Machine Learning: Probability and Statistics

Hao Tang

January 16, 2026

- I assume you have taken DMP (INFR08031).

- In this session, we will
  - quickly review the basic concepts (i.e., the language of probability)
  - cover the Gaussian distribution
  - introduce maximum likelihood estimation

Warm up: Rolling a dice

- What the probability to get an even number when we roll a (fair) dice?

Warm up: Rolling a dice

- What the probability to get an even number when we roll a (fair) dice?
- Before we talk about the answer, here's how we describe the probability

$$\mathbb{P}[X \in \{2, 4, 6\}] \tag{1}$$

  where X is a random variable where the outcome is the face of a dice.
- Think of this as the syntax of probability.

Warm up: Rolling a dice

- We know it's a fair dice, so the probability mass function $p(x)$ of rolling a dice is

x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6

- Using the general fact that

$$\mathbb{P}[X \in S] = \sum_{x \in S} p(x), \tag{2}$$

we now have

$$\mathbb{P}[X \in \{2, 4, 6\}] = p(2) + p(4) + p(6) = 1/2. \tag{3}$$

Probability measures

- The usual mathematical development of probability theory is to define random variables, probability measures, and then probability mass functions (or probability density functions in the continuous case).
- A probability measure is a function $\mathbb{P} : \Omega \to \mathbb{R}$ that satisfies
  - $0 \leq \mathbb{P}[X] \leq 1$ for any $X \subseteq \Omega$
  - $\mathbb{P}[\Omega] = 1$
  - $\mathbb{P}[A_1 \cup A_2] = \mathbb{P}[A_1] + \mathbb{P}[A_2]$ if $A_1 \cap A_2 = \emptyset$
- A probability mass function for a discrete random variable is defined as $p(x) = \mathbb{P}[X = x]$.
- A probability density function for a continuous random variable is defined as $p(x) = \frac{d}{dx}\mathbb{P}[X \leq x]$.

Probability distributions

- Instead of separating probability mass functions and probability density functions for discrete and continuous cases, we will just call them probability distributions.
- In other words, when we say that

$$p(x) \text{ is a distribution}$$

we mean that

$$p(x) \geq 0 \quad \text{and} \quad \sum_x p(x) = 1. \tag{4}$$

- In machine learning, we mostly work with probability distributions, and less with probability measures.

Parameterizing probability measures

- Consider learning a probability measure.
- In principle, we plan directly parameterize a probability measure, but it is a lot easier to parameterize its distribution.
- The probability measure is parameterized once the distribution is, because

$$\mathbb{P}[X \in S] = \sum_{x \in S} p(x) \tag{5}$$

- This might not make much sense right now, and we will come back to this in later sessions.

Expectation

Expectation

- The expectation of a random variable x is defined as

$$\mathbb{E}[x] = \sum_x xp(x). \tag{6}$$

- Note that the expectation $\mathbb{E}[x]$ is not a function of x.
- The equation

$$\mathbb{E}[f(x)] = \sum_x f(x)p(x) \tag{7}$$

is known as the law of unconscious statistician (LOTUS).

Mean and covariance

Mean and covariance

- The mean of a distribution is defined as

$$\mu = \mathbb{E}[x]. \tag{8}$$

Mean and covariance

- The mean of a distribution is defined as

$$\mu = \mathbb{E}[x]. \tag{8}$$

- In the 1D case, the variance is defined as

$$\sigma^2 = \mathbb{E}[(x - \mu)^2]. \tag{9}$$

Mean and covariance

- The mean of a distribution is defined as

$$\mu = \mathbb{E}[x]. \tag{8}$$

- In the 1D case, the variance is defined as

$$\sigma^2 = \mathbb{E}[(x - \mu)^2]. \tag{9}$$

- When x is a vector, the covariance matrix of a distribution is defined as

$$\Sigma = \mathbb{E}[(x - \mu)(x - \mu)^\top]. \tag{10}$$

- In particular, the diagonal entries of the covariance matrix Σ is called the variance.

