2. Review on Probability Theory

Dongwoo Kim

dongwookim.ac.kr

February 22, 2023

Table of Contents

Machine Learning and Probability Theory

- Recall that machine learning is to make a function from learning patterns in data, where pattern = a simple summary of data

Machine Learning and Probability Theory

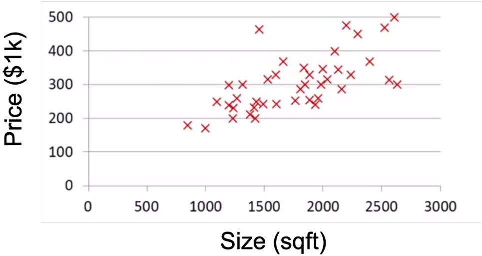
- Recall that machine learning is to make a function from learning patterns in data, where pattern = a simple summary of data
- Example: tossing a possibly unfair coin
  - Data: T T H H H T H T H T
  - Task: prediction of next outcome
  - A summary: H's and T's are fifty-fifty
  - In probability theory: the probability of seeing head is 0.5

Machine Learning and Probability Theory

- Recall that machine learning is to make a function from learning patterns in data, where pattern = a simple summary of data
- Example: tossing a possibly unfair coin
  - Data: T T H H H T H T H T
  - Task: prediction of next outcome
  - A summary: H's and T's are fifty-fifty
  - In probability theory: the probability of seeing head is 0.5
- The concept of probability/statistics is indeed quite useful to express patterns of data!
  - c.f., probability deals with predicting the likelihood of future events, while statistics involves the analysis of the frequency of past events.

Probability: a language of uncertainty/prediction

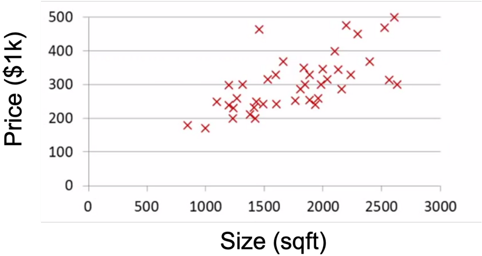
Life is full of surprises/uncertainty (so it is fun!)



- Task: predict the price of house of size 2,500 sqft
  - A1: $400k?
  - A2: $300k $\sim$ $500k?

Probability: a language of uncertainty/prediction

Life is full of surprises/uncertainty (so it is fun!)



- Task: predict the price of house of size 2,500 sqft
  - A1: \$400k?
  - A2: \$300k $\sim$ \$500k?
- c.f., for better prediction, we might need more features of the house (course tip: diversify learning source!)

Table of Contents

Probability Space

- A probability space is defined by triplet $(\Omega, \mathcal{F}, P)$:
  - Sample space Ω
  - Set of events (or event space)[1] $\mathcal{F}$
  - Probability measure P

[1] Technically called σ-field

Sample Space, Events, Field

- Sample space Ω is the set of all possible outcomes, where an outcome (incidence, or sample) is the result of a single execution of the model.
  - e.g., two successive coin tosses: $\Omega = \{hh, tt, ht, th\}$
- A subset E of Ω is called an event.
  - e.g., {hh}, {ht, th}, ...
- A collection $\mathcal{F}$ of subsets (or events) of Ω forms a field if
  - $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$;
  - $\forall E_1, E_2 \in \mathcal{F}$, $E_1 \cup E_2 \in \mathcal{F}$ and $E_1 \cap E_2 \in \mathcal{F}$;
  - $\forall E \in \mathcal{F}$, $\overline{E} := \Omega \setminus E \in \mathcal{F}$.
- A field $\mathcal{F}$ is σ-field if it is closed under any countable set of unions, intersections, and combinations.

Probability Measure

- Given a sample space Ω and σ-field $\mathcal{F} \subset 2^{\Omega}$, a function $P : \mathcal{F} \mapsto [0, 1]$ is a probability measure if
  - $P(A) \geq 0$ for any event $A \in \mathcal{F}$, $P(\emptyset) = 0$ and $P(\Omega) = 1$;
  - (countable additivity) For all countable collections $\{A_i\}_{i \in I}$ of pairwise disjoint events, i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j \in I$,

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i) \ .$$

- These properties are called the axioms of probability.

