COMP 551 - Applied Machine Learning
A Brief Tutorial on Probability Theory

Safa Alver
(slides adapted from Stanford's CS229 Probability Theory Review)

McGill University

January 16, 2021

About this tutorial

- This is not an comprehensive review of Probability Theory.

About this tutorial

- This is not an comprehensive review of Probability Theory.
- The focus is on the subset related to COMP 551.

About this tutorial

- This is not an comprehensive review of Probability Theory.
- The focus is on the subset related to COMP 551.
- More references can be found at the end of the slides.

About this tutorial

- This is not an comprehensive review of Probability Theory.
- The focus is on the subset related to COMP 551.
- More references can be found at the end of the slides.
- Also please shoot me an email if you find any typos or mistakes!

Outline

Outline

Why Probability Theory?

- There are lots of uncertainty in the world.

Why Probability Theory?

- There are lots of uncertainty in the world.
- Probability theory provides a consistent framework for quantification and manipulation of uncertainty.

Sample Space

- **Sample space**, denoted as Ω, is the set of all possible outcomes of an experiment.

Sample Space

- **Sample space**, denoted as Ω, is the set of all possible outcomes of an experiment.
- **Observations**, denoted as $\omega \in \Omega$, are points in the sample space. They are also called sampled outcomes.

Sample Space

- **Sample space**, denoted as Ω, is the set of all possible outcomes of an experiment.
- **Observations**, denoted as $\omega \in \Omega$, are points in the sample space. They are also called sampled outcomes.
- **Events**, denoted as $A \in \Omega$, are subsets of the sample space. Note that a set of events is also an event.

Sample Space

- **Sample space**, denoted as Ω, is the set of all possible outcomes of an experiment.
- **Observations**, denoted as $\omega \in \Omega$, are points in the sample space. They are also called sampled outcomes.
- **Events**, denoted as $A \in \Omega$, are subsets of the sample space. Note that a set of events is also an event.
- Example: Consider the experiment of flipping a coin twice:
  - $\Omega = \{HH, HT, TH, TT\}$.
  - The sampled outcomes can be $\omega = HH$ or $\omega = HT$.
  - One possible event is $A = \{HH, TT\}$, which corresponds to the event of both flips being the same.

Axioms of Probability Theory

- A probability measure is a function that maps events to the interval $[0, 1]$, i.e. $P : A \rightarrow [0, 1]$.

Axioms of Probability Theory

- A probability measure is a function that maps events to the interval $[0, 1]$, i.e. $P : A \to [0, 1]$.
- The probability measure satisfies the following properties:
  1. $P(A) \geq 0$ for all A.
  2. $P(\Omega) = 1$.
  3. If $A_1, A_2, \ldots$ are disjoint events ($A_i \cap A_j = \emptyset$), then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$

These three properties are called the **axioms of probability theory**.

Properties of the Probability Measure

- Properties:
  - If $A \subseteq B$, then $P(A) \leq P(B)$.
  - $P(A \cap B) \leq \min(P(A), P(B))$.
  - **Union Bound**:

$$P(A \cup B) \leq P(A) + P(B)$$

.

  - **Law of Total Probability**: If $A_1, \ldots, A_k$ are disjoint events such that $\bigcup_{i=1}^{k} A_i = \Omega$, then

$$\sum_{i=1}^{k} P(A_i) = 1.$$

  - For more see page 1 of CS229's Probability Theory Review.

Joint and Conditional Probabilities

- **Joint probability** of events A and B, denoted as $P(A, B)$, is the probability that both events will occur

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

Joint and Conditional Probabilities

- **Joint probability** of events A and B, denoted as $P(A, B)$, is the probability that both events will occur

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

- **Conditional probability** of A given B, denoted as $P(A|B)$, is the probability that A will occur given B has already occurred.
  - Assuming $P(B) > 0$, it is defined as

$$P(A|B) = \frac{P(A, B)}{P(B)}.$$

Joint and Conditional Probabilities

- **Joint probability** of events A and B, denoted as $P(A, B)$, is the probability that both events will occur

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

- **Conditional probability** of A given B, denoted as $P(A|B)$, is the probability that A will occur given B has already occurred.
  - Assuming $P(B) > 0$, it is defined as

$$P(A|B) = \frac{P(A, B)}{P(B)}.$$

  - This leads to the **product rule** in probability theory:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A).$$

Independence

- Events A and B are **independent** if and only if

$$P(A, B) = P(A)P(B),$$

or equivalently (using the product rule),

$$P(A|B) = P(A).$$

Intuitively, this says that observing B does not have any effect on the probability of A.

