

DDA3020 Machine Learning

Lecture 02 Probability and Information Theory

Jicong Fan
SDS, CUHK-SZ

September 04/09,2025

Outline

- 1 Probability, event, random variables
- 2 Probability of discrete random variables
- 3 Probability of continuous random variables
- 4 Some popular distributions
- 5 Information theory

- 1 Probability, event, random variables
- 2 Probability of discrete random variables
- 3 Probability of continuous random variables
- 4 Some popular distributions
- 5 Information theory

Random experiment, sample space, event

- **Random experiment**: we describe a random experiment by its **procedure** and observations of its **outcomes**. For example, we toss a coin 2 times, and observe which side is up after each toss.
- **Sample space**: All possible outcomes of the random experiment form a sample space S . For the above coin toss example, we define

$$S = \{(Head, Head), (Head, Tail), (Tail, Head), (Tail, Tail)\}.$$

- **Event**: A **subset** of sample space S , denoted as A , can be called as an event in a random experiment, *i.e.*, $A \subset S$. In the above example, we define an event A as *at least one head up*, then it can be represented by

$$A = \{(Head, Head), (Head, Tail), (Tail, Head)\} \subset S.$$

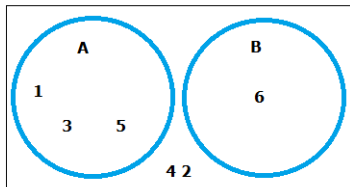
Probability of events

Assuming events $A \subset S$ and $B \subset S$, the probabilities of events related with and must satisfy,

- $P(A) \geq 0$
- $P(S) = 1$
- If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$;
otherwise, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

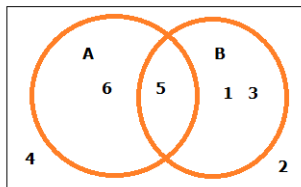
Disjoint Events

Event A: Get an odd Number
Event B: Get a 6



Overlapping Events

Event A: Get a number over 4
Event B: Get an odd number



Random variables

- A **random variable** is a real valued function from the sample space S to a real space \mathbb{R} , as follows:

$$X : S \rightarrow \mathbb{R}$$

- Still take the 2-times coin toss as an example, if we define the random variable as the number of tails, then we have

$$X((H, H)) = 0, X((H, T)) = 1, X((T, H)) = 1, X((T, T)) = 2.$$

Then, the output space of X is denoted as $\{0, 1, 2\}$, also called **state space** \mathcal{X} .

- There are two types of random variables:
 - **Discrete**: \mathcal{X} is discrete
 - **Continuous**: \mathcal{X} is continuous

- 1 Probability, event, random variables
- 2 Probability of discrete random variables
- 3 Probability of continuous random variables
- 4 Some popular distributions
- 5 Information theory

Probability of discrete random variables

- **Probability of discrete random variable** describes the chance of each state x in \mathcal{X} for random variable X in a random experiment, denoted as

$$P(X = x), x \in \mathcal{X}.$$

- It should satisfy

$$P(X = x) \geq 0, \forall x \in \mathcal{X}; \sum_{x \in \mathcal{X}} P(X = x) = 1.$$

Probability of discrete random variables

- Joint probability of two random variables X and Y :

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y) = P(Y = y|X = x)P(X = x)$$

- Marginal probability of X and Y :

$$P(X = x) = \sum_{y \in \mathcal{Y}} P(X = x, Y = y)$$

$$P(Y = y) = \sum_{x \in \mathcal{X}} P(X = x, Y = y)$$

- Conditional probability:

$$P(X = x|Y = y) = P(X = x, Y = y)/P(Y = y)$$

$$P(Y = y|X = x) = P(X = x, Y = y)/P(X = x)$$

Bayes rule and application on medical diagnosis

- **Bayes Rule:** Combining the definition of conditional probability with the product and sum rules yields Bayes rule, as follows:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y|X = x)}{\sum_{x' \in \mathcal{X}} P(X = x')P(Y = y|X = x')} \quad (1)$$

- Suppose that you do a medical test for breast cancer, the test result could be *positive* or *negative*. We denote $x = 1$ as the event of a positive test and $x = 0$ as the event of a negative test. We denote $y = 1$ as the event of having breast cancer and $y = 0$ as the event of no breast cancer.
- Suppose that if one has breast cancer, the test will be positive with the probability 0.8, *i.e.*,

$$P(x = 1|y = 1) = 0.8. \quad (2)$$

- Then, if one gets a positive test result, what is the probability of having breast cancer? $P(y = 1|x = 1) = 0.8?$

