DDA3020 Machine Learning Lecture 02 Probability Theory

JIA, Kui School of Data Science, CUHK-SZ

September 12/14, 2023

Outline

- A brief review of last week
- 2 Probability, event, random variables
- 3 Probability of discrete random variables
- Probability of continuous random variables
- **5** Some common discrete distributions
- 6 Some common continuous distributions
- Information theory

Definition and branches of machine learning



Figure: Machine learning is one of the most important branches of artificial intelligence.

Basic concepts, learning process



Figure: How supervised learning works.



Figure: How unsupervised learning works.

1 A brief review of last week

2 Probability, event, random variables

- 3 Probability of discrete random variables
- Probability of continuous random variables
- 5 Some common discrete distributions
- 5 Some common continuous distributions

7 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,

- $P(A) \ge 0$
- 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



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

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

• Still take the 2-times coin toss as 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: X is continuous

• Exercise 1: For the random experiment of throwing a pair of dice, please write down its sample space, and the event *at least one number 3 on two dice* as well as its probability.

• Exercise 2: For the above random experiment, we define the random variable as *the number summation of two dice*, please write down its state space.

- 1 A brief review of last week
- 2 Probability, event, random variables
- 3 Probability of discrete random variables
- Probability of continuous random variables
- 5 Some common discrete distributions
- 6 Some common continuous distributions
- 7 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

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

- Exercise 1: If we assume the coin is fair in the random experiment of 2-times coin toss, *i.e.*, $P(Head) = P(Tail) = \frac{1}{2}$ for toss, please compute the probability of different number of tails.
- Exercise 2: For the random experiment of throwing a pair of dice, please compute the probability of the state of 3 (*i.e.*, the number summation of two dice equals to 3).



Joint, marginal, conditional probability

• Probability of a union of two events: Given two events A and B, we define the probability of A or B as follows:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

$$= P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive.}$$
(1)

• Joint probabilities: The probability of the joint event A and B is defined as follows:

$$P(A, B) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A),$$
(2)

It is called the **product rule**.

• Marginal distribution: Given the above joint distribution, we can define the marginal distribution as follows:

$$P(A) = \sum_{b} P(A, B) = \sum_{b} P(A|B=b)P(B=b),$$
(3)

which sums over all possible states of B. It is called the sum rule.

Conditional probability and Bayes rule

• Conditional probability: Recalculating probability of event A after someone tells you that event B happened, as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

$$P(A \cap B) = P(A|B)P(B) \tag{5}$$

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

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)},\tag{6}$$

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y | X = x)}{\sum_{x' \in \mathcal{X}} P(X = x')P(Y = y | X = x')}$$
(7)

- 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 positive test, while x = 0 as the event of negative test. We denote y = 1 as the event of having breast cancer, while 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.*,

$$P(x=1|y=1) = 0.8.$$
 (8)

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

Application of Bayes rule: 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.*,

$$P(y=1) = 0.13. (9)$$

• 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):

$$P(x=1|y=0) = 0.1.$$
 (10)

• Combining all above probabilities using Bayes rule, we can compute

$$P(y=1|x=1) = \frac{P(x=1|y=1)P(y=1)}{P(x=1|y=1)P(y=1) + P(x=1|y=0)P(y=0)}$$
$$= \frac{0.8 \times 0.13}{0.8 \times 0.13 + 0.1 \times 0.87} = 0.5445.$$
(11)

It tells that if you test positive, you have 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.*,

$$X \perp Y \Longleftrightarrow P(X,Y) = P(X)P(Y). \tag{12}$$

- 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 the 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): $E(X) = \sum_{x \in \mathcal{X}} x P(X = x)$
- 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

$$Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2 = M_2 - M_1^2.$$
(13)

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

$$\operatorname{Std} = \sqrt{\operatorname{Var}(X)}.$$
 (14)

• 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 A brief review of last week
- 2 Probability, event, random variables
- **3** Probability of discrete random variables
- 4 Probability of continuous random variables
 - 5 Some common discrete distributions
 - 5 Some common continuous distributions
 - 7 Information theory

- 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

$$P(a < X < b) = \int_{a}^{b} p(x)dx \tag{15}$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \tag{16}$$

• 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 probablity 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 $p_{X,Y}(x,y) = p_X(x)p_Y(y)$



- Since the CDF F(·) is a monotonically increasing function, it has an inverse; let us denote this by F⁻¹(·).
- If F(x) is the CDF of X, then $F^{-1}(\alpha)$ is the value of x_{α} such that $P(X \leq x_{\alpha}) = \alpha$; this is called the a quantile of F. The value $F^{-1}(0.5)$ is the median of the distribution, with half of the probability mass on the left, and half on the right. The values $F^{-1}(0.25)$ and $F^{-1}(0.75)$ are the **lower** and **upper quartiles**.

Quantiles



• We can also use the inverse CDF to compute tail area probabilities.