Independence

- Events A and B are **independent** if and only if

$$P(A, B) = P(A)P(B),$$

or equivalently (using the product rule),

$$P(A|B) = P(A).$$

Intuitively, this says that observing B does not have any effect on the probability of A.

- Also events A and B are **conditionally independent** given event C if and only if

$$P(A, B|C) = P(A|C)P(B|C).$$

Outline

Random Variables

- Sample space is composed of events as five consequent coin tosses.

Random Variables

- Sample space is composed of events as five consequent coin tosses.
- But how do we go from these events to numerical outcomes that we actually care about? For example the number of heads?

Random Variables

- Sample space is composed of events as five consequent coin tosses.
- But how do we go from these events to numerical outcomes that we actually care about? For example the number of heads?
- A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ which assigns a real number $X(\omega)$ to each observed outcome $\omega \in \Omega$ in the sample space.

Random Variables

- Sample space is composed of events as five consequent coin tosses.
- But how do we go from these events to numerical outcomes that we actually care about? For example the number of heads?
- A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ which assigns a real number $X(\omega)$ to each observed outcome $\omega \in \Omega$ in the sample space.
- While the random variable is denoted with upper case letters $X(\omega)$ (or simply X), the values it can take are denoted with lower case ones x.

Random Variables

- Sample space is composed of events as five consequent coin tosses.
- But how do we go from these events to numerical outcomes that we actually care about? For example the number of heads?
- A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ which assigns a real number $X(\omega)$ to each observed outcome $\omega \in \Omega$ in the sample space.
- While the random variable is denoted with upper case letters $X(\omega)$ (or simply X), the values it can take are denoted with lower case ones x.
- Example: Consider the experiment of flipping a coin 5 times. $X(\omega)$ can be a counter that counts the number of heads. In this case if the outcome is $\omega_0 = HTTTH$, then $X(\omega_0) = 2$.

Discrete and Continuous Random Variables

- **Discrete** random variables
  - Can take only a finite number of values
  - Example: Number of heads in the sequence of 5 coin tosses
  - Here, the probability of the set associated with a random variable X taking on some specific value k is:

$$P(X = k) = P(\{\omega : X(\omega) = k\}).$$

Discrete and Continuous Random Variables

- **Discrete** random variables
  - Can take only a finite number of values
  - Example: Number of heads in the sequence of 5 coin tosses
  - Here, the probability of the set associated with a random variable X taking on some specific value k is:

$$P(X = k) = P(\{\omega : X(\omega) = k\}).$$

- **Continuous** random variables
  - Can take on a infinite number of possible values
  - Example: Time of bus arrivals in a bus station
  - Here, the probability of the set associated with a random variable X taking on some specific value between a and b ($a < b$) is:

$$P(a \leq X \leq b) = P(\{\omega : a \leq X(\omega) \leq b\}).$$

Probability Mass Functions

- When X is discrete random variable, we use probability mass functions to assign probabilities to random variables taking certain values.

Probability Mass Functions

- When X is discrete random variable, we use probability mass functions to assign probabilities to random variables taking certain values.
- The **probability mass function (PMF)** is a function $p_X : \mathbb{R} \to [0, 1]$ such that

$$p_X(x) = P(X = x).$$

Probability Mass Functions

- When X is discrete random variable, we use probability mass functions to assign probabilities to random variables taking certain values.
- The **probability mass function (PMF)** is a function $p_X : \mathbb{R} \to [0, 1]$ such that

$$p_X(x) = P(X = x).$$

- Intuitively, it is the probability that the random variable $X(\omega)$ will take the value x.

