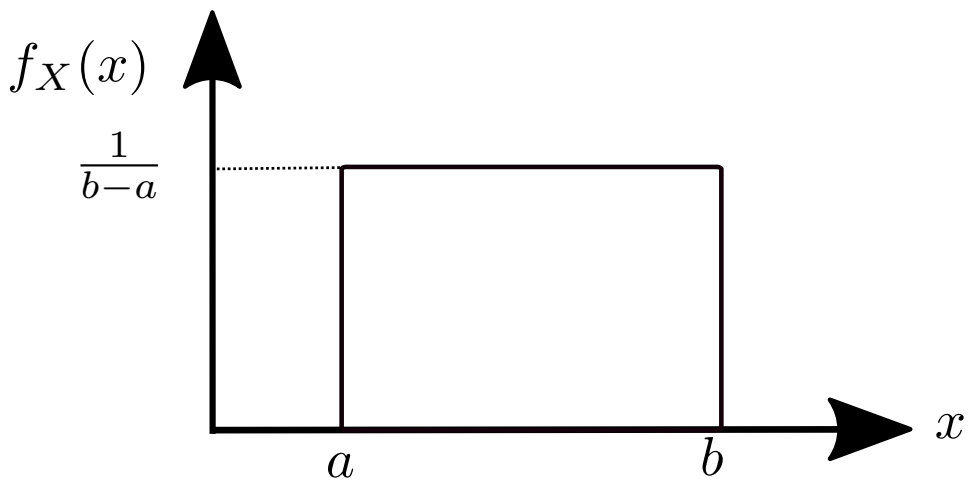
$$\begin{aligned} \mathbf{P}[x_1 \leq X \leq x_2] &= \mathbf{P}[X \leq x_2] - \mathbf{P}[X \leq x_1] \\ &= F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x) dx. \end{aligned}$$

5. Total area under PDF curve is one: $\int_{-\infty}^{\infty} f_X(y) dy = 1$

Example: Uniform Random variables

- A random variable X is said to be uniformly distributed over the interval (a, b) if its probability its PDF is given by

$$f_X(x) = \begin{cases} 0, & x \leq a \\ \frac{1}{b-a}, & a < X \leq b \\ 0, & X > b \end{cases}$$

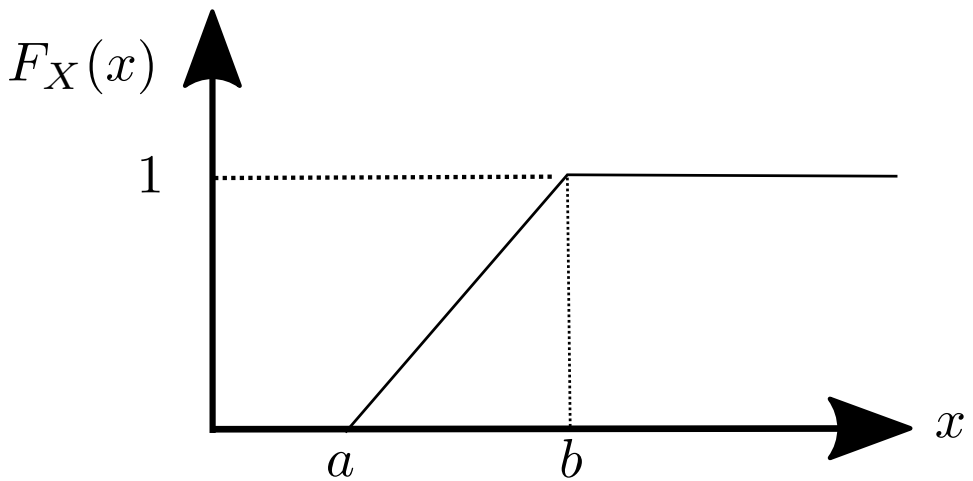


- Determine the CDF of the uniform random variable X.

Example: Uniform Random variables

- Determine the CDF of the uniform random variable X.
- The CDF of X is given by

$$F_X(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & X > b \end{cases}$$



Statistical Averages

- The expected value or mean of a random variable X is defined as

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

- Let $g(X)$ be a real-valued function defined on the real line. Then $Y = g(X)$ is also a random variable.

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy \qquad \Longrightarrow \qquad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Statistical Averages

- Example: Let $Y = g(X) = \cos(X)$, where X is a random variable defined in the interval $(-\pi, \pi)$:

$$f_X(x) = \begin{cases} \frac{1}{2\pi}, & -\pi < x < \pi \\ 0, & \text{otherwise} \end{cases}$$

Determine the expected value of Y.

$$\begin{aligned} \mathbb{E}[Y] &= \int_{-\pi}^{\pi} \cos(x) \times \frac{1}{2\pi} dx \\ &= \left(-\frac{1}{2\pi}\right) \times \sin(x)\Big|_{x=-\pi}^{\pi} = 0 \end{aligned}$$

Moments

- The nth moment of X can be written as follows:

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

- The first two moments are the most important moments of X.

$$\boxed{\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \qquad \text{and} \qquad \mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx}$$

- The second central moment of X is referred to as the variance of the random variable X;

$$\boxed{Var[X] = \sigma_X^2 = \mathbb{E}[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx}$$

- The variance of X can also be expressed as

$$\begin{aligned} \sigma_X^2 &= \mathbb{E}[(X - \mu_X)^2] \\ &= \mathbb{E}[X^2 - 2\mu_X X + \mu_X^2] \\ &= \mathbb{E}[X^2] - 2\mu_X \mathbb{E}[X] + \mu_X^2 = \mathbb{E}[X^2] - 2\mu_X^2 + \mu_X^2 \\ &= \mathbb{E}[X^2] - \mu_X^2 \end{aligned}$$