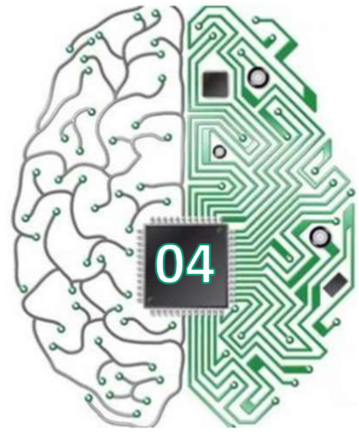
Open Elective Course [OE]
Course Code: CSO507
Winter 2023-24

Lecture#



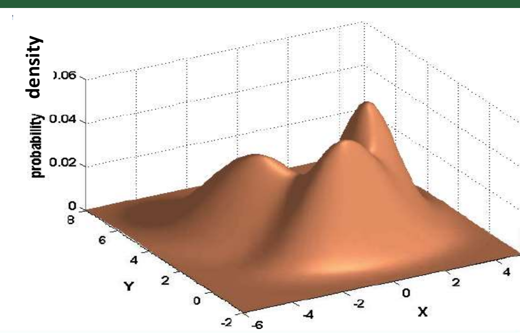
Deep Learning

Unit-1: Probability and Information Theory for Machine/Deep Learning

Course Instructor:
Dr. Monidipa Das
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology (Indian School of Mines) Dhanbad, Jharkhand 826004, India

Continuous Multivariate Distributions

- Same concepts of joint, marginal, and conditional probabilities apply for continuous random variables
- The probability distributions use integration of continuous random variables, instead of summation of discrete random variables



Definition (Joint PDF for continuous random variables)

Two random variables X and Y are *jointly continuous* if there exists a nonnegative function $f_{XY} : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$, such that, for any set $A \subseteq \mathbb{R} \times \mathbb{R}$, we have

$$\mathbb{P}\left((X,Y) \in A\right) = \iint_A f_{XY}(x,y)\, dxdy.$$

The function $f_{XY}(x,y)$ is called the *joint probability density function* (PDF) of X and Y.

Continuous Multivariate Distributions (2)

3

Marginal PDFs

Suppose $f_{X,Y}(x,y)$ is a joint PDF of X and Y, then the *marginal densities* of X and of Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

Definition (Conditional continuous random variable)

Suppose X and Y are jointly continuous, the *conditional probability density function (PDF)* of X given Y is given by

$$f_{X|Y=y}(x)) = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

Bayes' Theorem

4

- ***Bayes' theorem*** – allows to calculate conditional probabilities for one variable when conditional probabilities for another variable are known

$$P(X|\,Y) = \frac{P(Y|\,X)P(X)}{P(Y)}$$

- Also known as Bayes' rule
- Multiplication rule for the joint distribution is used: $P(X,Y) = P(Y|\,X)P(X)$
- By symmetry, we also have: $P(Y,X) = P(X|\,Y)P(Y)$
- The terms are referred to as:
  - $P(X)$, the prior probability, the initial degree of belief for X
  - $P(X|\,Y)$, the posterior probability, the degree of belief after incorporating the knowledge of Y
  - $P(Y|\,X)$, the likelihood of Y given X
  - P(Y), the evidence
  - Bayes' theorem: $\boldsymbol{posterior\ probability} = \frac{\boldsymbol{likelihood} \times \boldsymbol{prior\ probability}}{\boldsymbol{evidence}}$

Independence

5

- Two random variables X and Y are ***independent*** if the occurrence of Y does not reveal any information about the occurrence of X
  - E.g., two successive rolls of a die are independent
- Therefore, we can write: $P(X|\ Y) = P(X)$
  - The following notation is used: $X \perp Y$
  - Also note that for independent random variables: $P(X, Y) = P(X)P(Y)$
- In all other cases, the random variables are ***dependent***
- Two random variables X and Y are ***conditionally independent*** given another random variable Z if and only if $P(X, Y|Z) = P(X|Z)P(Y|Z)$
  - This is denoted as $X \perp Y|Z$