Probability Mass Functions

- Example: If $X(\omega)$ is a random variable indicating the number of H occurrences in 2 coin tosses, we have

$$p_X(2) = P(X = 2) = \frac{1}{4},$$

$$p_X(1) = P(X = 1) = \frac{1}{2},$$

$$p_X(0) = P(X = 0) = \frac{1}{4}.$$

Probability Mass Functions

- Example: If $X(\omega)$ is a random variable indicating the number of H occurrences in 2 coin tosses, we have

$$p_X(2) = P(X = 2) = \frac{1}{4},$$

$$p_X(1) = P(X = 1) = \frac{1}{2},$$

$$p_X(0) = P(X = 0) = \frac{1}{4}.$$

- For the properties PMFs see page 3 of CS229's Probability Theory Review.

Probability Density Functions

- When X is continuous random variable, we use probability density functions to assign probabilities to random variables taking certain values in a specific interval.

Probability Density Functions

- When X is continuous random variable, we use probability density functions to assign probabilities to random variables taking certain values in a specific interval.
- The **probability density function (PDF)** is a function $f_X : \mathbb{R} \to \mathbb{R}$ such that

$$\int_{x}^{x+\Delta x} f_X(x)dx = P(x \leq X \leq x + \Delta x).$$

Probability Density Functions

- When X is continuous random variable, we use probability density functions to assign probabilities to random variables taking certain values in a specific interval.
- The **probability density function (PDF)** is a function $f_X : \mathbb{R} \to \mathbb{R}$ such that

$$\int_{x}^{x+\Delta x} f_X(x)dx = P(x \leq X \leq x + \Delta x).$$

- It is important to note that $f_X(x)$ is not equal to $P(X = x)$, i.e.

$$f_X(x) \neq P(X = x).$$

  It is its integral over an interval that gives the probability.

Probability Density Functions

- When X is continuous random variable, we use probability density functions to assign probabilities to random variables taking certain values in a specific interval.
- The **probability density function (PDF)** is a function $f_X : \mathbb{R} \to \mathbb{R}$ such that

$$\int_x^{x+\Delta x} f_X(x)dx = P(x \leq X \leq x + \Delta x).$$

- It is important to note that $f_X(x)$ is not equal to $P(X = x)$, i.e.

$$f_X(x) \neq P(X = x).$$

  It is its integral over an interval that gives the probability.
- In fact, $f_X(x)$ can even be greater than 1 for certain x values. However, it cannot be negative.

Probability Density Functions

- When X is continuous random variable, we use probability density functions to assign probabilities to random variables taking certain values in a specific interval.
- The **probability density function (PDF)** is a function $f_X : \mathbb{R} \to \mathbb{R}$ such that

$$\int_x^{x+\Delta x} f_X(x)dx = P(x \leq X \leq x + \Delta x).$$

- It is important to note that $f_X(x)$ is not equal to $P(X = x)$, i.e.

$$f_X(x) \neq P(X = x).$$

  It is its integral over an interval that gives the probability.
- In fact, $f_X(x)$ can even be greater than 1 for certain x values. However, it cannot be negative.
- For the properties PDFs see page 4 of CS229's Probability Theory Review.

Expectation

- The **expectation** (**expected value** or **mean**) of a
  - discrete random variable X is defined as:

$$\mathbb{E}[X] = \sum_{x \in Val(X)} x p_X(x),$$

Expectation

- The **expectation** (**expected value** or **mean**) of a
  - discrete random variable X is defined as:

$$\mathbb{E}[X] = \sum_{x \in Val(X)} x p_X(x),$$

  - continuous random variable Y defined as:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

Expectation

- The **expectation** (**expected value** or **mean**) of a
  - discrete random variable X is defined as:

$$\mathbb{E}[X] = \sum_{x \in Val(X)} x p_X(x),$$

  - continuous random variable Y defined as:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

- Intuitively, the expectation of random variable can be thought of as a "weighted average" of the values it can take, where the weights are $p_X(x)$ or $f_Y(y)$.

Expectation

- The **expectation** (**expected value** or **mean**) of a
  - discrete random variable X is defined as:

$$\mathbb{E}[X] = \sum_{x \in Val(X)} x p_X(x),$$

  - continuous random variable Y defined as:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

- Intuitively, the expectation of random variable can be thought of as a "weighted average" of the values it can take, where the weights are $p_X(x)$ or $f_Y(y)$.
- For the properties expectations see page 4 of CS229's Probability Theory Review.