Bayes rule and application on medical diagnosis

- It is **WRONG**! It ignores the prior probability of having breast cancer.
- According to statistics, the average risk of a woman in the United States developing breast cancer sometime in her life is about 13%, *i.e.*,

$$P(y = 1) = 0.13. \quad (3)$$

- We also need to take into account the fact that the test may be a **false positive** or **false alarm**. Unfortunately, such false positives are quite likely (with current screening technology):

$$P(x = 1|y = 0) = 0.1. \quad (4)$$

- Combining all the above probabilities using the Bayes rule, we can compute

$$\begin{aligned} P(y = 1|x = 1) &= \frac{P(x = 1|y = 1)P(y = 1)}{P(x = 1|y = 1)P(y = 1) + P(x = 1|y = 0)P(y = 0)} \\ &= \frac{0.8 \times 0.13}{0.8 \times 0.13 + 0.1 \times 0.87} = 0.5445. \end{aligned} \quad (5)$$

It tells that if you test positive, you have about a 54% chance of really having breast cancer!

Independent random variables

- **Independent**: If X and Y are independent, denoted as $X \perp Y$, then the joint probability can be represented as the product of two marginals, *i.e.*,

$$X \perp Y \iff P(X, Y) = P(X)P(Y).$$

- Given the above independence, we can use fewer parameters to define a joint probability. Suppose that X has 3 states, Y has 4 states, then we need $3 - 1 = 2$ and $4 - 1 = 3$ free parameters to define $P(X)$ and $P(Y)$, respectively.
- If without independence, how many free parameters do we need to define the joint probability $P(X, Y)$? $(3 \times 4) - 1 = 11$.
- If given the independence, *i.e.*, $P(X, Y) = P(X)P(Y)$, how many free parameters do we need? $(3 - 1) + (4 - 1) = 5$.

Expectation and variance of discrete random variables

- **Expectation** (or mean): $E[X] = \sum_{x \in \mathcal{X}} xP(X = x)$
- Expectation of a function: $E[f(X)] = \sum_{x \in \mathcal{X}} f(x)P(X = x)$
- **Moments**: expectation of power of X : $M_k = E[X^k]$
- **Variance**: Average (squared) fluctuation from the mean

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2 = M_2 - M_1^2. \quad (6)$$

- **Standard deviation**: Square root of variance, *i.e.*,

$$\text{Std} = \sqrt{\text{Var}(X)}. \quad (7)$$

- **Exercise** : For the random variable of the number of tails in the random experiment of 2-times coin toss, please compute the above values.

- 1 Probability, event, random variables
- 2 Probability of discrete random variables
- 3 Probability of continuous random variables**
- 4 Some popular distributions
- 5 Information theory

Continuous random variables

- A random variable X is **continuous** if its state space \mathcal{X} is uncountable.
- In this case, $P(X = x) = 0$ for each x .
- If $p_X(x)$ is a **probability density function** (PDF) for X , then

$$P(a < X < b) = \int_a^b p(x)dx \quad (8)$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \quad (9)$$

- The **cumulative distribution function** (CDF) is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^x p(s)ds$.

Bivariate continuous distributions: Marginalization, Conditioning, and Independence

- $p_{X,Y}(x,y)$, joint probability density function of X and Y
- $\int_x \int_y p(x,y) dx dy = 1$
- **Marginal distribution:** $p(x) = \int_{-\infty}^{\infty} p(x,y) dy$
- **Conditional distribution:** $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0$! Formally, conditional probability in the continuous case can be derived using infinitesimal events.
- **Independence:** X and Y are independent if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$

Expectation and variance of continuous random variables

Similar to that of discrete random variables, only change the summation \sum to the integral \int .

- **Expectation** (or mean): $\mu = E[X] = \int_{\mathcal{X}} x \cdot p(x) dx$
- **Moments**: expectation of power of X : $M_k = E[X^k] = \int_{\mathcal{X}} x^k \cdot p(x) dx$
- **Variance**: Average (squared) fluctuation from the mean

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2 = M_2 - M_1^2. \quad (10)$$

- **Standard deviation**: Square root of variance, *i.e.*,

$$\text{Std} = \sqrt{\text{Var}(X)}. \quad (11)$$

- 1 Probability, event, random variables
- 2 Probability of discrete random variables
- 3 Probability of continuous random variables
- 4 Some popular distributions**
- 5 Information theory

Binary variables (discrete r.v.)

- We first consider the probability of a binary random variable $x \in \{0, 1\}$. Suppose that you toss a coin, and $x = 1$ denotes the event of ‘heads’, while $x = 0$ indicates the event of ‘tails’.
- The probability of $x = 1$ is described by a parameter μ ,

$$p(x = 1|\mu) = \mu, \quad (12)$$

where $\mu \in [0, 1]$, and we can obtain that $p(x = 0|\mu) = 1 - \mu$.

- The probability distribution over x can therefore be written in the form

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}, \quad (13)$$

which is called **Bernoulli** distribution.