Table of Contents

Important Properties of Probability (1)

- Joint probability: $P(A, B) := P(A \cap B)$
- Marginal probability: $P(A)$, $P(B)$
- Independence between A and B iff $P(A, B) = P(A)P(B)$
- Conditional probability: $P(A \mid B) := \frac{P(A,B)}{P(B)}$ if $P(B) \neq 0$
  - If A and B are independent, then $P(A \mid B) = P(A)$

Important Properties of Probability (2)

- Law of total probability (a.k.a. marginalization):

$$P(A) = \sum_{i=1}^{n} P(A, B_i) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

  - $\{B_i\}_{i=1,\ldots,n}$ is a partition of Ω, i.e., $\bigcup_{i=1}^{n} B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$
  - Marginalizing out unwanted data is a basic operation to process raw data

marginalize verb

Save Word

mar·gin·al·ize | \ ˈmärj-nə-ˌlīz, ˈmär-jə-nᵊl-ˌīz \

marginalized; **marginalizing**

Definition of *marginalize*

transitive verb

: to relegate (see RELEGATE sense 2) to an unimportant or powerless position within a society or group

Example of Marginalization

Suppose we have two unfair coins A and B, and observed:

(A, B)	# of observations
(H, H)	100
(H, T)	500
(T, H)	200
(T, T)	200

(A, B)	estimated $P(A, B)$
(H, H)	0.1
(H, T)	0.5
(T, H)	0.2
(T, T)	0.2

- $P(A = H) = P(A = H, B = H) + P(A = H, B = T) = 0.1 + 0.5$
- $P(B = H) = P(A = H, B = H) + P(A = T, B = H) = 0.1 + 0.2$

Example of Independence

Suppose we have two unfair coins A and B, and observed:

(A, B)	$\#$ of observations
(H, H)	100
(H, T)	500
(T, H)	200
(T, T)	200

(A, B)	estimated $P(A, B)$
(H, H)	0.1
(H, T)	0.5
(T, H)	0.2
(T, T)	0.2

Is it helpful to observe coin B to predict A?

- Recall $P(A = H) = 0.6$ and $P(B = H) = 0.3$
- Coins A and B are not independent as

$$P(A = H, B = H) = 0.1 \neq 0.6 \times 0.3 = 0.18$$

- This implies that observing B may provide some information on A!

Important Properties of Probability (3)

- Bayes' theorem:

$$P(A \mid B) := \frac{P(B \mid A)P(A)}{P(B)} \quad \text{if } P(B) \neq 0$$
$$\propto P(B \mid A)P(A) \quad \text{for any } A \text{ and fixed } B$$

- This is particularly useful in ML since...
  - We may want to find y[2] maximizing $P(\text{unobserved } y \mid \text{observation } x)$
  - But, a probabilistic model typically explains $P(x \mid y)P(y)$ rather than $P(y \mid x)$.

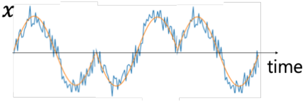
[2] often called latent

Example of Probabilistic Model

- Inference to maximize $P(\text{latent } y \mid \text{observation } x)$
- Typically, a causal model, which explains the generation of x from y, is employed,
- Example in a communication system:

Latent variables → Noisy Observation → Inference

$y = 101$

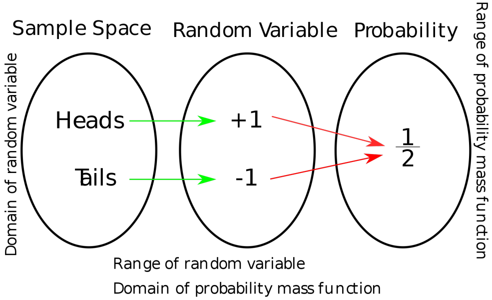


$\hat{y} = ?$

Table of Contents

Random Variable

- A random variable (r.v.) $X : \Omega \mapsto \mathbb{R}$ is a correspondence rule between a random outcome $\omega \in \Omega$ of an experiment and the real number $\mathbb{R}$.
  - For simplicity, we use X instead of $X(\omega)$.
  - (Probability) distribution of X is denoted by $P(X)$.
  - The range of random variable is often called support



Continuous Random Variables

- Since probability with continuous variable is defined for an infinite number of points over a continuous interval, a probability of a single point is always zero.
- Thus, the probabilities are measured over *intervals*, not a single point.