Variance

- The **variance** of a (discrete or continuous) random variable is defined as:
$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Variance

- The **variance** of a (discrete or continuous) random variable is defined as:
$$\mathsf{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

- With several steps of derivation, it can also be written as
$$\mathsf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Variance

- The **variance** of a (discrete or continuous) random variable is defined as:
$$\mathsf{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

- With several steps of derivation, it can also be written as
$$\mathsf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

- Intuitively, it is a measure of how concentrated the distribution of a random variables is around its mean.

Variance

- The **variance** of a (discrete or continuous) random variable is defined as:
$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$
- With several steps of derivation, it can also be written as
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$
- Intuitively, it is a measure of how concentrated the distribution of a random variables is around its mean.
- For the properties variance see page 4 of CS229's Probability Theory Review.

Common Random Variables

- Below are the distributions, means and variances of some common random variables[1]:

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

[1] Taken from CS229's Probability Theory Review.

Common Random Variables

- Below are the distributions, means and variances of some common random variables[1]:

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ $\forall x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

- Do not try to memorize them, they are put here for reference.

[1] Taken from CS229's Probability Theory Review.

Gaussian Distribution

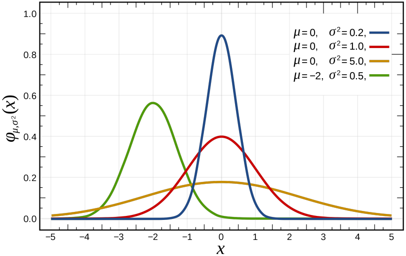
- Among these distributions the Gaussian distribution (also known as the Normal distribution) is particularly important as it shows up everywhere.

Gaussian Distribution

- Among these distributions the Gaussian distribution (also known as the Normal distribution) is particularly important as it shows up everywhere.
- It has the following PDF:

$$p_X(x) = \mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

and the following shape[2]:



[2] Taken from Wikipedia.

Outline

Two Random Variables

- Thus far, we have considered single random variables. In many situations, however, there may be more than one quantity that we are interested in knowing during a random experiment.

Two Random Variables

- Thus far, we have considered single random variables. In many situations, however, there may be more than one quantity that we are interested in knowing during a random experiment.
- For instance, in an experiment where we flip a coin ten times, we may care about both $X(\omega) =$ "the number of heads that come up" as well as $Y(\omega) =$ "the length of the longest run of consecutive heads".

Two Random Variables

- Thus far, we have considered single random variables. In many situations, however, there may be more than one quantity that we are interested in knowing during a random experiment.
- For instance, in an experiment where we flip a coin ten times, we may care about both $X(\omega) =$ "the number of heads that come up" as well as $Y(\omega) =$ "the length of the longest run of consecutive heads".
- In this part, we consider the setting of two random variables.

Joint and Marginal Probability Mass Functions

- When X and Y are discrete random variables, we use joint PMFs to assign probabilities to random variables taking certain values.

Joint and Marginal Probability Mass Functions

- When X and Y are discrete random variables, we use joint PMFs to assign probabilities to random variables taking certain values.
- The **joint PMF** is a function $p_{XY} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that

$$p_{XY}(x, y) = P(X = x, Y = y).$$

Joint and Marginal Probability Mass Functions

- When X and Y are discrete random variables, we use joint PMFs to assign probabilities to random variables taking certain values.
- The **joint PMF** is a function $p_{XY} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that

$$p_{XY}(x, y) = P(X = x, Y = y).$$

- Intuitively, it is the probability that the random variables $X(\omega)$ and $Y(\omega)$ will take the values x and y.

Joint and Marginal Probability Mass Functions...

- How does the joint PMF relate to PMFs for each variable separately? It turns out that

$$p_X(x) = \sum_y p_{XY}(x, y), \quad p_Y(y) = \sum_x p_{XY}(x, y).$$

Joint and Marginal Probability Mass Functions...

- How does the joint PMF relate to PMFs for each variable separately? It turns out that

$$p_X(x) = \sum_y p_{XY}(x, y), \quad p_Y(y) = \sum_x p_{XY}(x, y).$$

- $p_X(x)$ and $p_Y(y)$ are referred to **marginal PMFs** of X and Y, respectively.

Joint and Marginal Probability Mass Functions...