The joint density function factors for independent random variables

Jointly continuous random variables X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

(f indicates PDF here)

Expected Value

6

- The ***expected value*** or ***expectation*** of a function $f(x)$ with respect to a probability distribution $P(\mathrm{x})$ is the average (mean) when x is drawn from $P(\mathrm{x})$
- For a ***discrete random variable*** x, it is calculated as

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x)f(x)$$

- For a ***continuous random variable*** x, it is calculated as

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \int P(x)f(x)\,dx$$

Variance

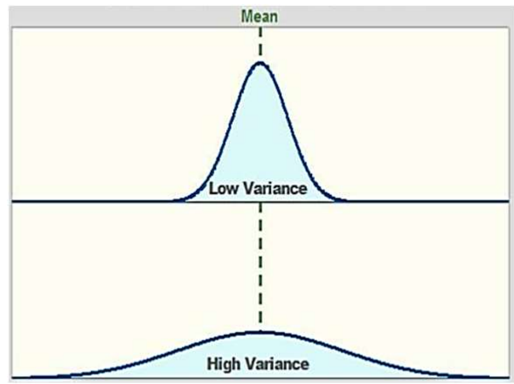
7

- ***Variance*** : how much the values of the function $f(x)$ deviate from the expected value as we sample values of x from $P(x)$

$$\mathrm{Var}\big(f(x)\big) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

- When the variance is low, the values of $f(x)$ cluster near the expected value

- The square root of the variance is the ***standard deviation***
  - Denoted $\sigma = \sqrt{\mathrm{Var}(f(x))}$



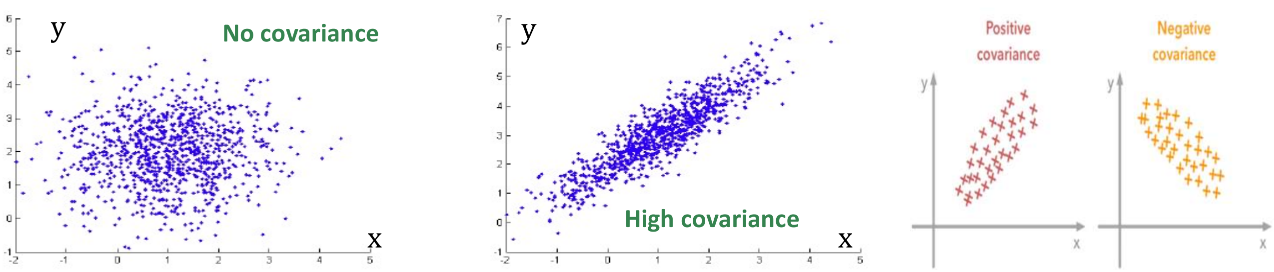
Covariance

8

- ***Covariance*** gives the measure of how much two random variables are linearly related to each other

$$\mathrm{Cov}\big(f(x), g(y)\big) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]$$

- The covariance measures the tendency for $f(x)$ and $g(y)$ to deviate from their means in same (or opposite) directions at same time



Correlation

9

- ***Correlation coefficient*** is the covariance normalized by the standard deviations

$$\mathrm{corr}\big(f(x), g(y)\big) = \frac{\mathrm{Cov}\big(f(x), g(y)\big)}{\sqrt{\mathrm{Var}(f(x))} \cdot \sqrt{\mathrm{Var}(g(y))}}$$

  - It is also called Pearson's correlation coefficient
  - The values are in the interval $[-1, 1]$
  - It only reflects linear dependence between variables, and it does not measure non-linear dependencies between the variables

Covariance Matrix

10

- ***Covariance matrix*** of a multivariate random variable $\mathbf{x}$ with states $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}$ is an $n \times n$ matrix, such that

$$\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}\big(\mathrm{x}_i, \mathrm{x}_j\big)$$

