BINUS UNIVERSITY

People
Innovation
Excellence

Course	: COMP6577 – Machine Learning
Effective Period	: December 2020

Probability and Stochastic Processes 1

Session 03 & 04

Learning Outcome

- LO2: Student be able to interpret the distribution of dataset using regression method

Outline

- Probability and random variables
- Probability
- Discrete random variables
- Continuous random variables
- Mean and variance
- Case Study

BINUS UNIVERSITY

Probability and Random Variables

- A random variable, *x*, is a variable whose variations are due to chance/randomness.
- A random variable can be considered as a function, which assigns a value to the outcome of an experiment.
- For example, in a coin tossing experiment, the corresponding random variable, *x*, can assume the values $x_1 = 0$ if the result of the experiment is “heads” and $x_2 = 1$ if the result is “tails.”
- A random variable is described in terms of a set of **probabilities** if its values are of a discrete nature, or in terms of a **probability density function (pdf)** if its values lie anywhere within an interval of the real axis (non-countably infinite set).

Probability

- Although the words “probability” and “probable” are quite common in our everyday vocabulary, the mathematical definition of probability is not a straightforward one, and there are a number of different definitions that have been proposed over the years.
- Needless to say, whatever definition is adopted, the end result is that the properties and rules, which are derived, remain the same.
- Two of the most commonly used definitions are: **Relative frequency definition** and **Axiomatic definition**

Relative Frequency Definition

- The probability, *P(A)*, of an event, *A*, is the limit

$$P(A) = \lim_{n \longrightarrow \infty} \frac{n_A}{n},$$

- Where *n* is the number of total trials and n_A the number of times event *A* occurred
- In practice in any physical experiment, the numbers n_A and *n* can be large, yet they are always finite. Thus, the limit can only be used as a hypothesis and not as something that can be attained experimentally.
- In practice, often, we use:

$$P(A) \approx \frac{n_A}{n}$$

for large values of *n*. However, this has to be used with caution, especially when the probability of an event is very small.

People
Innovation
Excellence

Axiomatic Definition

- This definition of probability is traced back to 1933 to the work of Andrey Kolmogorov, who found a close connection between probability theory and the mathematical theory of sets and functions of a real variable, in the context of measure theory.
- The probability, *P(A)*, of an event is a nonnegative number assigned to this event, or $P(A) \geq 0$.
- The probability of an event, *C*, which is certain to occur, equals to one $P(C) = 1$.
- If two events, A and B, are mutually exclusive (they cannot occur simultaneously), then the probability of occurrence of either A or B (denoted as A ∪ B) is given by $P(A \cup B) = P(A) + P(B)$.

Discrete Random Variables

- A discrete random variable, *x*, can take any value from a finite or countably infinite set *X* . The probability of the event, “x = *x* ∈ *X* ,” is denoted as *P(x = x)* or simply *P(x).*
- The function *P(·)* is known as the *probability mass function* (pmf). Being a probability, it has to satisfy the first axiom, so *P(x)* ≥ *0*.
- Assuming that no two values in *X* can occur simultaneously and that after any experiment a single value will always occur, the second and third axioms combined give:

$$\sum_{x \in \mathcal{X}} P(x) = 1$$

- The set *X* is also known as the sample or state space.

Joint and Conditional Probabilities

- The joint probability of two events, *A, B,* is the probability that both events occur simultaneously, and it is denoted as *P(A, B)*.
- Let us now consider two random variables, x, y, with sample spaces X = {x_1, … , x_{nx} } and Y = {y_1, … , y_{ny} }, respectively.
- Let us adopt the relative frequency definition and assume that we carry out *n* experiments and that each one of the values in *X* occurred n^x_1, … , n^x_{nx} times and each one of the va

$$P(x_i) \approx \frac{n^x_i}{n}, \ i = 1, 2, \ldots, n_x, \quad \text{and} \quad P(y_j) \approx \frac{n^y_j}{n}, \ j = 1, 2, \ldots, n_y.$$

Joint and Conditional Probabilities (2)

- Let us denote by n_{ij} the number of times the values x_i and y_j occurred simultaneously. Then, $P(x_i, y_j) \approx n_{ij} / n$. Simple reasoning dictates that the total number, n^x_i , that value x_i occurred, is equal to:

$$n_i^x = \sum_{j=1}^{n_y} n_{ij}$$

- Dividing both sides in the above by *n*, the following sum rule readily results.

$$\boxed{P(x) = \sum_{y \in \mathcal{Y}} P(x, y) : \quad \text{Sum Rule.}}$$

Joint and Conditional Probabilities (3)

- The conditional probability of an event, *A*, given another event, *B*, is denoted as *P(A|B)* and it is defined as

$$\boxed{P(A|B) := \frac{P(A,B)}{P(B)} : \quad \text{Conditional Probability,}}$$

provided *P(B) ≠ 0*

- Let n_{AB} be the number of times that both events occurred simultaneously, and n_B the times event *B* occurred, out of n experiments. Then, we have:

$$P(A|B) = \frac{n_{AB}}{n}\frac{n}{n_B} = \frac{n_{AB}}{n_B}$$

- In other words, the conditional probability of an event, *A*, given another one, *B*, is the relative frequency that *A* occurred, not with respect to the total number of experiments performed, but relative to the times event *B* occurred.

