

Fundamental results in probability theory

Zizhou Fang^{1, †}, Kaixi Tan^{2, †}, Ziyi Wang^{3, †}

¹ University of Toronto, Computer science and Mathematics Department, Ontario, Canada, M1C

² Physics Department, University of California Santa Barbara, CA, USA 93106

³ Mathematics Department, University of Richmond, Virginia, US, 23173

* Corresponding Author Email: guanghua.ren@gecacademy.cn

[†] These authors contributed equally.

Abstract. Probability theory is an area of mathematics that deals with the concept of likelihood. Probability theory is the mathematical foundation of statistical reasoning, and understanding how unpredictability impacts data is crucial for data scientists. Gaussian (normal) distribution is the most widely used distribution. It has two parameters which are mean and variance and easy to interpret. Also, the central limit theorem tells us that sums of independent random variables make the least number of assumptions. In addition, Poisson, Laplace, Beta, Pareto, Dirichlet, Binomial and Gamma Distributions are useful in different areas. The multivariate Gaussian is the most widely used joint probability density function. Covariance and correlation are used to measure the degree between two random variable's X and Y . Chebyshev Inequality defines a topological space, which includes a sequence of elements, and let the sequence be called y_n . Strong Law of Large Numbers Theorem use in large number of random variable in pairwise independent identically distributed and Renewal Theory is an example in Strong Law of Large Numbers Theorem.

Keywords: Strong Law of Large number theorem; random variable; Chebyshev inequality.

1. Introduction

Probability Theory is a branch of mathematics that deals with the statistical rules that govern random events. It is widely employed in different areas in our everyday lives [1-20] as an important subject in the field of mathematics in today's world. In general, it stems from the difficulty of gambling, and in the 17th century, mathematicians used probability to address the problem of "assigning the wager." Since then, the Probability Theory has progressed significantly, and it is now widely used in practical applications where random objects abound. The Probability Theory has a significant impact on a variety of aspects of our life. It is, for example, used to forecast admission scores in a variety of educational recruiting situations.

2. Main Works

Let A denote an event. $P(A)$: The probability that the event A will happen, or is true in more mathematical terms. $0 \leq P(A) \leq 1$. $P(A) = 0$ indicates that event A will never occur, and $P(A) = 1$ indicates that event A will definitely occur. $P(\bar{A})$: The probability of "not event A ". $0 \leq P(\bar{A}) \leq 1$. $P(\bar{A}) = 1 - P(A)$. A Discrete Random Variable X : any value from a set χ , which can be a finite set or sometimes countably infinite. The probability of $X = x$ is denoted $P(X = x)$ or $P(x)$. $P(x)$ is now a Probability Mass Function and its properties include:

$$0 \leq P(x) \leq 1 \quad (1)$$

$$\sum_{x \in \chi} P(x) = 1 \quad (2)$$

χ denotes a state space, and $I()$ is a binary indicator function.

2.1. Fundamental Rules of probability

Let A, B be two events.

The probability of A or B (the union of both events, the probability of either one event occurring) is defined as:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) \quad (3)$$

$$P(A \vee B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive.} \quad (4)$$

Joint probability:

$$\text{For a joint event } A, B, P(A, B) = P(A \wedge B) = P(A | B) P(B) \quad (5)$$

(A joint event A, B is true if both A and B are true)

The calculation of Joint Probability is occasionally addressed as the Product Rule.

Marginal Distribution is defined with a given Joint Distribution $P(A, B)$:

$$P(A) = \sum_b P(A, B) = \sum_b P(A | B = b) P(B = b) \quad (6)$$

The calculation of Marginal Distribution is occasionally addressed as the Sum Rule or the Rule of Total Possibility.

Chain rule of possibility:

$$P(x_{1:D}) = P(x_1)P(x_2|x_1)P(x_3|x_2, x_1)\dots P(x_D|x_{1:D-1}) \quad (7)$$

where $1:D$ denotes the set $\{1, 2, \dots, D\}$.

Conditional Probability:

Given that an event B is true, for an event A , we define its conditional probability to be:

$$P(A|B) = \frac{P(A, B)}{P(B)} \text{ if } P(B) > 0. \quad (8)$$

2.2. Bayes Rule

Bayes Rule or Bayes Theorem is a combination of product and sum rules, and is defined as below:

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y | X = x)}{\sum_{x'} P(X = x')P(Y = y | X = x')} \quad (9)$$

Generate Classifiers:

$$P(y = c | x, \theta) = \frac{P(y=c|\theta)P(x | y=c, \theta)}{\sum_{c'} P(y=c'|\theta)P(x | y=c', \theta)} \quad (10)$$

This is called a generative classifier with class conditional density $P(x|y = c)$ and class prior $P(y = c)$.