- I.e.,

$$\mathrm{Cov}(\mathbf{x}) = \begin{bmatrix} Cov(\mathrm{x}_1, \mathrm{x}_1) & Cov(\mathrm{x}_1, \mathrm{x}_2) & \cdots & Cov(\mathrm{x}_1, \mathrm{x}_n) \\ Cov(\mathrm{x}_2, \mathrm{x}_1) & \ddots & & Cov(\mathrm{x}_2, \mathrm{x}_n) \\ \vdots & & & \vdots \\ Cov(\mathrm{x}_n, \mathrm{x}_1) & Cov(\mathrm{x}_n, \mathrm{x}_2) & \cdots & Cov(\mathrm{x}_n, \mathrm{x}_n) \end{bmatrix}$$

- The diagonal elements of the covariance matrix are the variances of the elements of the vector

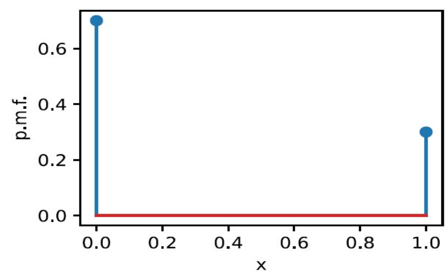
$$\mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \mathrm{Var}(\mathrm{x}_i)$$

- Also note that the covariance matrix is symmetric, since $Cov\big(\mathrm{x}_i, \mathrm{x}_j\big) = Cov\big(\mathrm{x}_j, \mathrm{x}_i\big)$

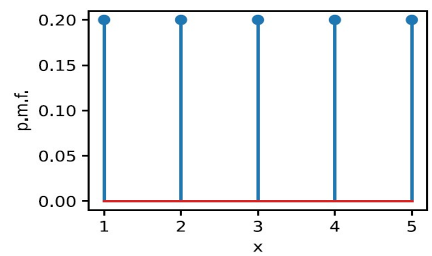
Probability Distributions

11

- ***Bernoulli distribution***
  - Binary random variable x with states $\{0, 1\}$
  - The random variable can encodes a coin flip which comes up 1 with probability p and 0 with probability $1 - p$
  - Notation: $\mathrm{x} \sim Bernoulli(p)$



- ***Uniform distribution***
  - The probability of each value $i \in \{1,2, \dots, n\}$ is $p_i = \frac{1}{n}$
  - Notation: $\mathrm{x} \sim U(n)$
  - Figure: $n = 5,\ p = 0.2$



Probability Distributions

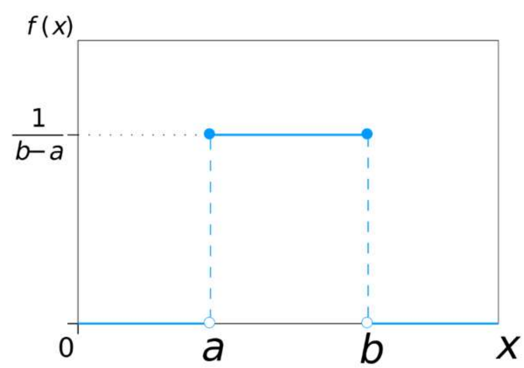
12

- ***Uniform distribution [continuous]***

The pdf of a uniform random variable with domain $[a, b]$, where $b > a$ are real numbers, is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & if\ a \le x \le b, \\ 0, & otherwise. \end{cases}$$

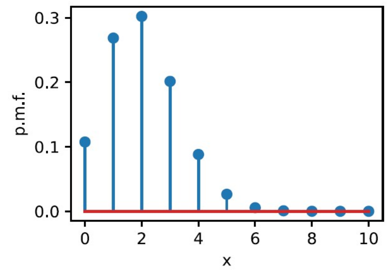
Notation: $\mathrm{x} \sim U(a, b)$



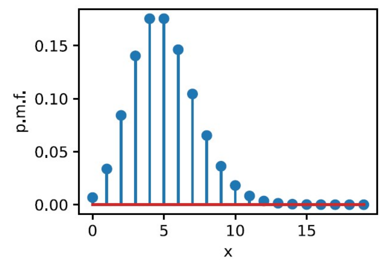
Probability Distributions

13

- ***Binomial distribution***
  - Performing a sequence of *n* independent experiments, each of which has probability *p* of succeeding, where $p \in \{0, 1\}$
  - The probability of getting *k* successes in *n* trials is $P(\mathrm{x} = k) = \binom{n}{k} p^k (1-p)^{n-k}$
  - Notation: $\mathrm{x} \sim Binomial(n, p)$