Definition (Probability density function)

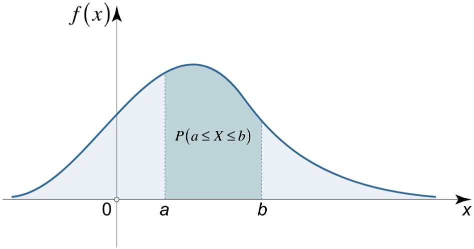
A function $f : \mathbb{R} \to \mathbb{R}$ is called a probability density function (pdf) if

1. $\forall x \in \mathbb{R} : f(x) \geq 0$
2. Its integral exists and $\int_{\mathbb{R}} f(x)dx = 1$.

Therefore,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

PDF example



- Area under the curve must be equal to 1.
- $f(x)$ can be greater than 1 at a particular point.

Cumulative Distribution Function

Definition (Cumulative Distribution Function)

A cumulative distribution function (CDF) is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as,

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

The cdf can be expressed also as the integral of the probability density function $f(\mathbf{x})$ so that

$$F_X(x) = \int_{-\infty}^{x} f(z)dz.$$

CDF and PDF

If CDF $F_X(x)$ is differentiable everywhere, we define PDF as the derivative of the CDF.

$$f(x) := \frac{dF_X(x)}{dx}$$

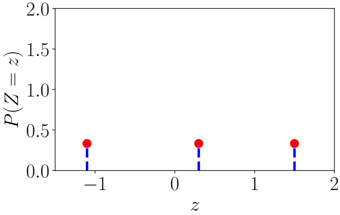
For small Δx

$$P(x \leq X \leq x + \Delta x) \approx f(x)\Delta x$$

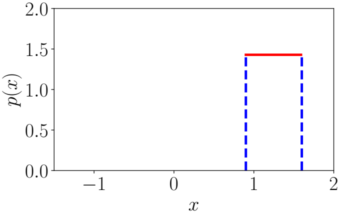
Both CDFs and PDFs can be used for calculating the probabilities of different events[3].

[3] Again, the value of PDF at any given point x is not the probability

Discrete vs Continuous



(a) Discrete distribution



(b) Continuous distribution

- (a) Uniform distribution over 3 variables
- (b) Uniform distribution over $[0.9, 1.6]$

Expectation and Moments

- Expectation[4] $\mathbb{E}[X] := \sum_x xP(X = x)$
- The k-th moment of X $\mathbb{E}[X^k] := \sum_x x^k P(X = x)$
- Mean (a.k.a. the first moment, average, expected value):

$$\mu_X := \mathbb{E}[X] = \sum_x xP(X = x) \, .$$

- Variance (a.k.a. the second central moment):

$$\sigma_X^2 = \mathbb{E}\left[(X - \mu_X)^2\right] = \mathbb{E}\left[X^2\right] - \mu_X^2 \, .$$

[4] $\mathbb{E}[X^k] := \int_{\mathcal{X}} xf(x)dx$ with continuous random variable.

Expectation Properties

Consider random variables X, constant c, and function $f : \mathcal{X} \to \mathbb{R}$

- $\mathbb{E}[c] = c$
- $\mathbb{E}[cX] = c\,\mathbb{E}[X]$
- $\mathbb{E}[cf(X)] = c\,\mathbb{E}[f(X)]$
- $\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]$
- $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$

Sample Mean

- Studying the summation of random variables is particularly interesting in machine learning
- As an estimation of mean, we often use sample mean

$$\mathbb{E}[X] \approx \frac{1}{n} \sum_{i=1}^{n} x_i$$

- The law of large numbers guarantees that the sample mean converges to the true expectation as (i) the number of samples increases if (ii) all samples are independent to each other (and the variance is bounded)

$$\frac{1}{n} \sum_{i=1}^{n} x_i \xrightarrow{p} \mathbb{E}[X] \quad \text{as } n \to \infty \ .$$