Joint and Conditional Probabilities (4)

- Viewed differently and adopting similar notation in terms of random variables, in conformity with Eq. (Sum rule), the definition of the conditional probability is also known as the product rule of probability, written as:

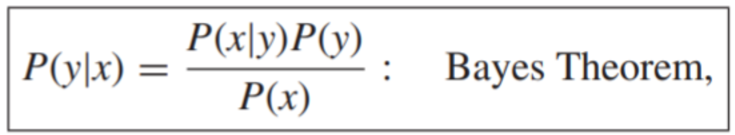
$$\boxed{P(x, y) = P(x|y)P(y) : \quad \text{Product Rule.}}$$

- To differentiate from the joint and conditional probabilities, probabilities, P(x) and P(y) are known as **marginal probabilities.**
- Statistical Independence: Two random variables are said to be statistically independent *if and only if* their joint probability is written as the product of the respective marginals,

$$P(x, y) = P(x)P(y)$$

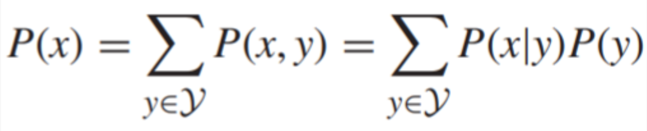
Bayes Theorem

- Bayes theorem is a direct consequence of the product rule and of the symmetry property of the joint probability, P(x, y) = P(y, x), and it is stated as:



$$\boxed{P(y|x) = \frac{P(x|y)P(y)}{P(x)} : \quad \text{Bayes Theorem,}}$$

- where the marginal, P(x), can be written as:



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

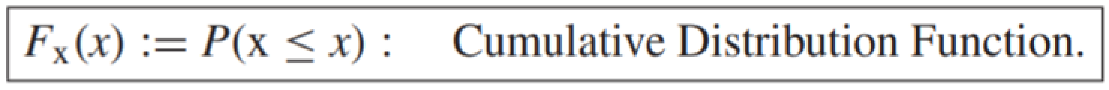
- and it can be considered as the normalizing constant of the numerator on the right-hand side in Eq. (Bayes Theorem), which guarantees that summing up *P(y|x)* with respect to all possible values of *y* ∈ *Y* results in one.
- Bayes theorem plays a central role in machine learning, and it will be the basis for developing Bayesian techniques for estimating the values of unknown parameters.

Continuous Random Variables

- Now turns to the extension of the notion of probability to random variables, which take values on the real axis, $\mathbb{R}$.
- The starting point is to compute the probability of a random variable, x, to lie in an interval, $x_1 < x \leq x_2$.
- Note that the two events, $x \leq x_1$ and $x_1 < x \leq x_2$, are mutually exclusive. Thus, we can write that:

$$P(x \leq x1) + P(x1 < x \leq x2) = P(x \leq x2)$$

- Define the cumulative distribution function (cdf) of x, as:



$$F_{\mathrm{x}}(x) := P(\mathrm{x} \leq x): \quad \text{Cumulative Distribution Function.}$$

- Then, we can write the equation above as:

$$P(x_1 < x \leq x_2) = F_x(x_2) - F_x(x_1).$$

People
Innovation
Excellence

Continuous Random Variables (2)

- Note that F_x is a monotonically increasing function. Furthermore, if it is continuous, the random variable *x* is said to be of a continuous type. Assuming that it is also differentiable, we can define the *pdf* (pdf) of x as:

$$\boxed{p_{\mathrm{x}}(x) := \frac{dF_{\mathrm{x}}(x)}{dx} : \quad \text{Probability Density Function,}}$$

- which then leads to:

$$P(x_1 < \mathrm{x} \le x_2) = \int_{x_1}^{x_2} p_{\mathrm{x}}(x)dx. \quad \text{and} \quad F_{\mathrm{x}}(x) = \int_{-\infty}^{x} p_{\mathrm{x}}(z)dz.$$

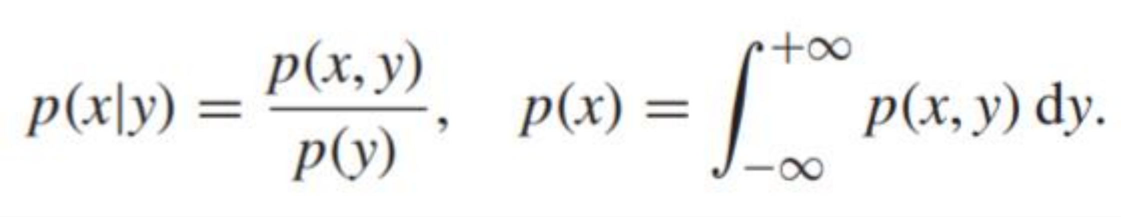
Using familiar logic from calculus arguments, the pdf can be interpreted as

$$\Delta P(x < \mathrm{x} \le x + \Delta x) \approx p_{\mathrm{x}}(x)\Delta x,$$

which justifies its name as a “density” function.

