ITCS 5356: Introduction to Machine Learning

Probability Theory

Razvan C. Bunescu

Department of Computer Science @ CCI

razvan.bunescu@charlotte.edu

Some of these slides are based on material from Dr. Minwoo Lee

Probability Theory

Image from: https://tawheedyousuf.medium.com/probability-theory-descriptive-statistics-2e35d1dab20a

Probabilities and Random Variables

- A **probability** is a number between 0 and 1 that indicates how likely some event is to occur or how likely it is that something happens.
  - A probability of 0 indicates that it is impossible.
  - A probability of 1 indicates that it must happen.
  - A probability of 0.5 indicates a 50% chance.
- A **random variable** a variable that can take on one of several values, each value having a probability of occurrence.
  - Two types:
    - Discrete
    - Continuous

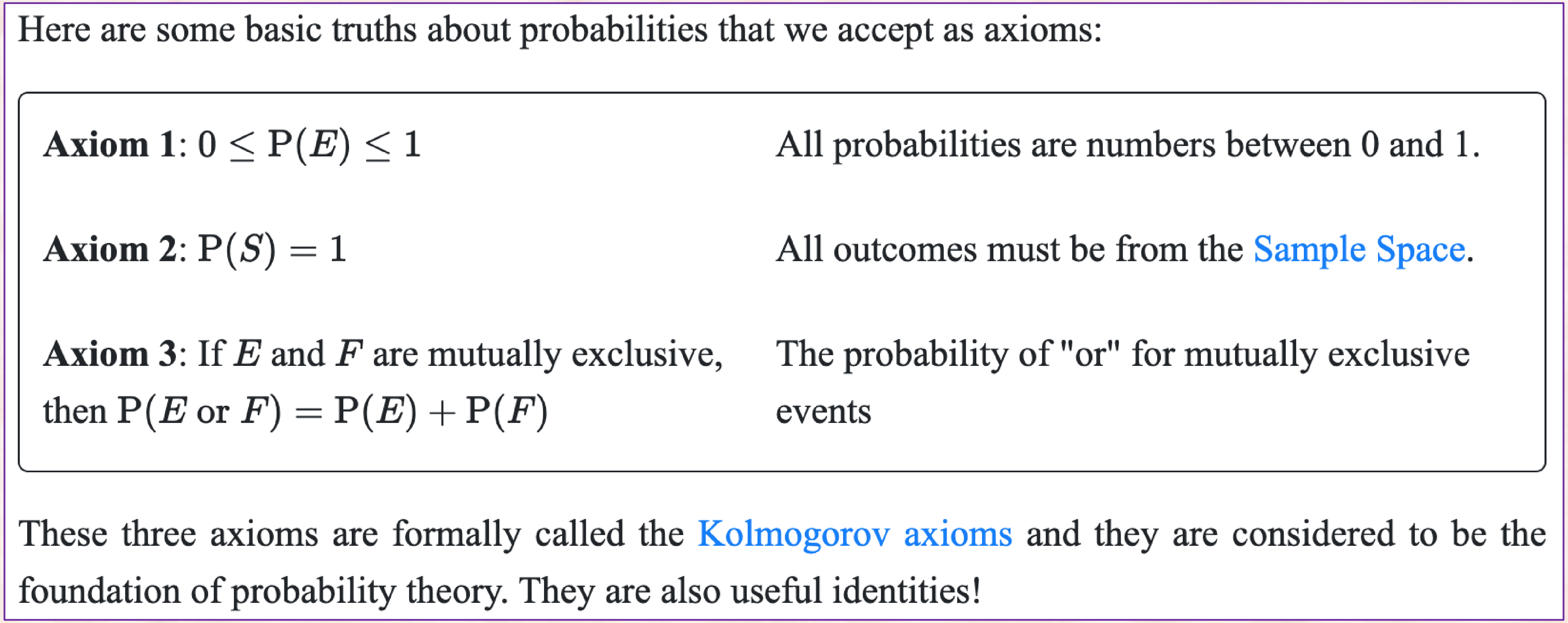
Sample Space, Events, Probability

- A **sample space** S is the set of all possible outcomes of an experiment:
  - A set S of possible outcomess for throwing a die, $S = \{1, 2, 3, 4, 5, 6\}$
  - A set S of possible outcomes for throwing a pair of dice, $S = \{(1,1), (1, 2), \ldots, (5, 6), (6, 6)\}$