Table of Contents

Bernoulli Distribution

Bernoulli distribution $\mathrm{Ber}(p)$ with parameter $p \in [0, 1]$

- Bernoulli r.v. $X \sim \mathrm{Ber}(p)$ has support $\{0, 1\}$ and

$$P(X = x) = p^x (1 - p)^{1-x} = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

- Mean $\mathbb{E}[X] = p$ and variance $\mathrm{Var}[X] = p(1 - p)$

Binomial Distribution

Binomial distribution $\text{Bin}(p, n)$ with parameters $p \in [0, 1]$ and $n \in \mathbb{N}$

- Binomial r.v. $X \sim \text{Bin}(p, n)$ has support $\{0, 1, ..., n\}$ and

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

  where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

- Note that a Binomial r.v. can be interpreted as the sum of n independent Bernoulli r.v.'s
- Mean $\mathbb{E}[X] = np$ and variance $\text{Var}[X] = np(1 - p)$
  - This computation can be considered as an example of law of large number as

$$\mathbb{E}\left[\frac{X}{n}\right] = p \quad \textit{and} \quad \text{Var}\left[\frac{X}{n}\right] = \frac{p(1-p)}{n} \underset{n\to\infty}{\rightarrow} 0$$

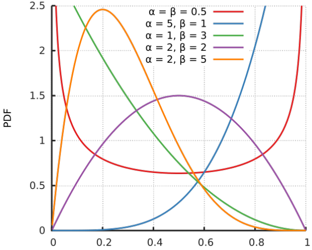
Beta Distribution

Beta distribution $\text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$

- Beta-distributed r.v. $X \sim \text{Beta}(\alpha, \beta)$ has support $[0, 1]$ and

$$P(X = x) \propto x^{\alpha-1}(1-x)^{\beta-1}$$

- Note that Beta distribution is often used to model parameter p of Bernoulli distribution
- Mean $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ and variance $\text{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



Example of Beta Distribution

Suppose that we want to predict next outcome of (possibly unfair) coin toss from observed sequence of very limited length: HHTHH (only five samples)

- We may assume the probability p of seeing head is some value close to $1/2$
- Such a prior model can be written as follows:

$$p \sim \text{Beta}(\alpha = 2, \beta = 2)$$

- Given the prior model, we may estimate the head probability as some value close to $1/2$ (Bayesian estimate) rather than $4/5$ (Frequentist estimate)

Gaussian Distribution or Normal Distribution

(Univariate) Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$

- Univariate Gaussian r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ has support $\mathbb{R}$ and

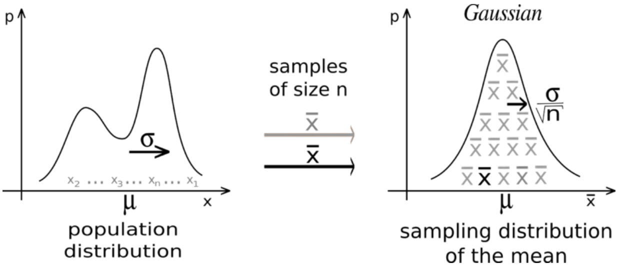
$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(X - \mu)^2\right) .$$

- Mean $\mathbb{E}[X] = \mu$ and variance $\mathrm{Var}[X] = \sigma^2$
- Elegant analytical properties, e.g., directly parameterized by mean and (co)-variance, central limit theorem, ...
- Maximum entropy, given values of the mean and the covariance matrix

Central Limit Theorem and Gaussian Distribution

Lindeberg-Levy central limit theorem (CLT)

- Let $(X_1, ..., X_n)$ be a random sequence of independent and identically distributed (i.i.d.) r.v.'s drawn from a distribution of expected value μ and finite variance σ^2
- Let $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean
- Then, $\sqrt{n}(\bar{X}_n - \mu)$ converges to $\mathcal{N}(0, \sigma^2)$ in distribution as $n \to \infty$



[from wikipedia]

Table of Contents

Joint distributions

Suppose that we have two random variables X and Y.

If we consider each of them separately, we will only need $F_X(x)$ and $F_Y(y)$.