- How does the joint PMF relate to PMFs for each variable separately? It turns out that

$$p_X(x) = \sum_y p_{XY}(x, y), \quad p_Y(y) = \sum_x p_{XY}(x, y).$$

- $p_X(x)$ and $p_Y(y)$ are referred to **marginal PMFs** of X and Y, respectively.
- This summation is referred to as **marginalization**.

Joint and Marginal Probability Density Functions

- When X and Y are continuous random variables, we use probability density functions to assign probabilities to random variables taking certain values in specific intervals.

Joint and Marginal Probability Density Functions

- When X and Y are continuous random variables, we use probability density functions to assign probabilities to random variables taking certain values in specific intervals.
- The **joint PDF** is a function $f_{XY} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that

$$\int_{x}^{x+\Delta x} \int_{y}^{y+\Delta y} f_{XY}(x, y) dx dy$$
$$= P(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y).$$

Joint and Marginal Probability Density Functions

- When X and Y are continuous random variables, we use probability density functions to assign probabilities to random variables taking certain values in specific intervals.
- The **joint PDF** is a function $f_{XY} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that

$$\int_{x}^{x+\Delta x} \int_{y}^{y+\Delta y} f_{XY}(x, y) dx dy$$
$$= P(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y).$$

- It is important to note again that

$$f_{XY}(x, y) \neq P(X = x, Y = y).$$

It is its integral over the intervals that gives the probability.

Joint and Marginal Probability Density Functions

- When X and Y are continuous random variables, we use probability density functions to assign probabilities to random variables taking certain values in specific intervals.
- The **joint PDF** is a function $f_{XY} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that

$$\int_{x}^{x+\Delta x} \int_{y}^{y+\Delta y} f_{XY}(x, y) dx dy$$
$$= P(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y).$$

- It is important to note again that

$$f_{XY}(x, y) \neq P(X = x, Y = y).$$

  It is its integral over the intervals that gives the probability.
- In fact, $f_{XY}(x, y)$ can even be greater than 1 for certain x and y values. However, it cannot be negative.

Joint and Marginal Probability Density Functions...

- Analogous to the discrete case we have:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx.$$

Joint and Marginal Probability Density Functions...

- Analogous to the discrete case we have:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx.$$

- $f_X(x)$ and $f_Y(y)$ are referred to **marginal PDFs** of X and Y, respectively.

Conditional Distributions

- In the discrete case, the **conditional PMF** of Y given X is defined as:
$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)},$$
assuming that $p_X(x) > 0$.

Conditional Distributions

- In the discrete case, the **conditional PMF** of Y given X is defined as:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)},$$

assuming that $p_X(x) > 0$.

- For the continuous case see page 8 of CS229's Probability Theory Review, but it has the same form.

Bayes' Rule

- A useful formula that often arises when trying to derive an expression for the conditional probability of one event given the other is **Bayes' Rule**.

Bayes' Rule

- A useful formula that often arises when trying to derive an expression for the conditional probability of one event given the other is **Bayes' Rule**.
- In the case of discrete random variables X and Y:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$(\text{Posterior} \propto \text{Likelihood} \times \text{Prior}).$$

Bayes' Rule

- A useful formula that often arises when trying to derive an expression for the conditional probability of one event given the other is **Bayes' Rule**.
- In the case of discrete random variables X and Y:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$(\text{Posterior} \propto \text{Likelihood} \times \text{Prior}).$$

- The same thing can be done with PDFs in the case of continuous random variables.

Independence

- Two discrete random variables X and Y are independent if

$$p_{XY}(x, y) = p_X(x)p_Y(y),$$

for all $x \in Val(X)$ and $y \in Val(Y)$.

Independence

- Two discrete random variables X and Y are independent if

$$p_{XY}(x, y) = p_X(x)p_Y(y),$$

for all $x \in Val(X)$ and $y \in Val(Y)$.

- For the continuous case see page 8 of CS229's Probability Theory Review, but it has the same form.

Independence

- Two discrete random variables X and Y are independent if

$$p_{XY}(x, y) = p_X(x) p_Y(y),$$

  for all $x \in Val(X)$ and $y \in Val(Y)$.
- For the continuous case see page 8 of CS229's Probability Theory Review, but it has the same form.
- Intuitively, two random variables are independent if knowing the value of one variable will never affect the probability of knowing the other.