Two events X, Y are defined to be unconditionally independent or marginally independent iff $P(X, Y) = P(X)P(Y)$, denoted $X \perp Y$.

$$\text{In other words, } P(X, Y) = P(X)P(Y) \Leftrightarrow X \perp Y. \quad (11)$$

Unconditional independence is infrequent as most variables will affect other variables, but the impact is through other variables. Therefore, conditional independence between two events X and Y is true when:

$$P(X, Y | Z) = P(X | Z)P(Y | Z) \Leftrightarrow X \perp Y | Z \quad (12)$$

Let X be an uncertain continuous quantity. For X to be between any intervals a, b , $P(a \leq X \leq b)$ is computed by the following:

Define the events $A = (X \leq a), B = (X \leq b), W = (a < X \leq b)$.

Thus, $B = A \vee W$.

Note that A and W are defined to be mutually exclusive, and thus the sum rule gives:

$$P(B) = P(A) + P(W) \quad (13)$$

And hence

$$P(W) = P(B) - P(A) \quad (14)$$

The empirical distribution is also called the empirical measure $P_{\text{emp}}(A) \triangleq \frac{1}{N} \sum_{i=1}^N \delta_{x_i}(A)$, where δ_{x_i} is Dirac measure: $\delta_{x_i}(A) = 0$ if $x \notin A$, $\delta_{x_i}(A) = 1$ if $x \in A$. Generally, we can link “weights” with each sample: $p(x) = \sum_{i=1}^N w_i \delta_{x_i}(A)$ where $0 \leq w_i \leq 1$ and $\sum_{i=1}^N w_i = 1$. The image should look in the form of a bar graph, points X_i being its peaks or spikes, the heights of which is decided by w_i . For any point not included in the data set, the image would give a probability of 0.

2.3. Gaussian (normal) distribution

This distribution has perhaps the widest appearance and usage. Its pdf is: $N(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, $\mu = E[X]$ is the mean, $\sigma^2 = \text{var}[X]$ is the variance, $\frac{1}{\sqrt{2\pi\sigma^2}}$ is the normalization constant and the density integrates to 1. We write $X \sim N(\mu, \sigma^2)$ to denote that $p(X=x) = N(x|\mu, \sigma^2)$ and $X \sim N(0, 1)$ is a standard normal distribution. The plot of this pdf can also be addressed as the bell curve.

The precision of a Gaussian means the inverse variance: $\lambda = 1/\sigma^2$. The implication behind a high precision is a relatively narrower distribution than its lower precision counterparts. This is also known as a Low Variance, and is centered on μ . Since this is a pdf, $p(x) > 1$ is possible. For example, evaluating the density at its center, $x = \mu$, $N(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, so if $\sigma < 1/\sqrt{2\pi}$ and $p(x) > 1$. CDF: $\Phi(x; \mu, \sigma^2) \triangleq \int_{-\infty}^x N(z|\mu, \sigma^2) dz$. ERF(error function): $\Phi(x; \mu, \sigma) = \frac{1}{2} [1 + \text{erf}(z/\sqrt{2\pi})]$, where $z = (x - \mu)/\sigma$ and $\text{erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. The most broadly applied distribution in statistics is often considered to be the Gaussian distribution. First, the two parameters of this distribution, namely mean and variance, would pose a relatively low difficulty in terms of interpretation. Second, independent random variables would sum to approximately formulate a Gaussian distribution, as given in the central limit theorem. Therefore, the Gaussian distribution is an excellent model to assist the study of residual errors. Third, compared to other distributions, the Gaussian distribution has the most reduced number of speculations, and with mean and variance always specified, this distribution fits many scenarios naturally. Finally, its simple mathematical form implies methods of both simple implementation and high efficiency. Centring at μ , the Gaussian becomes a spike with both height and thinness reaching to infinity. $\lim_{\sigma^2 \rightarrow 0} N(x|\mu, \sigma^2) = \delta(x - \mu)$: here, δ represents a Dirac delta function, defined as: $\delta(x) = \infty$ if $x = 0$; $\delta(x) = 0$ if $x \neq 0$ such that: $\int_{-\infty}^{\infty} \delta(x) dx = 1$. The sifting property is the following: $\int_{-\infty}^{\infty} f(x) \delta(x - \mu) dx = f(\mu)$. Gaussian distribution is sensitive to outliers, a more stable distribution is the Student distribution, write pdf as: $T(x|\mu, \sigma^2, \nu) \propto \frac{1}{\sqrt{\nu\sigma^2}} \left(1 + \frac{(x - \mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$, where mean = μ = mode, var = $\nu\sigma^2/(\nu - 2)$. For the variance to be defined, ν must be greater than 2. Similarly, for the mean to be defined, ν must be greater than 1.