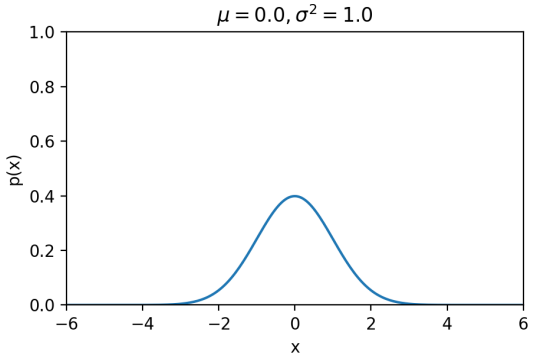
1D Gaussian

- A variable $x \in \mathbb{R}$ is said to follow a Gaussian distribution if

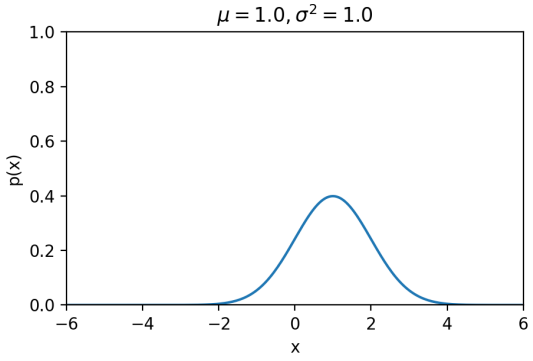
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \tag{11}$$

  where $\mu \in \mathbb{R}$ is the mean and $\sigma^2 \in \mathbb{R}$ is the variance.

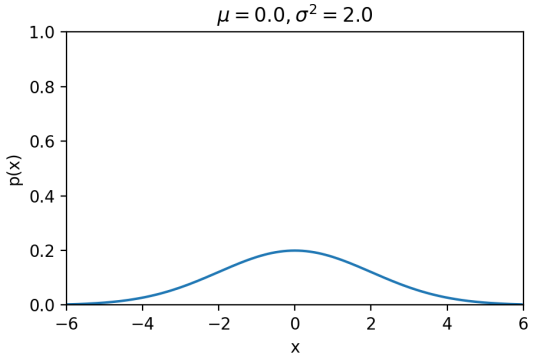
1D Gaussian



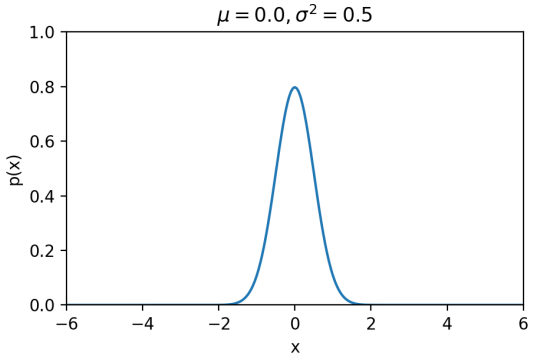
1D Gaussian



1D Gaussian



1D Gaussian



Sampling notation

- We say that a is drawn from a Gaussian if

$$a \sim \mathcal{N}(\mu, \sigma^2). \tag{12}$$

  It simply means

$$p(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(a-\mu)^2\right). \tag{13}$$

- This notation also commonly appears in expectations, such as in

$$\mathbb{E}_{x\sim p(x)}[f(x)], \tag{14}$$

  to make explicit what variables are being integrated.
- In particular $\mathbb{E}[f(x)]$, $\mathbb{E}_x[f(x)]$, $\mathbb{E}_p[f(x)]$, and $\mathbb{E}_{x\sim p(x)}[f(x)]$, all mean the same thing.

Joint, marginal, and conditional distribution

- The distribution $p(x, y)$ is referred to as the joint distribution of x and y.
- Given $p(x, y)$, the distribution $p(x) = \sum_y p(x, y)$ and is referred to as the marginal distribution. The act of computing the sum is known as marginalization.
- The conditional distribution $p(y|x)$ is defined as $\frac{p(x,y)}{p(x)}$.

Bayes rule

Bayes rule

- The equation

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{15}$$

  is known as the Bayes rule.

- It is commonly used to invert conditional probabilities, i.e., from $p(y|x)$ to $p(x|y)$.

Bayes rule

- The equation

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{15}$$

  is known as the Bayes rule.

- It is commonly used to invert conditional probabilities, i.e., from $p(y|x)$ to $p(x|y)$.

- The equation

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_{y'} p(x|y')p(y')} \tag{16}$$

  is actually more useful in practice, because we can compute $p(y|x)$ as long as we have $p(x|y)$ and $p(y)$ defined.