- For example, if Φ is the CDF of the Gaussian distribution $\mathcal{N}(0, 1)$, then points to the left of $\Phi^{-1}(\alpha/2)$ contain $\alpha/2$ probability mass. By symmetry, points to the right of $\Phi^{-1}(1 \alpha/2)$ also contain $\alpha/2$ probability mass.
- Hence, the central interval (Φ⁻¹(α/2), Φ⁻¹(1 − α/2)) contains 1 − α of the mass. If we set α = 0.05, the central 95% interval is covered by the range

$$(\Phi^{-1}(0.025), \Phi^{-1}(0.975)) = (-1.96, 1.96).$$
 (17)

For a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the central 95% interval is $(\mu - 1.96\sigma, \mu + 1.96\sigma)$.

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

$$Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2 = M_2 - M_1^2.$$
(18)

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

$$\operatorname{Std} = \sqrt{\operatorname{Var}(X)}.$$
 (19)

- 1 A brief review of last week
- 2 Probability, event, random variables
- 3 Probability of discrete random variables
- Probability of continuous random variables
- **5** Some common discrete distributions
 - 5 Some common continuous distributions
 - 7 Information theory

Binary variables (discrete r.v.)

- We firstly 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 ,

$$p(x=1|\mu) = \mu, \tag{20}$$

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

Bern
$$(x|\mu) = \mu^x (1-\mu)^{1-x},$$
 (21)

which is called **Bernoulli** distribution.

• Its mean and variance are

$$\mathbb{E}[x] = \sum_{x} x \operatorname{Bern}(x|\mu) = \mu, \qquad (22)$$

$$\operatorname{var}[x] = \mathbb{E}[(x - \mu)^2] = \mu(1 - \mu)$$
 (23)

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 numbers of heads, then its distribution is formulated as follows:

$$\operatorname{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}, \qquad (24)$$

which is called **Binomial** distribution, where

$$\binom{N}{m} = \frac{N!}{(N-m)!m!}.$$
(25)

• Its mean and variance are

$$\mathbb{E}[m] = \sum_{m=0}^{N} m \operatorname{Bin}(m|N,\mu) = N\mu, \qquad (26)$$

$$\operatorname{var}[m] = \mathbb{E}[(m - N\mu)^2] = N\mu(1 - \mu).$$
 (27)

Binomial distribution



- 1 A brief review of last week
- 2 Probability, event, random variables
- 3 Probability of discrete random variables
- Probability of continuous random variables
- 5 Some common discrete distributions
- 6 Some common continuous distributions

7 Information theory

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

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$
(28)

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

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

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right), \quad (29)$$

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 A brief review of last week
- 2 Probability, event, random variables
- 3 Probability of discrete random variables
- Probability of continuous random variables
- 5 Some common discrete distributions
- 5 Some common continuous distributions
- **7** 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: https://www.youtube.com/ watch?v=z7bVw7lMtUg
- Shannon defined the measure that quantifies the uncertainty of an event with given probability a bit.
- For a discrete random variable (a source) with finite alphabet, as follows

$$\mathcal{X} = \{x_0, \cdots, x_k, \cdots 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 logarithm is base 2, information is given in bits.

• Note that $I(x_k) \ge 0$, *i.e.*, non-negative. The equality only holds when $p_k = 1$, which means there is no uncertainty, then no 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	pprox 1/76 million	pprox 26 bits
gamma ray burst mass extinction today	$< 2.7 \cdot 10^{-12}$	> 38 bits

Larger surprise/uncertainty, more information/bits.

Entropy

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

$$H_P(\mathcal{X}) = E[I(x_k)]$$

= $\sum_{x_k \in \mathcal{X}} p_k \cdot I(x_k)$
= $-\sum_{x_k \in \mathcal{X}} p_k \cdot \log(p_k).$

• 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

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

Entropy of binary source



• 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

$$H_{P,Q}(\mathcal{X}) = -p_0 \log q_0 - p_1 \log q_1$$

= $-p_0 \log q_0 - (1 - p_0) \log (1 - q_0).$

There are two properties of cross-entropy:

- It is a **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/

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

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

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

Relative entropy: Kullback-Leibler divergence

- Exercise 1: What is a KL divergence for a discrete r.v. ?
- There are two properties of KL divergence:
 - Non-negativity: $D_{P,Q}(\mathcal{X}) \geq 0$
 - Asymmetry: $D_{P,Q}(\mathcal{X}) \neq D_{Q,P}(\mathcal{X})$
- \bullet Exercise 2: Prove the above two properties (hint: Jensen inequal 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 additional KL divergence $D_{P,Q}(\mathcal{X})$.

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

• Exercise 3: Can you prove the equality of the equation?

Reference: https://machinelearningmastery.com/cross-entropy-for-machine-learning/ https://stats.stackexchange.com/questions/335197/why-kl-divergence-is-non-negative