Continuous Random Variables (3)

- All previously stated rules for the probability are readily carried out for the case of pdfs, in the following way



$$p(x|y) = \frac{p(x,y)}{p(y)}, \quad p(x) = \int_{-\infty}^{+\infty} p(x,y)\,\mathrm{d}y.$$

- Note: the lower case “p” to denote a pdf and the capital “P” to denote a probability.

Mean and Variance

- Two of the most common and useful quantities associated with any random variable are the respective mean value and variance.
- The mean value (or sometimes called expected value) is denoted as:

$$\mathbb{E}[\mathrm{x}] := \int_{-\infty}^{+\infty} x p(x) \, \mathrm{d}x : \quad \text{Mean Value},$$

- where for discrete random variables the integration is replaced by summation $\left(\mathbb{E}[\mathrm{x}] = \sum_{x \in \mathcal{X}} x P(x)\right)$
- The variance is denoted as σ_x^2 and it is defined as:

$$\sigma_x^2 := \int_{-\infty}^{+\infty} (x - \mathbb{E}[\mathrm{x}])^2 p(x) \, \mathrm{d}x : \quad \text{Variance},$$

- where integration is replaced by summation for discrete variables. The variance is a measure of the spread of the values of the random variable around its mean value.

People
Innovation
Excellence

Mean and Variance (2)

- The definition of the mean value is generalized for any function, f(x), i.e.,

$$\mathbb{E}[f(\mathrm{x})] := \int_{-\infty}^{+\infty} f(x)p(x)dx.$$

- It is readily shown that the mean value with respect to two random variables, y, x, can be written as the product

$$\mathbb{E}_{\mathrm{x},\mathrm{y}}[f(\mathrm{x},\mathrm{y})] = \mathbb{E}_{\mathrm{x}}\left[\mathbb{E}_{\mathrm{y}|\mathrm{x}}[f(\mathrm{x},\mathrm{y})]\right]$$

- This is a direct consequence of the definition of the mean value and the product rule of probability.

Mean and Variance (3)

- Given two random variables x, y, their covariance is defined as

$$\mathrm{cov}(\mathrm{x},\mathrm{y}) := \mathbb{E}[(\mathrm{x}-\mathbb{E}[\mathrm{x}])(\mathrm{y}-\mathbb{E}[\mathrm{y}])]$$

- and their (

$$r_{xy} := \mathbb{E}[\mathrm{xy}] = \mathrm{cov}(\mathrm{x},\mathrm{y}) + \mathbb{E}[\mathrm{x}]\,\mathbb{E}[\mathrm{y}]$$

- A random vector is a collection of random variables, x = $[x_1, \ldots, x_l]^T$, and *p(x)* is the joint pdf (probability for discrete variables), *p(x) = p(x_1, ... , x_l).*
- The ned as

$$\boxed{\mathrm{Cov}(\mathbf{x}) := \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T\right] : \quad \text{Covariance Matrix,}}$$

$$\mathrm{Cov}(\mathbf{x}) = \begin{bmatrix} \mathrm{cov}(\mathrm{x}_1,\mathrm{x}_1) & \ldots & \mathrm{cov}(\mathrm{x}_1,\mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathrm{x}_l,\mathrm{x}_1) & \ldots & \mathrm{cov}(\mathrm{x}_l,\mathrm{x}_l) \end{bmatrix}$$

Mean and Variance (4)

- Similarly, the correlation matrix of a random vector, *x,* is defined as

$$\boxed{R_x := \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right] : \quad \text{Correlation Matrix,}}$$

$$R_x = \begin{bmatrix} \mathbb{E}[\mathrm{x}_1, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_1, \mathrm{x}_l] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\mathrm{x}_l, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_l, \mathrm{x}_l] \end{bmatrix}$$
$$= \mathrm{Cov}(\mathbf{x}) + \mathbb{E}[\mathbf{x}]\,\mathbb{E}[\mathbf{x}^T].$$

Case Study

Given data of Singapore Airbnb which can be downloaded in this link

https://www.kaggle.com/jojoker/singapore-airbnb

1. From the downloaded data we can identify discrete and continuous random variables. Discuss about those variables.

2. We can also calculate mean and variance each for discrete and continuous random variables.

End of Session 03 & 04

References

- Sergios Theodoridis. (2015). *Machine Learning: a Bayesian and Optimization Perspective*. Jonathan Simpson. ISBN: 978-0-12-801522-3. Chapter 2.
- https://www.kaggle.com/jojoker/singapore-airbnb