- An event $\mathcal{F} \subset S$ is some relevant subset of S we ascribe meaning to:
  - “odd roll” is the event $\mathcal{F} = \{1,3,5\}$

- $P : \mathcal{F} \rightarrow \mathbb{R}$ is the probability function, assigning probabilities to events:
  - $P(\text{odd roll}) = \frac{1}{2}$

Axioms of Probability



Here are some basic truths about probabilities that we accept as axioms:

Axiom 1: $0 \leq \mathrm{P}(E) \leq 1$	All probabilities are numbers between 0 and 1.
Axiom 2: $\mathrm{P}(S) = 1$	All outcomes must be from the Sample Space.
Axiom 3: If E and F are mutually exclusive, then $\mathrm{P}(E \text{ or } F) = \mathrm{P}(E) + \mathrm{P}(F)$	The probability of "or" for mutually exclusive events

These three axioms are formally called the Kolmogorov axioms and they are considered to be the foundation of probability theory. They are also useful identities!

If A, B are not mutually disjoint, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Prior Probability

The *unconditional* or **prior probability** of an event is the degree of belief that the event will happen in the absence of any other evidence.

$$P(doubles) = \frac{6}{36} = \frac{1}{6}$$

Out of 36 possible combinations, only six are doubles.

Posterior Probability

The *conditional* or **posterior probability** of an event is the degree of belief that the event will happen given some evidence.

$$P(\underbrace{doubles}_{\text{event}} \mid D_1 = 4) = \frac{1}{6}$$

Posterior Probability

The *conditional* or **posterior probability** of an event is the degree of belief that the event will happen given some evidence.

$$P(doubles \underbrace{\,|\,}_{\text{“given that…”}} D_1 = 4) = \frac{1}{6}$$

Posterior Probability

The *conditional* or **posterior probability** of an event is the degree of belief that the event will happen given some evidence.

$$P(doubles \mid \underbrace{D_1 = 4}_{\text{evidence}}) = \frac{1}{6}$$

Posterior Probability

The *conditional* or **posterior probability** of an event is the degree of belief that the event will happen given some evidence.

$$P(doubles|D_1 = 4) = \frac{1}{6}$$

"What is the probability that doubles will be rolled given that the first die already shows a 4?"

Posterior Probability

The *conditional* or **posterior probability** of an event is the degree of belief that the event will happen given some evidence.

$$P(doubles|D_1 = 4) = \frac{1}{6}$$

Posterior Probability

Posterior probabilities are defined in terms of **evidence** probabilities.

$$P(event|evidence) = \frac{P(event \cap evidence)}{P(evidence)}$$

What is the **probability of an event** occurring, **given some existing evidence**?

Posterior Probability

Posterior probabilities are defined in terms of **evidence** probabilities.

$$P(doubles|D_1 = 4) = \frac{P(doubles \cap D_1 = 4)}{P(D_1 = 4)}$$

Posterior Probability

Posterior probabilities are defined in terms of **evidence** probabilities.

$$P(doubles|D_1 = 4) = \frac{P(doubles \cap D_1 = 4)}{1/6}$$

Posterior Probability

Posterior probabilities are defined in terms of evidence probabilities.

$$P(doubles|D_1 = 4) = \frac{1/36}{1/6}$$

Posterior Probability

Posterior probabilities are defined in terms of evidence probabilities.

$$P(doubles|D_1 = 4) = \frac{1}{6}$$

Independence

- A is **independent** of B if knowing B does not change our belief about X:
  - $P(X = x \mid Y = y) = P(X = x)$
  - $P(X = x, Y = y) = P(X = x)P(Y = y)\ \forall x, y$
  - Independence is symmetric and written as $A \perp B$ or $B \perp A$.

Conditional Probability and Independence

- What is the probability that both A and B occur together?

