Machine learning

Overview of probability theory

Hamid Beigy

Sharif University of Technology

October 31, 2021

Table of contents

Probability

- Probability theory is the study of uncertainty.
- Elements of probability
  - Sample space Ω : The set of all the outcomes of a random experiment.
  - Event space $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ (called events) are subsets of Ω.
  - Probability measure : A function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties,
    1. $P(A) \geq 0$, for all $A \in \mathcal{F}$.
    2. $P(\Omega) = 1$.
    3. If $A_1, A_2, \ldots$ are disjoint events (i.e.,$A_i \cap A_j = \emptyset$ whenever $i \neq j$),then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

  - Consider the following example.

Example (Tossing two coins)

In tossing two coins, we have

- The sample space equals to $\Omega = \{HH, HT, TT, TH\}$
- An event space $\mathcal{F}$ that only one head is a subset of Ω such as $\mathcal{F} = \{TH, HT\}$

Properties of probability

- If $A \subseteq B \implies P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$.
- $P(A \cup B) \leq P(A) + P(B)$. This property is called union bound.
- $P(\Omega \setminus A) = 1 - P(A)$.
- If $A_1, A_2, \ldots, A_k$ are disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$,then

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

  This property is called law of total probability.

Conditional probability and independence

- Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

  In other words, $P(A \mid B)$ is the probability measure of the event A after observing the occurrence of event B.

- Two events are called independent if and only if

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

  or equivalently, $P(A \mid B) = P(A)$.
  Therefore, independence is equivalent to saying that observing B does not have any effect on the probability of A.

What is probability?

- Classical definition (Laplace, 1814)

$$P(A) = \frac{N_A}{N}$$

  where N mutually exclusive equally likely outcomes, N_A of which result in the occurrence of A.

- Frequentist definition

$$P(A) = \lim_{N \to \infty} \frac{N_A}{N}$$

  or relative frequency of occurrence of A in infinite number of trials.

- Bayesian definition(de Finetti, 1930s)
  $P(A)$ is a degree of belief.

What is probability? (example)

- Suppose that you have a coin that has an unknown probability θ of coming up heads.
- We must determine this probability as accurately as possible using experimentation.
- Experimentation is to repeatedly tossing the coin. Let us denote the two possible outcomes of a single toss by 1 (for HEADS) and 0 (for TAILS).
- If you toss the coin m times, then you can record the outcomes as $x_1, \ldots, x_m$, where each $x_i \in \{0, 1\}$ and $P[x_i = 1] = \theta$ independently of all other x_i's.
- What would be a reasonable estimate of θ?
- In Frequentist view, by Law of Large Numbers, in a long sequence of independent coin tosses, the relative frequency of heads will eventually approach the true value of θ with high probability. Hence,

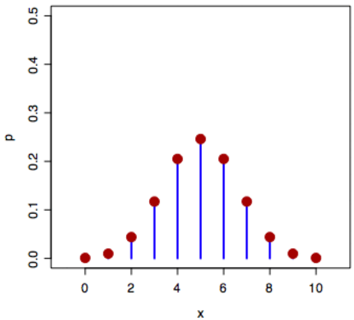
$$\hat{\theta} = \frac{1}{m} \sum_i x_i$$

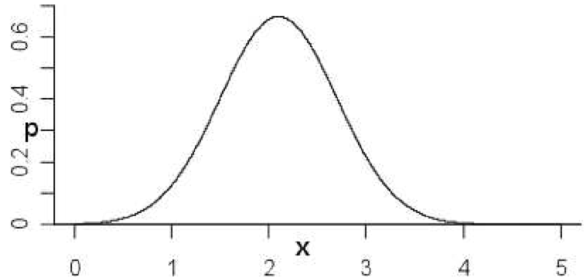
- In Bayesian view, θ is a random variable and has a distribution.

Random variables

- Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads.
- Here, the elements of the sample space Ω are 10-length sequences of heads and tails.
- However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails.
- Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails.
- These functions, under some technical conditions, are known as random variables.
- More formally, a random variable X is a function $X : \Omega \to \mathbb{R}$. Typically, we will denote random variables using upper case letters $X(\omega)$ or more simply X, where ω is an event.
- We will denote the value that a random variable X may take on using lower case letter x.

- A random variable can be discrete or continuous.