Statistical independence

Statistical independence

- We say that x and y are statistically independent if

$$p(x, y) = p(x)p(y). \tag{17}$$

Statistical independence

- We say that x and y are statistically independent if

$$p(x, y) = p(x)p(y). \tag{17}$$

- In other words,

$$p(y|x) = p(y) \tag{18}$$

  if x and y are independent.
- To put this in words, x and y are statistically independent if knowing x does not tell us anything more about y.

Independent and identically distributed variables

- The samples $x_1, x_2, \ldots, x_n$ are called independent and identically distributed (i.i.d.) samples if $x_1, x_2, \ldots, x_n$ are mutually independent and are drawn from the same distribution.

- In particular,

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n) = \prod_{i=1}^{n} p(x_i). \tag{19}$$

Distributions on vectors

- We will work with vectors a lot in this course.
- A distribution on a d-dimensional vector is simply a joint distribution on d random variables.
- In other words, instead of writing

$$p(x_1, x_2, \ldots, x_d) \tag{20}$$

we will just write

$$p(x), \tag{21}$$

where $x = \begin{bmatrix} x_1 & x_2 & \ldots & x_d \end{bmatrix}^\top$.

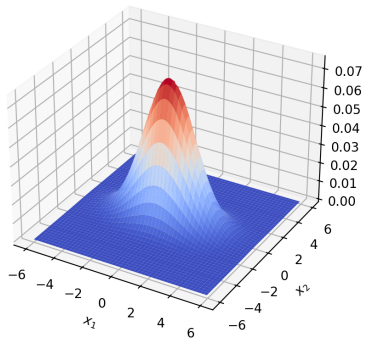
Multivariate Gaussian

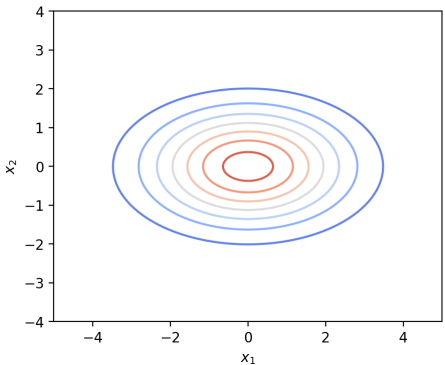
- A variable $x \in \mathbb{R}^d$ is said to follow a Gaussian distribution if

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right), \tag{22}$$

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

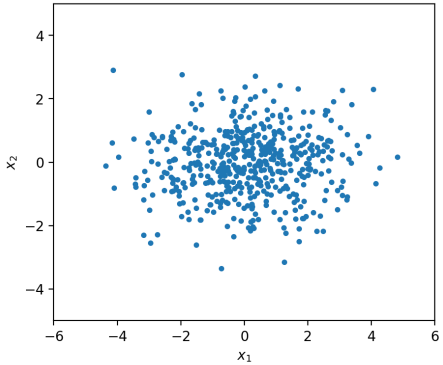
2D Gaussian


0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−6
−4
−2
0
2
4
6
x_1
x_2


4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
x_1
x_2

Estimation

Estimation



Estimation

- When given a data set, we often want to know how the data is generated.
- Guessing the underlying distribution for a data set is often called estimation.
- This question is often solved by first assuming a distribution and estimating the parameters of the distribution.

Maximum likelihood

- Maximum likelihood is a way to estimate parameters.
- A likelihood is a function of the parameter, not the data.
- A maximum likelihood estimator is the parameter that maximizes the likelihood.

Maximum likelihood estimation of a Gaussian mean

What is the maximum likelihood estimate of the Gaussian mean μ given n i.i.d. 1D Gaussian samples $x_1, \ldots, x_n$?

Maximum likelihood estimation of a Gaussian mean

- After going through all the definitions, we can finally understand what the question means.
- The value $p(x_1, x_2, \ldots, x_n)$ is called the likelihood, and it's a function fo the parameter, in this case, the mean μ.
- Due to the i.i.d. property, the likelihood can be written as

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) \tag{23}$$

- To actually solve it requires doing some calculus, and we will come back to this in later sessions.

Further reading

- Statistics 110: Probability
  https://www.youtube.com/playlist?list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo
- Blitzstein and Hwang, "Introduction to Probability," CRC Press, 2019
- Capinski et al., "Measure, Integral and Probability," Springer, 2004
- Stigler, "The epic story of maximum likelihood," Statistical Science, 2007