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \Rightarrow P(A,B) = P(A \mid B)\mathrm{P(B)}$$

- When A and B are statistically independent ($A \perp B$):

$$P(A,B) = P(A)\mathrm{P(B)}$$

- In general, if events A_i are independent:

$$P(\cap_i A_i) = \prod_i P(A_i)$$

Independence

- $X_1, \dots, X_n$ are independent if and only if:

$$P(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

- If $X_1, \dots, X_n$ are *independent* and *identically distributed*, we say they are i.i.d. (or that they are a random sample) and we write:

$$X_1, \dots, X_n \sim p$$

Conditional Independence

- Random variables are rarely independent.
- We can still leverage local structural properties like **conditional independence** (CI).
  - $X \perp Y \mid Z$ if once Z is observed, knowing the value of Y does not change our belief about X.
- Let X = *carry umbrella* and Y = *ground is wet* and Z = *it rains*:
  - X and Y are **not independent**:
    - If the ground is wet, then there is a higher chance it rains, which means there is a higher change that one will carry an umbrella.
  - X and Y are **conditionally independent** given Z:
    - They are **independent given** the presence or absence of rain.

Conditional Independence $X \perp Y \mid Z$

$$P(X = x \mid Z = z, Y = y) = P(X = x \mid Z = z)$$

$$P(Y = y \mid Z = z, X = x) = P(Y = y \mid Z = z)$$

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$$

Bayes Rule

$$\left.\begin{array}{l} \bullet\ P(A,B) = P(\mathrm{B})P(A \mid B) \\[2ex] \bullet\ P(\mathrm{A|B}) = P(\mathrm{A})P(B \mid A) \end{array}\right\} \Rightarrow P(A \mid B)\mathrm{P(B)} = P(B \mid A)\mathrm{P(A)}$$

$$\boxed{\Rightarrow P(A \mid B) = \frac{\mathrm{P(A)}P(B \mid A)}{\mathrm{P(B)}}}$$

- Examples: $P(\mathit{Disease} \mid \mathit{Symptom}) = \dfrac{P(\mathit{Symptom} \mid \mathit{Disease})\mathrm{P}(\mathit{Disease})}{\mathrm{P}(\mathit{Symptom})}$

$$P(y \mid \mathbf{x}) = \frac{\mathrm{P}(y)P(\mathbf{x}|y)}{\mathrm{P}(\mathbf{x})} \qquad P(\mathbf{w} \mid \mathrm{X}, \mathbf{y}) = \frac{\mathrm{P}(\mathbf{w})P(\mathrm{X}, \mathbf{y}|\mathbf{w})}{\mathrm{P}(\mathrm{X}, \mathbf{y})}$$

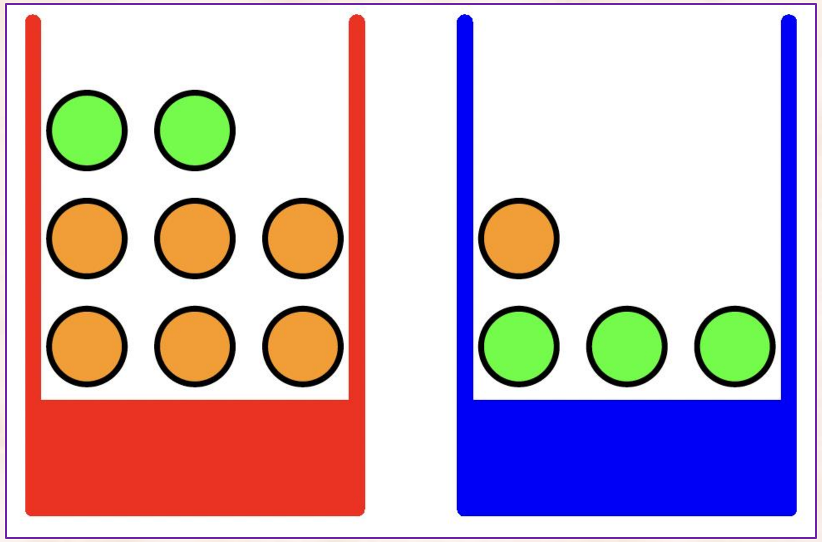
label *feature vector* *parameters* *training examples*

Homework: Bayes Rule

There are two boxes, one *red* and one *blue*. The red box contains 2 *limes* and 6 *oranges*. The blue box contains 3 *limes* and 1 *orange*.

A box is chosen at random with probability 0.8 for the red box and 0.2 for the blue box.

One fruit is picked from the chosen box, uniformly at random.



If we observe that the fruit is a lime, what is the probability that it came from the red box?