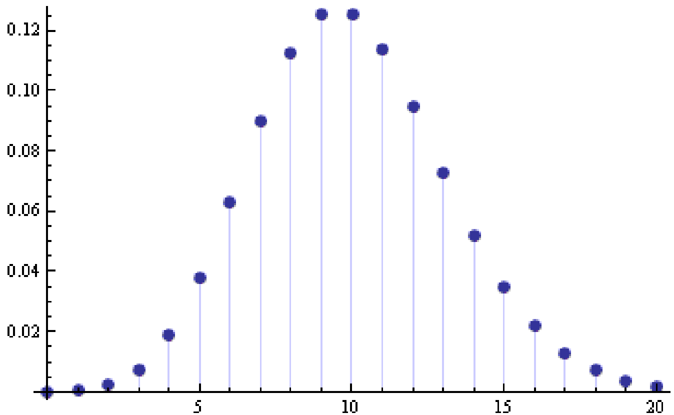




- A random variable is associated with a probability mass function or *probability distribution*.

- For a discrete random variable X, $p(x)$ denotes the probability that $p(X = x)$.
- $p(x)$ is called the probability mass function (PMF). This function has the following properties:

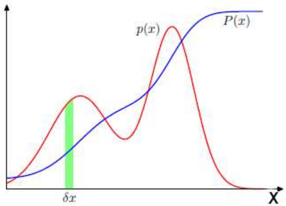
$$p(x) \geq 0$$
$$p(x) \leq 1$$
$$\sum_{x} p(x) = 1$$



- For a continuous random variable X, a probability $p(X = x)$ is meaningless.
- Instead we use $p(x)$ to denote the probability density function (PDF).

$$p(x) \geq 0$$
$$\int_x p(x) = 1$$

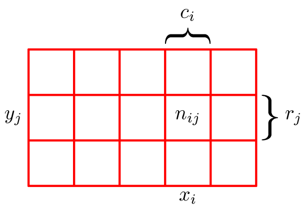
- Probability that a continuous random variable $X \in (x, x + \delta x)$ is $p(x)\delta x$ as $\delta x \to 0$.



- Probability that $X \in (-\infty, z)$ is given by the cumulative distribution function (CDF) $P(z)$, where

$$P(z) = p(X \leq z) = \int_{-\infty}^{z} p(x)dx$$
$$p(x) = z \left| \frac{dP(z)}{dz} \right|_{z=x}$$

- Joint probability $p(X, Y)$ models probability of co-occurrence of two random variables X and Y.
- Let n_{ij} be the number of times events x_i and y_j simultaneously occur.



- Let $N = \sum_i \sum_j n_{ij}$.
- Joint probability is

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}.$$

- Let $c_i = \sum_j n_{ij}$, and $r_j = \sum_i n_{ij}$.
- The probability of X irrespective of Y is

$$p(X = x_i) = \frac{c_i}{N}.$$

- Therefore, we can marginalize or sum over Y, i.e. $p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$.
- For discrete random variables, we have $\sum_x \sum_y p(X = x, Y = y) = 1$.
- For continuous random variables, we have $\int_x \int_y p(X = x, Y = y) = 1$.

Marginalization

- Consider only instances where the fraction of instances $Y = y_j$ when $X = x_i$.
- This is conditional probability and is written $p(Y = y_j | X = x_i)$, the probability of Y given X.

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

- Now consider

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

- If two events are independent, $p(X, Y) = p(X)p(Y)$ and $p(X|Y) = p(X)$
- Sum rule $p(X) = \sum_Y p(X, Y)$
- Product rule $p(X, Y) = p(Y|X)p(X)$

- Expectation, expected value, or mean of a random variable X, denoted by $\mathbb{E}[X]$, is the average value of X in a large number of experiments.

$$\mathbb{E}[X] = \sum_x p(x)x$$

or

$$\mathbb{E}[X] = \int p(x)xdx$$

- The definition of Expectation also applies to functions of random variables (e.g., $\mathbb{E}[f(x)]$)
- Linearity of expectation

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

Variance and and Covariance

- Variance (σ^2) measures how much X varies around the expected value and is defined as.

$$Var(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mu^2$$

- Standard deviation : $std[X] = \sqrt{Var[X]} = \sigma$.
- Covariance of two random variables X and Y indicates the relationship between two random variables X and Y.

$$Cov(X, Y) = \mathbb{E}_{X,Y}\left[(X - \mathbb{E}[X])^\top (Y - \mathbb{E}[Y])\right]$$

Probability distributions

We will use these probability distributions extensively to model data as well as parameters