Expectation and Covariance

- Suppose that we have two discrete random variables X and Y, and $g(.,.)$ is a function of these two random variables.

Expectation and Covariance

- Suppose that we have two discrete random variables X and Y, and $g(.,.)$ is a function of these two random variables.
- Then the expected value of g is defined in the following way:

$$\mathbb{E}[g(X,Y)] = \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y) p_{XY}(x,y).$$

Expectation and Covariance

- Suppose that we have two discrete random variables X and Y, and $g(.,.)$ is a function of these two random variables.
- Then the expected value of g is defined in the following way:

$$\mathbb{E}[g(X,Y)] = \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y) p_{XY}(x,y).$$

- For continuous random variables X and Y the analogous expression is:

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

Expectation and Covariance...

- Using the expectation definition in the previous slide, the **covariance** of two random variables X and Y is defined as:

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Expectation and Covariance...

- Using the expectation definition in the previous slide, the **covariance** of two random variables X and Y is defined as:

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

- With a few steps of derivation it can also be written as:

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Expectation and Covariance...

- Using the expectation definition in the previous slide, the **covariance** of two random variables X and Y is defined as:

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

- With a few steps of derivation it can also be written as:

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

- When $Cov[X, Y] = 0$, we say that X and Y are **uncorrelated**.

Expectation and Covariance...

- Using the expectation definition in the previous slide, the **covariance** of two random variables X and Y is defined as:

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

- With a few steps of derivation it can also be written as:

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

- When $Cov[X, Y] = 0$, we say that X and Y are **uncorrelated**.
- When X and Y are independent, then $Cov[X, Y] = 0$. However, the opposite is not always true.

Expectation and Covariance...

- Using the expectation definition in the previous slide, the **covariance** of two random variables X and Y is defined as:

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

- With a few steps of derivation it can also be written as:

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

- When $Cov[X, Y] = 0$, we say that X and Y are **uncorrelated**.
- When X and Y are independent, then $Cov[X, Y] = 0$. However, the opposite is not always true.
- For the properties covariance see page 9 of CS229's Probability Theory Review.

Outline

Multiple Random Variables

- The notions and ideas introduced in the previous section can be generalized to more than two random variables.

Multiple Random Variables

- The notions and ideas introduced in the previous section can be generalized to more than two random variables.
- In particular, suppose that we have n random variables

$$X_1(\omega), X_2(\omega), \ldots, X_n(\omega).$$

Example of Generalization

- In the case of having two discrete random variables X and Y, the joint PMF was defined as:

$$p_{XY}(x, y) = P(X = x, Y = y),$$

and marginalization was done as follows:

$$p_X(x) = \sum_y p_{XY}(x, y), \quad p_Y(y) = \sum_x p_{XY}(x, y).$$

Example of Generalization

- In the case of having two discrete random variables X and Y, the joint PMF was defined as:

$$p_{XY}(x, y) = P(X = x, Y = y),$$

and marginalization was done as follows:

$$p_X(x) = \sum_y p_{XY}(x, y), \quad p_Y(y) = \sum_x p_{XY}(x, y).$$

- In the case of having n discrete random variables $X_1, X_2, \ldots, X_n$, the joint PMF becomes:

$$p_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n),$$

and marginalization becomes:

$$p_{X_1}(x_1) = \sum_{x_2, x_3, \ldots, x_n} p_{X_2 \ldots X_n}(x_2, \ldots, x_n).$$

Example of Generalization

- In the case of having two discrete random variables X and Y, the joint PMF was defined as:

$$p_{XY}(x, y) = P(X = x, Y = y),$$

and marginalization was done as follows:

$$p_X(x) = \sum_y p_{XY}(x, y), \quad p_Y(y) = \sum_x p_{XY}(x, y).$$

- In the case of having n discrete random variables $X_1, X_2, \dots, X_n$, the joint PMF becomes:

$$p_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n),$$

and marginalization becomes:

$$p_{X_1}(x_1) = \sum_{x_2, x_3, \dots, x_n} p_{X_2 \dots X_n}(x_2, \dots, x_n).$$

Note that the same can be done for $x_2, \dots, x_n$.