Discrete Random Variables

- Let $\{x_1, x_2, \ldots, x_K\}$ to the set of all possible values for discrete random variable X.
- A **probability mass function** (pmf) assignes a probability for each possible value:

$$P(X = x_i)$$

- Properties (based on axioms of probability):
  - $\sum_i P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0 \quad \text{if } i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j) \quad \text{if } i \neq j$
  - $P(X = x_1 \cup X = x_2 \cdots \cup X = x_K) = 1$

Common Discrete Distributions

- **Uniform** $X \sim U[1, \dots, N]$
  - X takes values 1, 2, …, N
  - $P(X = i) = 1/N$
    - N = 6 for rolling a fair dice.

- **Binomial** $X \sim Bin(N, p)$
  - X takes values $0, 1, \dots, N$
  - $\mathrm{P}(X = i) = \binom{N}{i} p^i (1-p)^{N-i}$
    - N coin flips, *i* heads observed.
    - *p* is the probability of heads in a flip.

- **Bernoulli** $X \sim Bernoulli(p)$

  X takes values $0, 1$

  $$\mathrm{P}(X = 1) = p$$

For example:

p is the probability of heads in a coin flip.

X takes values in {H, T}

P(X = H) = *p*

If the coin is fair, *p* = 0.5

Continuous Random Variables

- A **probability density function** (pdf) $f(x)$ indicates how dense the probability is around x.
  - High $f(x)$ means a high concentration of probability around x.

- Properties of a pdf f:
  - $f(x) \geq 0, \ \forall x$
  - $\int_{-\infty}^{\infty} f(x) = 1$

- Probability calculation is done using the integral of the pdf:
  - The probability that X falls in an interval $[a, b]$:

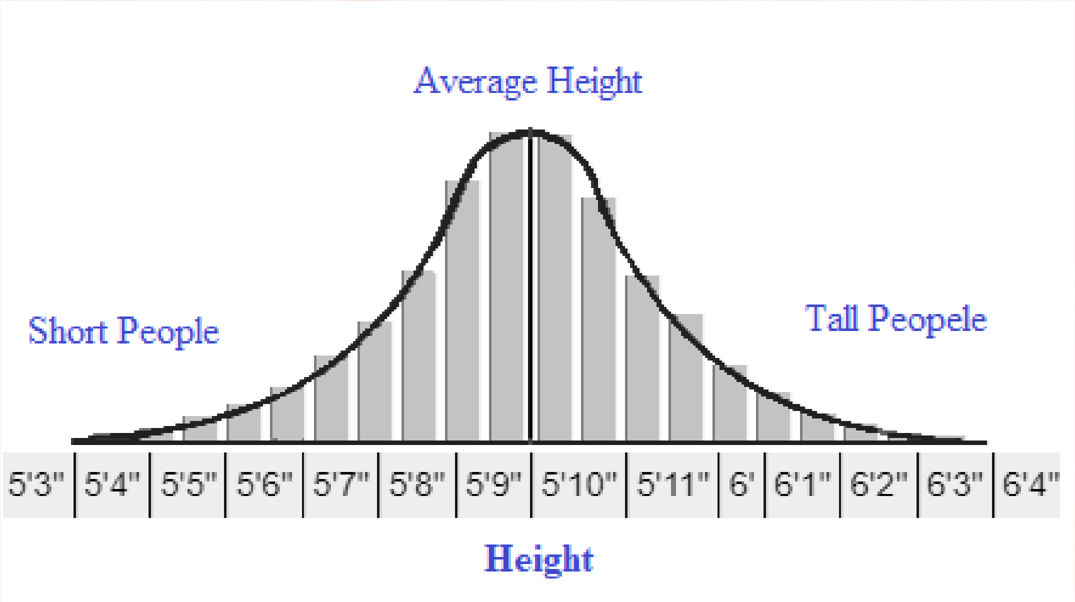
$$P(a \leq X \leq b) = \int_{a}^{b} f(x) dx$$

Common Distributions

- Normal or Gaussins distribution $X \sim \mathcal{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  - For example, X is the height of adult humans.



