CPEN 455 26W1 Tutorial

Probability and Statistics[1]

Created by: Qi Yan and Felix Fu

Presented by: Felix Fu

University of British Columbia

September 14th, 2026

[1] Based on materials from Goodfellow et al., 2016; Murphy, 2022; Zhang et al., 2023.

Contents

PROBABILITY

Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

- Probability: quantitative degree of belief.
  - an image classifier outputs a probability distribution given an input image.
  - a large language model (e.g., chatGPT) outputs a probability distribution over the next word when making predictions.
  - a generative model (VAEs/GANs/diffusion) starts generation from known prior distributions.

Probability

Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

- Probability: quantitative degree of belief.
  - an image classifier outputs a probability distribution given an input image.
  - a large language model (e.g., chatGPT) outputs a probability distribution over the next word when making predictions.
  - a generative model (VAEs/GANs/diffusion) starts generation from known prior distributions.
- **Frequentist** perspective: probabilities represent long run frequencies of **events** that can happen multiple times.
  - if we toss the coin many times, we expect it to land heads about half the time.
- **Bayesian** perspective: probability is used to quantify our **uncertainty** or ignorance about something; hence it is fundamentally related to information rather than repeated trials
  - we believe the coin is equally likely to land heads or tails on the next toss.

PROBABILITY

- **Probability Space**:
  - A mathematical model of a random experiment.

PROBABILITY

- **Probability Space**:
  - A mathematical model of a random experiment.
- **Sample Space**:
  - The collection of all possible outcomes of the experiment.

PROBABILITY

- **Probability Space**:
  - A mathematical model of a random experiment.
- **Sample Space**:
  - The collection of all possible outcomes of the experiment.
- **Events**:
  - Subsets of the sample space.
  - An event *occurs* if the experiment's outcome belongs to that subset.
  - Events are those subsets whose occurrence is noteworthy and decidable.

PROBABILITY

- **Probability Space**:
  - A mathematical model of a random experiment.
- **Sample Space**:
  - The collection of all possible outcomes of the experiment.
- **Events**:
  - Subsets of the sample space.
  - An event *occurs* if the experiment's outcome belongs to that subset.
  - Events are those subsets whose occurrence is noteworthy and decidable.
- **Axioms of Probability**:
  - **Non-negativit**y: the probability of an event is a non-negative real number.
  - **Unit measure**: the probability of the event representing the entire sample space occurs is 1.
  - σ**-additivity**: any countable sequence of disjoint sets (mutually exclusive events) satisfies the fact that the probability of their union equals to the sum of their individual probabilities.

Quick Question 1: Events

A fair six-sided die is rolled once. Let

$$A = \{2, 4, 6\}, \qquad B = \{4, 5, 6\}.$$

What are $P(A \cap B)$ and $P(A \cup B)$?

Quick Question 1: Events

A fair six-sided die is rolled once. Let

$$A = \{2, 4, 6\}, \qquad B = \{4, 5, 6\}.$$

What are $P(A \cap B)$ and $P(A \cup B)$?

Answer

The intersection is $A \cap B = \{4, 6\}$, so $P(A \cap B) = 2/6 = 1/3$.
The union is $A \cup B = \{2, 4, 5, 6\}$, so

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{2}{3}.$$

Adding $P(A)$ and $P(B)$ counts the overlap twice.

Random Variables

- **Random Variable**:
  - A variable that maps events with potential various random values.
  - Distinct from its possible values.

Random Variables

- **Random Variable**:
  - A variable that maps events with potential various random values.
  - Distinct from its possible values.
- **Nature of Random Variables**:
  - A quantitative description of possible states. to define likelihood of each state.

Random Variables

- **Random Variable**:
  - A variable that maps events with potential various random values.
  - Distinct from its possible values.
- **Nature of Random Variables**:
  - A quantitative description of possible states. to define likelihood of each state.
- **Types of Random Variables**:
  - *Discrete Random Variable*:
    - Finite or countably infinite states.
    - States could be numerical or non-numerical.
  - *Continuous Random Variable*:
    - Associated with real number values.

Discrete Random Variables

Described using a **Probability Mass Function (PMF)**, typically denoted by P. Associates each possible state with a probability.

PMF Characteristics

- Maps a state of a random variable to its probability of occurrence.
- PMFs can describe joint probabilities for multiple variables, e.g., $P(\mathrm{x}, \mathrm{y})$.
- Must satisfy two conditions:
    1. $0 \leq P(\mathrm{x}) \leq 1$ for all states x.
    2. $\sum_{\mathrm{x}} P(\mathrm{x}) = 1$ (Normalization).

Discrete Random Variables

Described using a **Probability Mass Function (PMF)**, typically denoted by P. Associates each possible state with a probability.