Chain Rule and Mutual Independence

- From the definition of conditional probabilities, one can show that:

$$p_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = p_{X_1}(x_1) \prod_{i=2}^{n} p_{X_i}(x_i | x_1, \dots, x_{i-1}).$$

This is called the **chain rule** in probability theory.

Chain Rule and Mutual Independence

- From the definition of conditional probabilities, one can show that:

$$p_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = p_{X_1}(x_1) \prod_{i=2}^{n} p_{X_i}(x_i | x_1, \dots, x_{i-1}).$$

  This is called the **chain rule** in probability theory.
- In this way, a joint probability is written in terms of conditional probabilities.

Chain Rule and Mutual Independence

- From the definition of conditional probabilities, one can show that:

$$p_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1) \prod_{i=2}^{n} p_{X_i}(x_i | x_1, \ldots, x_{i-1}).$$

  This is called the **chain rule** in probability theory.
- In this way, a joint probability is written in terms of conditional probabilities.
- We say that $X_1, X_2, \ldots, X_n$ are **mutually independent** if:

$$p_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

Independent and Identically Distributed (IID)

- A set of random variables are **independent and identically distributed (i.i.d.)** if
  - they are sampled from the same distribution
  - and are mutually independent.

Independent and Identically Distributed (IID)

- A set of random variables are **independent and identically distributed (i.i.d.)** if
  - they are sampled from the same distribution
  - and are mutually independent.
- Example: Consecutive coin flips are assumed to be i.i.d.:
  - The probability of H is the same for each flip.
  - One flip's outcome doesn't affect the others.

Random Vectors

- When working with multiple random variables $X_1, X_2, \ldots, X_n$, it is often convenient to put them in a vector:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

This resulting vector is called a **random vector** (a mapping $\boldsymbol{X} : \Omega \to \mathbb{R}^n$).

Random Vectors

- When working with multiple random variables $X_1, X_2, \ldots, X_n$, it is often convenient to put them in a vector:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

  This resulting vector is called a **random vector** (a mapping $\boldsymbol{X} : \Omega \to \mathbb{R}^n$).
- It should be noted that this is just another notation, and nothing more.

Covariance Matrix

- For a given random vector $\boldsymbol{X}$, its covariance matrix $\boldsymbol{\Sigma}$ is an $n \times n$ square matrix and is defined as:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\top].$$

And its (i, j)th entry is

$$\boldsymbol{\Sigma}_{ij} = Cov[X_i, X_j].$$

Covariance Matrix

- For a given random vector $\boldsymbol{X}$, its covariance matrix $\boldsymbol{\Sigma}$ is an $n \times n$ square matrix and is defined as:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\top].$$

  And its (i, j)th entry is

$$\boldsymbol{\Sigma}_{ij} = Cov[X_i, X_j].$$

- With a few steps of derivation it can also be written as:

$$\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top.$$

Covariance Matrix

- For a given random vector $\boldsymbol{X}$, its covariance matrix $\boldsymbol{\Sigma}$ is an $n \times n$ square matrix and is defined as:

  $$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\top].$$

  And its (i, j)th entry is

  $$\boldsymbol{\Sigma}_{ij} = Cov[X_i, X_j].$$

- With a few steps of derivation it can also be written as:

  $$\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top.$$

- It should be noted that $\boldsymbol{\Sigma}$ is positive semidefinite and symmetric.

Multivariate Gaussian Distribution

- A particularly important example of a probability distribution over random vectors is the **multivariate Gaussian (normal) distribution**.

Multivariate Gaussian Distribution

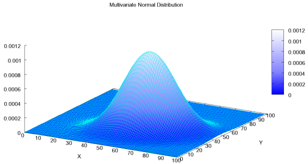
- A particularly important example of a probability distribution over random vectors is the **multivariate Gaussian (normal) distribution**.
- It is a generalization of the univariate Guassian to higher dimensions.

Multivariate Gaussian Distribution

- A particularly important example of a probability distribution over random vectors is the **multivariate Gaussian (normal) distribution**.
- It is a generalization of the univariate Guassian to higher dimensions.
- A multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ has the following PDF:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}},$$

and the following shape[3] (in 2D):



[3]Taken from Wikipedia.

References

- This tutorial is mainly adapted from Stanford's CS229 Probability Theory Review. However, it doesn't give all the details. For the details please refer to this source.