Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Supplementary Material: Basic Probability

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2024)

License Information for Mathematics of Data Slides

Outline

- Review of probability theory
  1. Basic concepts
  2. Probability distributions

Basic concepts in probability theory

Definition (Sample space)

The sample space Ω of an experiment is the set of all possible outcomes of that experiment.

Example

If the experiment is tossing a coin, the sample set is the set $\{\text{head, tail}\}$.

Definition (Event)

An event E corresponds to a subset of the sample space; i.e., $E \subseteq \Omega$.

Definition (Probability measure)

Probability measure $P(E)$ maps event E from Ω onto the interval $[0, 1]$ and satisfies the following Kolmogorov axioms:

- $P(E) \geq 0$,
- $P(\Omega) = 1$ and
- $P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$, where $E_1, ..., E_n$ are mutually exclusive (i.e. $\bigcap_{i=1}^{n} E_i = \emptyset$). Such events are called *independent*.

Union of non-disjoint events

Definition (Principle of inclusion-exclusion)

The probability of the union of n events is

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq n} P(E_{i_1} \cap \ldots \cap E_{i_k}),$$

where the second sum is over all subsets of k events.

Union of non-disjoint events

Definition (Principle of inclusion-exclusion)

The probability of the union of n events is

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq n} P(E_{i_1} \cap \ldots \cap E_{i_k}),$$

where the second sum is over all subsets of k events.

Example

Suppose we throw two dices and ask what is the probability that the outcome is even or larger than 7. Let A and B denote the event of having an even number and the event of getting the number that exceeds 7, respectively. Then, $P(A) = \frac{1}{2}$, $P(B) = \frac{15}{36}$ and $P(A \cap B) = \frac{9}{36}$.

By the inclusion-exclusion principle, $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{2}{3}$.

The rules of probability

Let A and B denote two events in a sample space Ω, and let $P(B) \neq 0$.

Definition (Marginal probability)

The probability of an event (A) occuring ($P(A)$).

Definition (Joint probability)

$P(A, B)$ is the probability of event A and event B occuring. Symmetry property holds, i.e. $P(A, B) = P(B, A)$.

Definition (Conditional probability)

$P(B|A)$ is the probability that B will occur given that A has occurred.

Rules

- Sum rule: $P(A) = \sum_B P(A, B)$
- Product rule: $P(A, B) = P(B|A)P(A)$.

Bayes' rule

Bayes' rule

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

Constituents:

- $P(A)$, the prior probability, is the probability of A before B is observed.
- $P(A|B)$, the posterior probability, is the probability of A given B, i.e., after B is observed.
- $P(B|A)$ is the probability of observing B given A. As a function of A with B fixed, this is the likelihood.

Random variable

Definition

A real-valued random variable is a **function** that associates a value to the outcome of a randomized experiment $X : \Omega \to \mathbb{R}$.

Example

- Whether a coin flip was heads: a function from $\Omega = \{H, T\}$ to $\{0, 1\}$
- Number of heads in a sequence of n throws: function from $\Omega = \{H, T\}^n$ to $\{0, 1, ..., n\}$.

Discrete random variable

Probability mass function (Pmf)

The probability mass function is the function from values to its probability, $P_X(x) = P(X = x)$ for $x \in \mathcal{X}$ (i.e., a countable subset of the reals) with properties:

- $P_X(x) \geq 0$ for every $x \in \mathcal{X}$,
- $\sum_{x \in \mathcal{X}} P_X(x) = 1$

Example

Discrete distributions:

- Bernoulli distribution - distribution of a binary variable $x \in \{0, 1\}$; single parameter $\mu \in [0, 1]$ represents the probability of $x = 1$:
$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}.$$
- Binomial distribution - probability of observing m occurrences of 1 in a set of N samples from a Bernoulli distribution:
$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{1-m}.$$
- Other important discrete distributions: Categorical, Multinomial, Poisson, Geometric, Negative binomial, etc.

Probability density function (pdf)

◦ A continuous random variable can have uncountably infinite possible values.

Probability density function (pdf)

The probability density function of a continuous random variable X is an integrable function $p(x)$ satisfying the following:

1. The density is nonnegative: i.e., $p(x) \geq 0$ for any x,
2. Probabilities integrate to 1: i.e., $\int_{-\infty}^{\infty} p(x)dx = 1$,
3. The probability that x belongs to the interval $[a, b]$ is given by the integral of $p(x)$ over that interval: i.e.,

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

Basic rules of probability

1. Analog of sum rule: $p(x) = \int p(x, y)dy$
2. Product rule: $p(x, y) = p(y|x)p(x)$.

Expectations and variances

Definition (Expectation (1^{st} moment, mean))

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} xP(X = x) & \text{discrete} \\ \int_{-\infty}^{\infty} xp(x)dx & \text{continuous} \end{cases}$$

Definition (Variance (2^{nd} moment))

$$\mathbb{V}[X] = \begin{cases} \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 P(X = x) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 p(x)dx & \text{continuous} \end{cases}$$

Definition (Conditional expectation and Covariance)

$$\mathbb{E}[X|Y = y] = \sum_{x \in \mathcal{X}} xP(X = x|Y = y)$$

$$\text{cov}[x, y] = \mathbb{E}\Big[\Big(x - \mathbb{E}[X]\Big)\Big(y - \mathbb{E}[Y]\Big)\Big]$$

Probability distributions for continuous variables

Common distributions:

- Uniform
- Normal / Gaussian
- Beta
- Chi-Squared
- Exponential
- Gamma
- Laplace

Normal (Gaussian) Distribution

Gaussian distribution

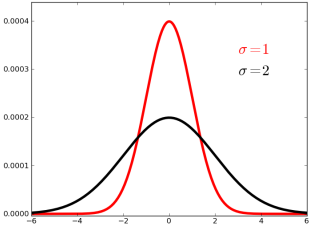
For $\mathbf{x} \in \mathbb{R}^d$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is the covariance matrix and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

- In the case of a single variable

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



Law of large numbers and central limit theorem

Theorem (Strong Law of Large Numbers)

Let X be a real-valued random variable with the finite first moment $\mathbb{E}[X]$, and let $X_1, X_2, ..., X_n$ be an infinite sequence of independent and identically distributed copies of X. Then the empirical average of this sequence $\bar{X}_n := \frac{1}{n}(X_1 + ... + X_n)$ converges almost surely to $\mathbb{E}[X]$ i.e., $P\Big(\text{lim}_{n\to\infty}\bar{X}_n = \mathbb{E}[X]\Big) = 1$.

Theorem (Central Limit Theorem)

Let $X_1, ...X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then the distribution of $\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}$ tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$,

$$P\left(\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right) \to \frac{1}{2\pi}\int_{-\infty} ae^{-x^2/2}dx$$

as $n \to \infty$.

- Intuitively, the sampling distribution of the mean will be close to Gaussian, if you just take enough independent samples.