PMF Characteristics

- Maps a state of a random variable to its probability of occurrence.
- PMFs can describe joint probabilities for multiple variables, e.g., $P(\mathrm{x}, \mathrm{y})$.
- Must satisfy two conditions:
  1. $0 \leq P(\mathrm{x}) \leq 1$ for all states x.
  2. $\sum_{\mathrm{x}} P(\mathrm{x}) = 1$ (Normalization).

Example 2.1 (Uniform Distribution PMF)

Given a discrete random variable x with k states, a uniform distribution assigns:

$$P(\mathrm{x} = x_i) = \frac{1}{k}$$

This satisfies the normalization condition since $\sum_i \frac{1}{k} = 1$.

Quick Question 2: A Valid PMF

A random variable X takes values $0, 1, 2$, with

$$P(X = 0) = 0.2, \qquad P(X = 1) = c, \qquad P(X = 2) = 0.5.$$

What must c be? What is $P(X \geq 1)$?

Quick Question 2: A Valid PMF

A random variable X takes values $0, 1, 2$, with

$$P(X = 0) = 0.2, \qquad P(X = 1) = c, \qquad P(X = 2) = 0.5.$$

What must c be? What is $P(X \geq 1)$?

Answer

A PMF sums to one:

$$c = 1 - 0.2 - 0.5 = 0.3.$$

The events $X = 1$ and $X = 2$ are disjoint, so

$$P(X \geq 1) = 0.3 + 0.5 = 0.8.$$

Random Variables

Continuous Variables and PDFs

Continuous Random Variables

Probability Density Functions (PDFs) p describe probability distributions for continuous variables:

- Domain is all possible states of x.
- $p(\mathrm{x}) \geq 0$ for all x in the domain.
- $\int p(\mathrm{x})\, d\mathrm{x} = 1$ (Total probability is 1).

Understanding PDFs

- PDFs give the probability density, not the probability of specific states.
- Probability for an infinitesimal region is $p(\mathrm{x})\, d\mathrm{x}$.
- The probability that x is in a set S is the integral of $p(\mathrm{x})$ over S.

Continuous Random Variables

Probability Density Functions (PDFs) p describe probability distributions for continuous variables:

- Domain is all possible states of x.
- $p(\mathrm{x}) \geq 0$ for all x in the domain.
- $\int p(\mathrm{x})\, d\mathrm{x} = 1$ (Total probability is 1).

Understanding PDFs

- PDFs give the probability density, not the probability of specific states.
- Probability for an infinitesimal region is $p(\mathrm{x})\, d\mathrm{x}$.
- The probability that x is in a set S is the integral of $p(\mathrm{x})$ over S.

Example 2.2 (Uniform Distribution PDF)

Consider $u(\mathrm{x}; a, b)$ for a uniform distribution over $[a, b]$, where $a < b$:

$$u(\mathrm{x}; a, b) = \begin{cases} 0 & \text{for } \mathrm{x} \notin [a, b] \\ \frac{1}{b-a} & \text{for } \mathrm{x} \in [a, b] \end{cases}$$

This function is always nonnegative and integrates to 1, representing the uniform distribution $\mathrm{x} \sim U(a, b)$.

Distribution and Density

Probability Distribution

The distribution of X assigns a probability to each event $X \in S$. For a real-valued X, the **cumulative distribution function (CDF)**

$$F_X(x) = P(X \leq x)$$

fully describes its distribution, whether discrete or continuous.

PMF and PDF

A discrete distribution has a **PMF**: $p_X(x) = P(X = x)$.
An absolutely continuous distribution has a **density (PDF)** f_X:

$$P(a < X \leq b) = \int_a^b f_X(t)\, dt = F_X(b) - F_X(a).$$

The CDF accumulates probability. The PDF gives probability per unit of x.

A density can exceed 1. Its total area is 1, and $P(X = x) = 0$. Not every distribution has a density.

Probability Distributions

Marginal Probability

Concept of Marginal Probability

The probability distribution over a subset of variables, derived from a joint distribution of multiple variables.

Discrete Random Variables

Given discrete random variables x and y, and the joint distribution $P(\mathrm{x}, \mathrm{y})$, the marginal probability $P(\mathrm{x})$ is calculated as:

$$\forall \mathrm{x} \in X, P(\mathrm{x} = x) = \sum_{y \in Y} P(\mathrm{x} = x, \mathrm{y} = y)$$

Continuous Random Variables

For continuous variables, marginal probability is found using integration:

$$p(\mathrm{x}) = \int p(\mathrm{x}, \mathrm{y}) \, d\mathrm{y}$$

Origin of the Term

The term "marginal" refers to the practice of summing probabilities in a table and writing the totals in the margins.

Definition

The probability of an event given that another event has occurred, denoted as $P(\mathrm{y} = y \mid \mathrm{x} = x)$.

Computation Formula

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

Note: Defined only if $P(\mathrm{x} = x) > 0$.