- Some discrete distributions and what they can model:
  1. Bernoulli : Binary numbers, e.g., outcome (head/tail, 0/1) of a coin toss
  2. Binomial : Bounded non-negative integers, e.g., the number of heads in n coin tosses
  3. Multinomial : One of $K(> 2)$ possibilities, e.g., outcome of a dice roll
  4. Poisson : Non-negative integers, e.g., the number of words in a document
- Some continuous distributions and what they can model:
  1. Uniform: Numbers defined over a fixed range
  2. Beta: Numbers between 0 and 1, e.g., probability of head for a biased coin
  3. Gamma: Positive unbounded real numbers
  4. Dirichlet : Vectors that sum of 1 (fraction of data points in different clusters)
  5. Gaussian: Real-valued numbers or real-valued vectors

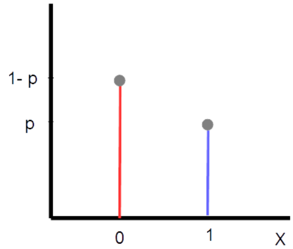
Probability distributions

Discrete distributions

- Distribution over a binary random variable $x \in \{0, 1\}$, like a coin-toss outcome
- Defined by a probability parameter $p \in (0, 1)$.

$$p[X = 1] = p$$
$$p[X = 0] = 1 - p$$

- Distribution defined as: $Bernoulli(x; p) = p^x(1 - p)^{1-x}$

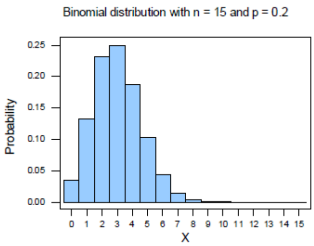


- The expected value and the variance of X are equal to

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

- Distribution over number of successes m in a number of trials
- Defined by two parameters: total number of trials (N) and probability of each success $p \in (0, 1)$.
- We can think Binomial as multiple independent Bernoulli trials
- Distribution defined as

$$Binomial(m; N, p) = \binom{N}{m} p^m (1-p)^{N-m}$$



- The expected value and the variance of m are equal to

$$\mathbb{E}[m] = Np$$
$$Var(m) = Np(1-p)$$

- Consider a generalization of Bernoulli where the outcome of a random event is one of K mutually exclusive and exhaustive states, each of which has a probability of occurring q_i where $\sum_{i=1}^{K} q_i = 1$.
- Suppose that n such trials are made where outcome i occurred n_i times with $\sum_{i=1}^{K} n_i = n$.
- The joint distribution of $n_1, n_2, \ldots, n_K$ is multinomial

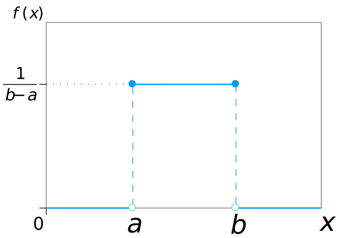
$$P(n_1, n_2, \ldots, n_K) = n! \prod_{i=1}^{K} \frac{q_i^{n_i}}{n_i!}$$

Probability distributions

Continuous distributions

- Models a continuous random variable X distributed uniformly over a finite interval $[a, b]$.

$$Uniform(X; a, b) = \frac{1}{b-a}$$



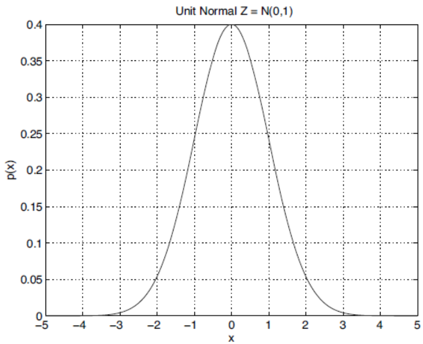
- The expected value and the variance of X are equal to

$$\mathbb{E}[X] = \frac{b+a}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

- For 1-dimensional normal or Gaussian distributed variable x with mean μ and variance σ^2, denoted as $\mathcal{N}(x; \mu, \sigma^2)$, we have

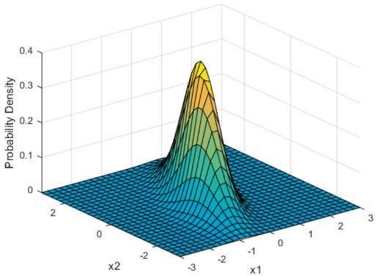
$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



- Mean: $\mathbb{E}[X] = \mu$
- Variance: $var[X] = \sigma^2$
- Precision (inverse variance): $\beta = \frac{1}{\sigma^2}$

- Distribution over a multivariate random variables vector $x \in \mathbb{R}^D$ of real numbers
- Defined by a mean vector $\mu \in \mathbb{R}^D$ and a $D \times D$ covariance matrix Σ

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$



- The covariance matrix Σ must be symmetric and positive definite
  1. All eigenvalues are positive
  2. $z^\top \Sigma z > 0$ for any real vector z.
- Often we parameterize a multivariate Gaussian using the inverse of the covariance matrix, i.e., the precision matrix $\Lambda = \Sigma^{-1}$.