- Its mean and variance are

$$E[x] = \sum_x x \text{Bern}(x|\mu) = \mu, \quad (14)$$

$$\text{var}[x] = E[(x - \mu)^2] = \mu(1 - \mu) \quad (15)$$

Binomial variables

- Imagine that you toss the coin N times, and each tossing follows the Bernoulli distribution $p(x|\mu)$. We denote the variable m as the number of heads, then its distribution is formulated as follows:

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}, \quad (16)$$

which is called **Binomial** distribution, where

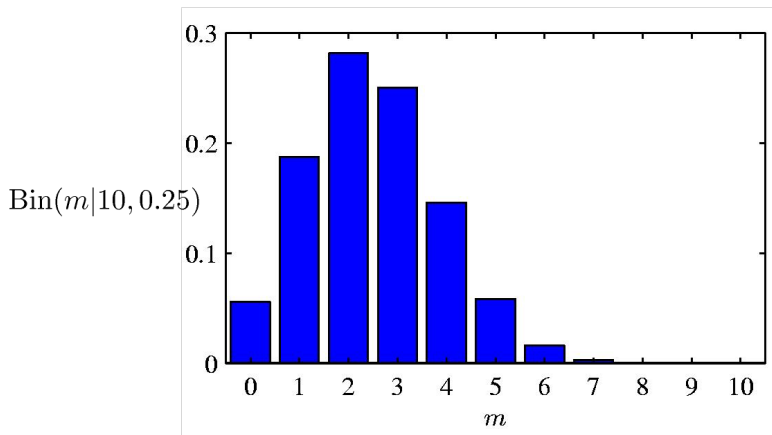
$$\binom{N}{m} = \frac{N!}{(N-m)!m!}. \quad (17)$$

- Its mean and variance are

$$E[m] = \sum_{m=0}^N m \text{Bin}(m|N, \mu) = N\mu, \quad (18)$$

$$\text{var}[m] = E[(m - N\mu)^2] = N\mu(1 - \mu). \quad (19)$$

Binomial distribution



Gaussian distribution (continuous r.v.)

- The **Gaussian**, also known as the **normal** distribution, is a widely used model for the distribution of **continuous** variables. In the case of a single variable x , the Gaussian distribution can be written in the form

$$\mathcal{N}(x|\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad (20)$$

where μ is the **mean** and σ^2 is the **variance**.

- For a D -dimensional vector \mathbf{x} , the **multivariate Gaussian** distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}\right), \quad (21)$$

where $\boldsymbol{\mu}$ is a D -dimensional **mean vector**, and $\boldsymbol{\Sigma}$ is a $D \times D$ **covariance matrix**, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

- 1 Probability, event, random variables
- 2 Probability of discrete random variables
- 3 Probability of continuous random variables
- 4 Some popular distributions
- 5 Information theory

What is information

- To know what is **information**, we need to know a bit who is **Claude Shannon** and his information theory (**let's see a video**).
- Shannon defined the measure that quantifies the **uncertainty** of an event with given probability - **a bit**.
- For a discrete random variable (a source) with a finite alphabet, as follows

$$\mathcal{X} = \{x_0, \dots, x_k, \dots, x_S\},$$

where the probability of each symbol is given by $P(X = x_k) = p_k$.

- We define the **information** as

$$I(x_k) = \log \frac{1}{p_k} = -\log(p_k).$$

If the logarithm is base 2, information is given in bits.

- Note that $I(x_k) \geq 0$, *i.e.*, non-negative. The equality only holds when $p_k = 1$, which means there is **no uncertainty**, then **no information**.

What is information

- It represents the *surprise* of seeing the outcome (a highly probable outcome is not surprising).

event	probability	surprise
one equals one	1	0 bits
wrong guess on a 4-choice question	$3/4$	0.415 bits
correct guess on true-false question	$1/2$	1 bit
correct guess on a 4-choice question	$1/4$	2 bits
seven on a pair of dice	$6/36$	2.58 bits
win any prize at Euromilhões	$1/24$	4.585 bits
win Euromilhões Jackpot	$\approx 1/76$ million	≈ 26 bits
gamma ray burst mass extinction today	$< 2.7 \cdot 10^{-12}$	> 38 bits

Larger surprise/uncertainty, more information/bits.