Understanding Conditional Probability

- Not to be confused with the consequences of actions or interventions.
- The conditional probability of an event y given x is different from the probability that y would happen if x were to be caused by some action (called intervention query for causal modeling).

Chain Rule

A joint probability distribution can be decomposed into conditional probabilities:

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \prod_{i=2}^{n} P(\mathrm{x}^{(i)} | \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

Application of the Chain Rule

- Simplifies the computation of joint distributions.
- Derived from the definition of conditional probability.

Chain Rule

A joint probability distribution can be decomposed into conditional probabilities:

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \prod_{i=2}^{n} P(\mathrm{x}^{(i)} | \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

Application of the Chain Rule

- Simplifies the computation of joint distributions.
- Derived from the definition of conditional probability.

Example 3.1

Applying the rule to three variables a, b, and c:

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a}|\mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b}|\mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a}|\mathrm{b}, \mathrm{c}) P(\mathrm{b}|\mathrm{c}) P(\mathrm{c}) \end{aligned}$$

Independence of Random Variables

Two random variables x and y are **independent** if:

$$\forall x \in X, y \in Y, \quad p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$$

Conditional Independence

x and y are **conditionally independent** given z if:

$$\forall x \in X, y \in Y, z \in Z, \quad p(\mathrm{x} = x, \mathrm{y} = y | \mathrm{z} = z) = p(\mathrm{x} = x | \mathrm{z} = z)p(\mathrm{y} = y | \mathrm{z} = z)$$

Notation

Independence is denoted by $\mathrm{x} \perp \mathrm{y}$, while conditional independence given z is denoted by $\mathrm{x} \perp \mathrm{y} | \mathrm{z}$.

Quick Question 3: Independence

For one fair die roll, let $A = \{1, 2\}$ and $B = \{5, 6\}$. The events are mutually exclusive. Are they independent?

Quick Question 3: Independence

For one fair die roll, let $A = \{1, 2\}$ and $B = \{5, 6\}$. The events are mutually exclusive. Are they independent?

Answer

No. They cannot happen together:

$$P(A \cap B) = 0, \qquad P(A)P(B) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}.$$

These quantities differ, so independence fails. Knowing that A occurred makes B impossible: $P(B \mid A) = 0$, whereas $P(B) = 1/3$.

Definition

The expectation or expected value of a function $f(\mathrm{x})$ with respect to a distribution $P(\mathrm{x})$ is the mean value f takes when x is drawn from P.

Computation

For discrete variables:

$$\mathbb{E}_{\mathrm{x}\sim P}[f(\mathrm{x})] = \sum_{\mathrm{x}} P(\mathrm{x})f(\mathrm{x}),$$

For continuous variables:

$$\mathbb{E}_{\mathrm{x}\sim P}[f(\mathrm{x})] = \int p(\mathrm{x})f(\mathrm{x})d\mathrm{x}.$$

Properties of Expectation

- Expectations are linear:

$$\mathbb{E}_{\mathrm{x}}[\alpha f(\mathrm{x}) + \beta g(\mathrm{x})] = \alpha\mathbb{E}_{\mathrm{x}}[f(\mathrm{x})] + \beta\mathbb{E}_{\mathrm{x}}[g(\mathrm{x})],$$

  where α and β are constants.
- Notation can be simplified to $\mathbb{E}[f(\mathrm{x})]$ when the context is clear.

Variance

The measure of the spread of a function $f(\mathrm{x})$ of a random variable:

$$\mathrm{Var}(f(\mathrm{x})) = \mathbb{E}[(f(\mathrm{x}) - \mathbb{E}[f(\mathrm{x})])^2]$$

Standard deviation is the square root of the variance.

Covariance

Indicates how two variables $f(\mathrm{x})$ and $g(\mathrm{y})$ linearly relate to each other:

$$\mathrm{Cov}(f(\mathrm{x}), g(\mathrm{y})) = \mathbb{E}[(f(\mathrm{x}) - \mathbb{E}[f(\mathrm{x})])(g(\mathrm{y}) - \mathbb{E}[g(\mathrm{y})])]$$

Positive (negative) covariance implies a positive (negative) linear relationship.

Variance

The measure of the spread of a function $f(\mathrm{x})$ of a random variable:

$$\mathrm{Var}(f(\mathrm{x})) = \mathbb{E}[(f(\mathrm{x}) - \mathbb{E}[f(\mathrm{x})])^2]$$

Standard deviation is the square root of the variance.

Covariance

Indicates how two variables $f(\mathrm{x})$ and $g(\mathrm{y})$ linearly relate to each other:

$$\mathrm{Cov}(f(\mathrm{x}), g(\mathrm{y})) = \mathbb{E}[(f(\mathrm{x}) - \mathbb{E}[f(\mathrm{x})])(g(\mathrm{y}) - \mathbb{E}[g(\mathrm{y})])]$$

Positive (negative) covariance implies a positive (negative) linear relationship.