Bayes theorem

- Bayes theorem

$$
\begin{aligned}
p(Y|X) &= \frac{P(X|Y)P(Y)}{P(X)} \\
&= \frac{P(X|Y)P(Y)}{\sum_Y p(X|Y)p(Y)}
\end{aligned}
$$

- $p(Y)$ is called prior of Y. This is information we have before observing anything about the Y that was drawn.
- $p(Y|X)$ is called posterior probability, or simply posterior. This is the distribution of Y after observing X.
- $p(X|Y)$ is called likelihood of X and is the conditional probability that an event Y has the associated observation X.
- $p(X)$ is called evidence and is the marginal probability that an observation X is seen.
- In other words

$$
posterior = \frac{prior \times likelihood}{evidence}.
$$

- In many learning scenarios, the learner considers some set $\mathcal{Y}$ and is interested in finding the most probable $Y \in \mathcal{Y}$ given observed data X.
- This is called maximum a posteriori estimation (MAP) and can be estimated using Bayes theorem.

$$
\begin{aligned}
Y_{MAP} &= \underset{Y \in \mathcal{Y}}{argmax} \; p(Y|X) \\
&= \underset{Y \in \mathcal{Y}}{argmax} \; \frac{P(X|Y)P(Y)}{P(X)} \\
&= \underset{Y \in \mathcal{Y}}{argmax} \; P(X|Y)P(Y)
\end{aligned}
$$

- $P(X)$ is dropped because it is constant and independent of Y.

$$
\begin{aligned}
Y_{MAP} &= \underset{Y \in \mathcal{Y}}{argmax} \; P(X|Y)P(Y) \\
&= \underset{Y \in \mathcal{Y}}{argmax} \; \{\log P(X|Y) + \log P(Y)\} \\
&= \underset{Y \in \mathcal{Y}}{argmin} \; \{-\log P(X|Y) - \log P(Y)\}
\end{aligned}
$$

Maximum likelihood estimation

- In some cases, we will assume that every $Y \in \mathcal{Y}$ is equally probable.
- This is called maximum likelihood estimation.

$$\begin{aligned} Y_{ML} &= \underset{Y \in \mathcal{Y}}{argmax} \quad P(X|Y) \\ &= \underset{Y \in \mathcal{Y}}{argmax} \quad \log P(X|Y) \\ &= \underset{Y \in \mathcal{Y}}{argmin} \quad \{-\log P(X|Y)\} \end{aligned}$$

- Let $x_1, x_2, \ldots, x_N$ be random samples drawn from $p(X, Y)$.
- Assuming statistical independence between the different samples,we can form $p(X|Y)$ as

$$p(X|Y) = p(x_1, x_2, \ldots, x_N|Y) = \prod_{n=1}^{N} p(x_n|Y)$$

- This method estimates Y so that $p(X|Y)$ takes its maximum value.

$$Y_{ML} = \underset{Y \in \mathcal{Y}}{argmax} \quad \prod_{n=1}^{N} p(x_n|Y)$$

- A necessary condition that Y_{ML} must satisfy in order to be a maximum is the gradient of the likelihood function with respect to Y to be zero.

$$\frac{\partial \prod_{n=1}^{N} p(x_n|Y)}{\partial Y} = 0$$

- Because of the monotonicity of the logarithmic function, we define the log likelihood function as

$$L(Y) = \ln \prod_{n=1}^{N} p(x_n|Y)$$

- Equivalently, we have

$$\begin{aligned} \frac{\partial L(Y)}{\partial Y} &= \sum_{n=1}^{N} \frac{\partial \ln p(x_n|Y)}{\partial Y} \\ &= \sum_{n=1}^{N} \frac{1}{p(x_n|Y)} \frac{\partial p(x_n|Y)}{\partial Y} = 0 \end{aligned}$$

Readings

1. Chapter 2 of Pattern Recognition and Machine Learning Book (Bishop 2006).
2. Chapter 2 of Machine Learning: A probabilistic perspective (Murphy 2012).
3. Chapter 1 of Probabilistic Machine Learning: An introduction (Murphy 2022).

- Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.
- Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- – (2022). *Probabilistic Machine Learning: An introduction*. MIT Press.

Questions?