Cumulative Distribution Function

- The **cumulative distribution function** (cdf) provides the probability that the r.v. *X* is less than or equal to a certain value *v*.

$$F_X(v) = P(\, X \le v)$$

- Discrete random variables:
  - $F_X(v) = \sum_{v_i \le v} P(X = v_i)$

- Continuous random variables:
  - $F_X(v) = \int_{-\infty}^{v} f(x)dx$
  - $\frac{d}{dx} F_X(x) = f(x)$

Multivariate Joint Distributions

$\mathbf{P}(Weather, Car)$

		Variable	
		drive Car to work	
		Value	
Variable	**Value**	$true$	$false$
$Weather$ is	$sunny$	0.120	0.480
	$rainy$	0.020	0.080
	$cloudy$	0.058	0.232
	$snowy$	0.002	0.008

This is the full table for the *joint probability distribution* of variables $Weather$ and Car.

$$\mathbf{P}(W = s) = P(W = s, C = t) + P(W = s, C = f) = 0.12 + 0.48 = 0.6$$

The Chain Rule

Generalization of $P(x,y) = P(x|y)P(y) = P(y|x)P(x)$

$$P(x,y,z) = P(x)P(y \mid x)P(z \mid x,y)$$

Logical AND is commutative, hence the order does not matter.

$$P(x,y,z) = P(z)P(y \mid z)P(x \mid y,z)$$

$$P(x,y,z) = P(z)P(x \mid z)P(y \mid x,z)$$

$$P(x,y,z) = P(y)P(z \mid y)P(x \mid y,z)$$

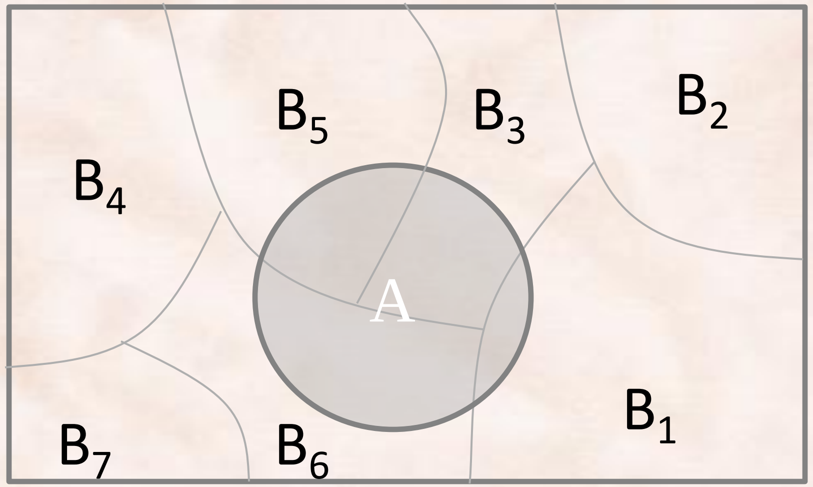
$$\dots$$

Marginalization

- Give a joint distribution, e.g. $P(X, Y)$, compute simpler distributions, e.g. P(X) or P(y).
  - Use the Sum Rule of Probability (Law of Total Probability).

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

$$= \textstyle\sum_y P(X{=}x \mid Y{=}y)\mathrm{P}(\mathrm{Y}{=}\mathrm{y})$$



$P(T, W)$

		Temperature		
		hot	cold	
Weather	sun	0.45	0.15	0.60
	rain	0.02	0.08	0.10
	fog	0.03	0.27	0.30
	meteor	0.00	0.00	0.00
		0.50	0.50	

$P(W)$

$P(T)$

Summary

- Probability as a measure of belief or the chance of an event [0,1]
- Sample space, events, and the probability axioms.
- Conditional probability $P(\, A \mid B \,) = \frac{P(A \cap B)}{P(B)}$
- Absolute and conditional independence.
- Bayes Rule $P(\, A \mid B \,) = \frac{P(B|A)P(A)}{P(B)}$
- Probability distributions:
  - Discrete (Uniform, Binomial, Bernoulli) and Continuous (Normal).
- Chain Rule $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{<i})$
- Multivariate joint distribution and marginalization.

Required Readings

1. Section 1.2 on Probability Theory (pp. 12-28) in the textbook Pattern Recognition and Machine Learning textbook.
2. Chapter 3 on Probability and Information Theory (pp. 51-68) in the Deep Learning textbook.
3. Probability for Computer Scientists course reader from Stanford CS 109:
   1. Part 1 on Core Probability.
   2. Part 2 on Random Variables.
   3. Part 3 on Probabilistic Models, first 3 sections on Joint Probability, Marginalization, Multinomial.

The introductory resources above are listed in decreasing order of familiarity and expertise. If you are already familiar with probability theory and just need a quick refresher, you can read 1 or 2. If you have not had a formal introduction to probability theory or struggle with using it, you are strongly advised to go through the readings are 3 above, which introduce the major concepts in detail and offer extensive examples and exercises.