Correlation

Correlation normalizes covariance to measure the strength of the linear relationship (scale-invariant).

Independence v.s. Covariance

Independence implies zero covariance but not vice versa. Zero covariance does not imply independence.

Covariance Matrix

For a random vector $\mathbf{x} \in \mathbb{R}^n$, the covariance matrix is $n \times n$ with:

$$\mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j) = \mathrm{Var}(\mathrm{x}_i, \mathrm{x}_j)$$

Diagonal elements represent the variance.

Quick Question 4: Expectation and Variance

Let X be 1 with probability p and 0 with probability $1 - p$. What are $\mathbb{E}[X]$ and $\mathrm{Var}(X)$?

QUICK QUESTION 4: EXPECTATION AND VARIANCE

Let X be 1 with probability p and 0 with probability $1 - p$. What are $\mathbb{E}[X]$ and $\mathrm{Var}(X)$?

Answer

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Since $X^2 = X$, we also have $\mathbb{E}[X^2] = p$. Therefore,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p).$$

For a fair coin, $p = 1/2$, so the mean is $1/2$ and the variance is $1/4$.

A Bit of Information Theory

Entropy

Definition

Entropy, denoted as $H(\mathrm{X})$, is a measure of the uncertainty or unpredictability in a random variable X.

Shannon Entropy

For a discrete random variable X with possible values $\{x_1, x_2, \ldots, x_n\}$ and probability mass function $P(\mathrm{X})$, entropy is defined as:

$$H(\mathrm{X}) = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(\mathrm{x})] = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

where the logarithm is base 2 for bits.

A Bit of Information Theory

Entropy

Definition

Entropy, denoted as $H(\mathrm{X})$, is a measure of the uncertainty or unpredictability in a random variable X.

Shannon Entropy

For a discrete random variable X with possible values $\{x_1, x_2, \ldots, x_n\}$ and probability mass function $P(\mathrm{X})$, entropy is defined as:

$$H(\mathrm{X}) = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(\mathrm{x})] = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

where the logarithm is base 2 for bits.

Interpretation

Entropy quantifies the expected amount of information conveyed by identifying the outcome of X. Higher entropy implies a more uncertain outcome, while lower entropy implies a more predictable outcome.

Properties of Entropy

- $H(\mathrm{X}) \geq 0$ for all X.
- $H(\mathrm{X}) = 0$ if and only if one outcome has a probability of 1 (no uncertainty).

Kullback-Leibler (KL) Divergence

Measures how one probability distribution P diverges from a second, reference probability distribution Q:

$$D_{\mathrm{KL}}(P \| Q) = \mathbb{E}_{\mathrm{x} \sim P} \left[\log \frac{P(\mathrm{x})}{Q(\mathrm{x})} \right]$$

It represents the extra amount of information needed to code samples from P using a code optimized for Q.

Properties of KL Divergence

- Non-negative: $D_{\mathrm{KL}}(P \| Q) \geq 0$
- Zero if and only if P and Q are the same distribution.
- Non-symmetric: $D_{\mathrm{KL}}(P \| Q) \neq D_{\mathrm{KL}}(Q \| P)$

Cross-Entropy

Related to KL divergence, cross-entropy combines the entropy of P with the KL divergence between P and Q:

$$H(P, Q) = H(P) + D_{\mathrm{KL}}(P \| Q)$$

Interpretation

Minimizing cross-entropy with respect to Q equates to minimizing the KL divergence, often used in optimization problems such as training machine learning models.

Example 5.1 (Cross-Entropy and KL Divergence in Image Classification)

In image classification, a model predicts a probability distribution Q over classes for a given image, while the true distribution P is typically one-hot encoded (1 for the correct class, 0 for others).
Cross-entropy in this context measures the difference between the distributions P and Q:

$$H(P, Q) = -\sum_{\mathrm{c}} P(\mathrm{c}) \log Q(\mathrm{c})$$

where c indexes over the classes.

Connection to KL Divergence

Cross-entropy decomposes into the sum of the true distribution's entropy and the KL divergence:

$$H(P, Q) = H(P) + D_{\mathrm{KL}}(P \| Q)$$

Since $H(P)$ is constant (the true label is fixed), minimizing cross-entropy $H(P, Q)$ with respect to Q is equivalent to minimizing $D_{\mathrm{KL}}(P \| Q)$.
During training, optimizing cross-entropy encourages the model to make predictions Q that match the true distribution P, effectively reducing the divergence from the true label distribution.

References

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). ***Deep learning.*** MIT press.
- Murphy, K. P. (2022). ***Probabilistic machine learning: An introduction.*** MIT Press.
- Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). ***Dive into deep learning.*** Cambridge University Press.