But if we want to understand their relation, we need a more complicated structure known as the joint cumulative distribution of X and Y:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

where

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y) dy$$
$$F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y) dx$$

Joint and marginal probability density functions

- We can define the joint probability density function as

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

where $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) = 1$

- From the joint pdf, we can obtain marginal pdf (or marginal density) of X as

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

Conditionals and Bayes's rule

- We define the conditional probability density of Y given $X = x$ to be

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)}.$$

- From the conditional, we can derive Bayes's rule as

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}\left(x \mid y'\right) f_Y\left(y'\right) dy'}$$

Chain Rule

Chain rule: from the definition of conditional probabilities for random variables, one can show that

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n) &= f(x_n \mid x_1, x_2 \ldots, x_{n-1})\, f(x_1, x_2 \ldots, x_{n-1}) \\
&= f(x_n \mid x_1, x_2 \ldots, x_{n-1})\, f(x_{n-1} \mid x_1, x_2 \ldots, x_{n-2}) \\
&\qquad \times f(x_1, x_2 \ldots, x_{n-2}) \\
&= \ldots f(x_1) \prod_{i=2}^{n} f(x_i \mid x_1, \ldots, x_{i-1})
\end{aligned}
$$

Independence

- Two random variables X and Y are independent if

$$F_{XY}(x, y) = F_X(x)F_Y(y).$$

- Equivalently,

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

- or

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

Expectation and covariance

Suppose $g : \mathbb{R}^2 \to \mathbb{R}$ is a function of two random variables.

- Expected value of g is defined as

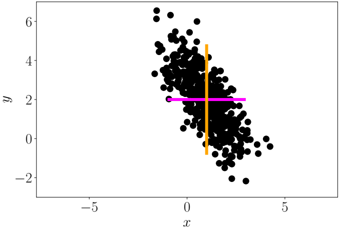
$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\,dxdy$$

- Covariance of two random variables is defined as

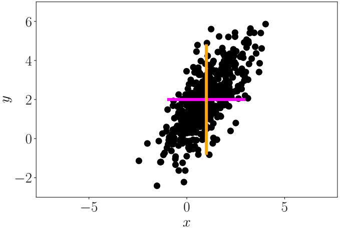
$$\mathrm{Cov}[X,Y] \triangleq E[(X - E[X])(Y - E[Y])]$$

- When $\mathrm{Cov}[X,Y] = 0$, we say that X and Y are uncorrelated.

Illustration of Covariance



(a) x and y are negatively correlated.



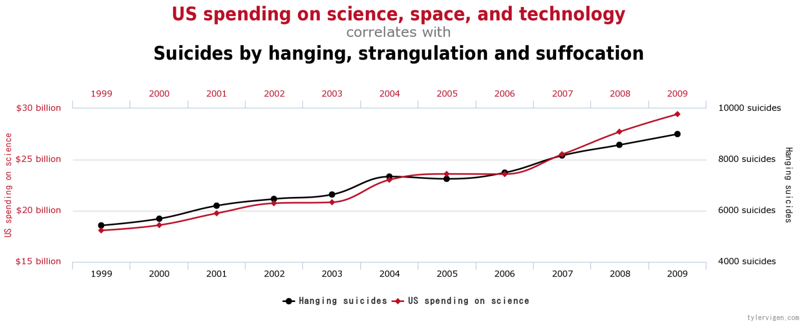
(b) x and y are positively correlated.

Remark (Correlation)

The correlation is the normalized covariance.

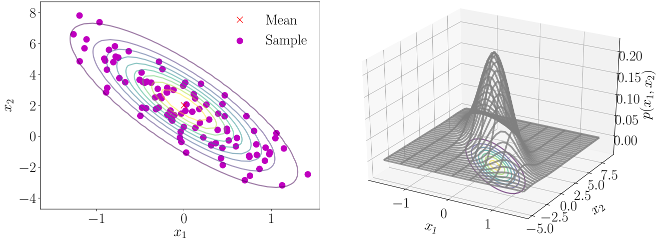
$$\mathrm{corr}[X, Y] = \frac{\mathrm{cov}[X, Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}} \in [-1, 1]$$

Correlation does not imply causation



[source: `https://tylervigen.com/spurious-correlations`]

Multivariate Gaussian PDF



- The PDF of multivariate Gaussian distribution is

$$p(\boldsymbol{X}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-D/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}))$$

Additional Reading

Chapter 2 of Textbook (Probabilistic Machine Learning: An Introduction)