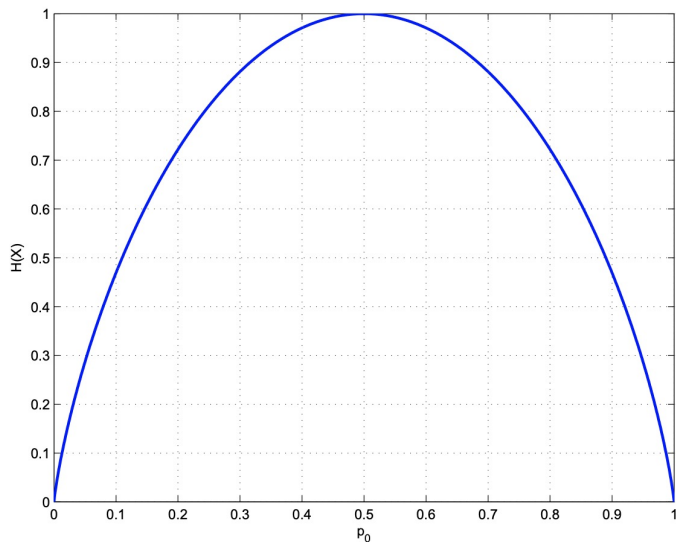
- **Entropy** is defined as the **expected value of information** from a source,

$$\begin{aligned}H_P(\mathcal{X}) &= E[I(x_k)] \\&= \sum_{x_k \in \mathcal{X}} p_k \cdot I(x_k) \\&= - \sum_{x_k \in \mathcal{X}} p_k \cdot \log(p_k).\end{aligned}$$

- Let $\mathcal{X} = \{0, 1\}$ be a binary source with p_0 and p_1 being the probability of symbols $x_0 = 0$ and $x_1 = 1$, respectively, we have

$$\begin{aligned}H_P(\mathcal{X}) &= E[I(x_k)] \\&= -p_0 \log p_0 - p_1 \log p_1 \\&= -p_0 \log p_0 - (1 - p_0) \log (1 - p_0)\end{aligned}$$

Entropy of binary source



Cross Entropy

- **Cross-entropy** builds upon the idea of entropy. It calculates the number of bits required to represent or transmit an average event from one distribution $P(X)$, compared to another distribution $Q(X)$,

$$H_{P,Q}(\mathcal{X}) = - \sum_{x_k \in \mathcal{X}} P(X = x_k) \cdot \log(Q(X = x_k)),$$

where $P(X = x_k)$ is the probability of the event x_k in $P(X)$, $Q(X = x_k)$ is the probability of event x_k in $Q(X)$.

- Let $\mathcal{X} = \{0, 1\}$ be a binary source with p_0 and p_1 being the probability of symbols $x_0 = 0$ and $x_1 = 1$, respectively, then we have

$$\begin{aligned} H_{P,Q}(\mathcal{X}) &= -p_0 \log q_0 - p_1 \log q_1 \\ &= -p_0 \log q_0 - (1 - p_0) \log (1 - q_0). \end{aligned}$$

Cross Entropy

There are **two properties** of cross-entropy:

- It is **non-negative**, *i.e.*, $H_{P,Q}(\mathcal{X}) \geq 0$
- Cross-entropy is **not smaller than** entropy, *i.e.*, $H_{P,Q}(\mathcal{X}) \geq H_P(\mathcal{X})$, and the equality holds only when $P = Q$.

Exercise: Prove the above two properties
(hint: Jensen inequality and $\log(\cdot)$ is concave)

Reference: <https://machinelearningmastery.com/cross-entropy-for-machine-learning/>

Relative entropy: Kullback-Leibler divergence

- The **relative entropy** between two discrete probability distributions P and Q is

$$D_{P,Q}(\mathcal{X}) = \sum_{x_k \in \mathcal{X}} P(X = x_k) \cdot \log \left(\frac{P(X = x_k)}{Q(X = x_k)} \right),$$

- The **relative entropy** between two continuous probability density functions $p_X(x)$ and $q_X(x)$ is defined as follows

$$D_{P,Q}(\mathcal{X}) = \int_{x \in \mathcal{X}} p_X(x) \log \left(\frac{p_X(x)}{q_X(x)} \right) dx$$

- It is also called **KL divergence**, which measures the distance between two distributions.

Relative entropy: Kullback-Leibler divergence

- There are **two properties** of KL divergence:
 - **Non-negativity:** $D_{P,Q}(\mathcal{X}) \geq 0$
 - **Asymmetry:** $D_{P,Q}(\mathcal{X}) \neq D_{Q,P}(\mathcal{X})$
- **Exercise 1: Prove the above two properties.**
(Hint: Jensen inequality and $\log(\cdot)$ is concave)
- The cross-entropy $H_{P,Q}(\mathcal{X})$ is the entropy of the distribution $H_P(\mathcal{X})$ plus the KL divergence $D_{P,Q}(\mathcal{X})$.

$$H_{P,Q}(\mathcal{X}) = H_P(\mathcal{X}) + D_{P,Q}(\mathcal{X})$$

- **Exercise 2: Prove the above equation.**

Reference:

<https://machinelearningmastery.com/cross-entropy-for-machine-learning/>

<https://stats.stackexchange.com/questions/335197/why-kl-divergence-is-non-negative>