- ***Poisson distribution***
  - A number of events occurring independently in a fixed interval of time with a known rate λ
  - A discrete random variable x with states $k \in \{0, 1, 2, \ldots\}$ has probability $P(\mathrm{x} = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$
  - The rate λ is the average number of occurrences of the event
  - Notation: $\mathrm{x} \sim Poisson(\lambda)$

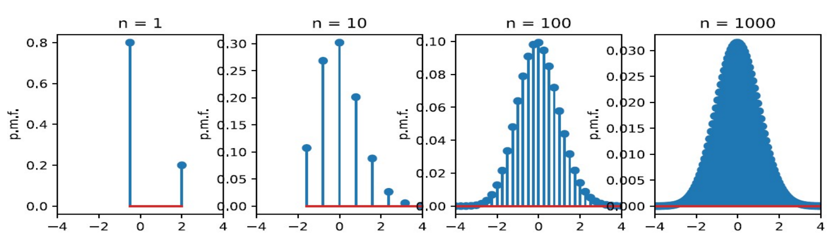


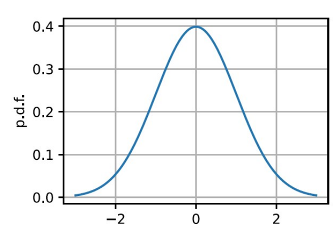
Probability Distributions

14

- ***Gaussian distribution***
  - The most well-studied distribution
    - Referred to as normal distribution or informally bell-shaped distribution
  - Defined with the mean μ and variance σ^2
  - Notation: $\mathrm{x} \sim \mathcal{N}(\mu, \sigma^2)$
  - For a random variable x with *n* independent measurements, the density is

$$P_{\mathrm{x}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$





Probability Distributions

15

- ***Multinoulli distribution***
  - It is an extension of the Bernoulli distribution, from binary class to multi-class
  - Multinoulli distribution is also called categorical distribution or generalized Bernoulli distribution
  - For example, in multi-class classification in machine learning, we have a set of data examples $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, and corresponding to the data example $\mathbf{x}_i$ is a k-class label $\mathbf{y}_i = \{y_{i1}, y_{i2}, \dots, y_{ik}\}$ representing one-hot encoding
    - Let's denote the probabilities for assigning the class labels to a data example by $\{p_1, p_2, \dots, p_k\}$
    - The multinoulli probability of the data example $\mathbf{x}_i$ is $P(\mathbf{x}_i) = {p_1}^{y_{i1}} \cdot {p_2}^{y_{i2}} \cdots {p_k}^{y_{ik}} = \prod_j {p_j}^{y_{ij}}$
    - Similarly, we can calculate the probability of all data examples as $\prod_i \prod_j {p_j}^{y_{ij}}$

16

Information Theory

Information Theory

17

- ***Information theory*** studies encoding, decoding, transmitting, and manipulating information

- Father of information theory: *Claude Elwood Shannon*

- As such, information theory provides fundamental language for discussing the information processing in computer systems
  - E.g., machine learning applications use the cross-entropy loss [to be discussed in the next lecture], derived from information theoretic considerations

Self-information

18

- Shannon defined the ***self-information*** of an event X as

$$I(X) = -\log\big(P(X)\big)$$

  - $I(X)$ is the self-information, and $P(X)$ is the probability of the event X

- The self-information outputs the bits of information received for the event X
  - For example, if we want to send the code "0010" over a channel
  - The event "0010" is a series of codes of length n (in this case, the length is n =4)
  - Each code is a bit (0 or 1), and occurs with probability of $\frac{1}{2}$; for this event $P = \frac{1}{2^n}$

$$I(\text{"0010"}) = -\log\big(P(\text{"0010"})\big) = -\log\left(\frac{1}{2^4}\right) = -\log_2(1) + \log_2(2^4) = 0 + 4 = 4 \text{ bits}$$

[Log base e => unit is nats   Log base 2 => unit is bits]

Prof. Monidipa Das, Department of CSE, IIT (ISM) Dhanbad

Entropy

19

- For a discrete random variable X that follows a probability distribution P with a probability mass function $P(X)$, the expected amount of self-information is ***entropy*** (or Shannon entropy):
$$H(X) = \mathbb{E}_{X\sim P}[I(X)] = -\mathbb{E}_{X\sim P}[\log P(X)]$$

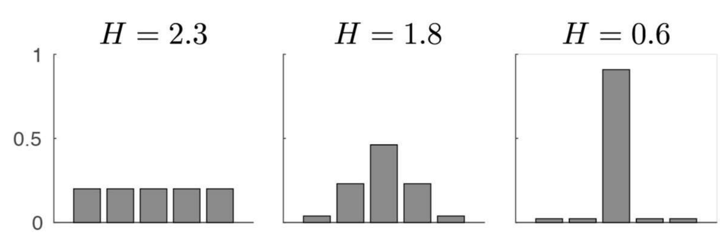
- Based on the expectation definition $\mathbb{E}_{X\sim P}[f(X)] = \sum_X P(X)f(X)$, we can rewrite the entropy as
$$H(X) = -\sum_X P(X)\log P(X)$$

- If X is a continuous random variable that follows a probability distribution P with a probability density function $P(X)$, the entropy is
$$H(X) = -\int_X P(X)\log P(X)\, dX$$
  - For continuous random variables, the entropy is also called differential entropy

Entropy

20

- Distributions that are closer to a uniform distribution have high entropy
- Because there is little surprise when we draw samples from a uniform distribution, since all samples have similar values



Kullback–Leibler Divergence

21

- ***Kullback-Leibler (KL) divergence*** (or relative entropy)
- provides a measure of how different two probability distribution are
- For two probability distributions $P(X)$ and $Q(X)$ over the same random variable X, the KL divergence is

$$D_{KL}(P||Q) = \mathbb{E}_{X\sim P}\left[\log\frac{P(X)}{Q(X)}\right] = \mathbb{E}_{X\sim P}[\log P(X) - \log Q(X)] = -\mathbb{E}_{X\sim P}\left[\log\frac{Q(X)}{P(X)}\right]$$

- For discrete random variables, this formula is equivalent to

$$D_{KL}(P||Q) = \sum_X P(X)\log\frac{P(X)}{Q(X)} = -\sum_X P(X)\log\frac{Q(X)}{P(X)}$$

- KL divergence is **non-negative**

Jensen-Shannon divergence

22

- KL divergence is not a true distance metric, because it is **not symmetric**

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

- An alternative divergence which is non-negative and symmetric is the ***Jensen-Shannon divergence***, defined as

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$$

  – In the above, M is the average of the two distributions, $M = \frac{1}{2}(P + Q)$

Exercise

23

- What is the gradient ($\nabla_x \sigma(\boldsymbol{x})$) of the sigmoid function $\sigma(\boldsymbol{x})$ as defined below?

$$\sigma(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T \boldsymbol{x})}$$

- What is the SVD decomposition of the following matrix:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 2 & 4 \\ 1 & 2 \end{bmatrix}$$

(write using the following form : $U\Sigma V^T$ where U and V are orthogonal, and Σ is diagonal)

Exercise

24

- Compute expectation and variance of the random variable X following:
  - $P(X = k) = \phi(1-\phi)^{k-1} \quad \text{for} \quad k = 1, 2, \ldots \quad \phi \in [0, 1]$
  - $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for} \quad k = 1, 2, \ldots \quad \lambda > 0$
- Compute expectation and variance of the random variable X following:
  - $p(X = x) = \frac{1}{b - a} \quad \forall x \in (a, b)$
  - $p(X = x) = \lambda e^{-\lambda x} \quad x \geq 0, \quad \lambda > 0$

Exercise

25

- What can be the maximum value of D_{KL}?
- What does $D_{KL} = 0$ indicate?

26

Questions?