2.4. The Laplace distribution

The characteristic of this distribution is the existence of long tails. Another common name to address the Laplace distribution is the double sided exponential distribution, its PDF: $\text{Lap}(x|\mu, b) \triangleq \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$, with mean = μ .

2.4.1 The gamma distribution

The flexibility of Gamma distribution is retained to rv's of real values, $x > 0$. Its parameters are the shape $a > 0$ and the rate $b > 0$: $\text{Ga}(T|\text{shape} = a, \text{rate} = b) \triangleq \frac{b^a}{\Gamma(a)} T^{a-1} e^{-Tb}$, where $\Gamma(a) \triangleq \int_0^{\infty} u^{a-1} e^{-u} du$, with properties: mean = $\frac{a}{b}$, mode = $\frac{a-1}{b}$, var = $\frac{a}{b^2}$.

Some special cases of the Gamma:

Exponential distribution: $\text{Expon}(x|\lambda) \triangleq \text{Ga}(x|1, \lambda)$, where λ is the rate parameter.

Erlang distribution: $\text{Erlang}(x|\lambda) = \text{Ga}(x|2, \lambda)$, where λ is the rate parameter, a is an integer common fix to $a = 2$.

Chi-squared distribution: $\chi^2(x|\nu) \triangleq \text{Ga}(x|\frac{\nu}{2}, \frac{1}{2})$. More precisely, if $Z_i \sim N(0, 1)$, and $S = \sum_{i=1}^{\nu} Z_i^2$, then $S \sim \chi^2$.

inverse gamma: $\text{IG}(x|\text{shape} = a, \text{scale} = b) \triangleq \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{-b/x}$, with properties:
mean = $\frac{b}{a-1}$, mode = $\frac{b}{a+1}$, var = $\frac{b^2}{(a-1)^2 (a-2)}$.

2.4.2 The beta distribution

The beta distribution has support over the interval $[0, 1]$:

$$\text{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}, B(a, b) \triangleq \Gamma(a)\Gamma(b)/\Gamma(a+b)$$

Case 1: $a = b = 1$. This case results in the uniform distribution.

Case 2: $a < 1, b < 1$. Then the distribution is bimodal, with 0 and 1 being the position of its spikes.

Casw 3: $a > 1, b > 1$. Then the distribution becomes unimodal with qualities as follows: mean = $\frac{a}{a+b}$, mode = $\frac{a-1}{a+b-2}$, var = $\frac{ab}{(a+b)^2 (a+b+1)}$.

Pareto distribution

This is often applied to represent the distribution of quantities that unfold the so-called long tails or heavy tails. Its PDF: $\text{Pareto}(x|k, m) = kmx^{-(k+1)} \mathbb{I}(x \geq m)$. Its plot composes a straight line, $\log p(x) = a \log x + c$ for some constants a and c . For this distribution, mean = $\frac{km}{k-1}$ if $k > 1$, mode = m ,
var = $\frac{m^2 k}{(k-1)^2 (k-2)}$ (when $k > 2$).

2.5. Joint probability distributions

This distribution appears in the form $p(x_1, \dots, x_D)$ with respect to a set of variables, namely $x_1, \dots, x_D$, where $D > 1$. Among the variables, it also provides assistance in finding out the (stochastic) relationships. Using this distribution, the joint distribution of a discrete distribution can be expressed in the form of a large array with several dimensions, each dimension including one variable.

2.5.1 Covariance and correlation

Measures the degree between two rv's X and Y . Defined as:

$$\text{cov}[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

If x is a d -dimensional random vector: $\text{cov}[x] = E[(x - E[x])(x - E[x])^T]$

correlation coefficient is more convenient to work with a normalized measure, with a finite upper bound. $\text{corr}[X, Y] = \frac{\text{cov}[X, Y]}{\sqrt{\text{var}[X] \text{var}[Y]}}$, $-1 \leq \text{corr}[X, Y] \leq 1$. If X and Y are independent, $\text{cov}[X, Y] = 0$ = $\text{corr}[X, Y]$ and the converse is not true. For $\text{corr}[X, Y] = 1$ if and only if $Y = aX + b$ for some parameters a and b are linear relationship

The multivariate Gaussian, designed for continuous variables, has probably the widest usage among all functions of joint probability density. It is sometimes referred to as the multivariate normal function, or MVN for short. The pdf of the MVN in D dimensions:

$N(x|\mu, \Sigma) \triangleq \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp[-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)]$ where $\mu = E[x] \in \mathbb{R}^D$, $\Sigma = \text{cov}[x]$ is the $D \times D$ covariance matrix. Precision matrix or concentration matrix $\Lambda = \Sigma^{-1}$

2.6. Linear transformations

Let the function $f(x)$ be linear: $y = f(x) = Ax + b$

If we input this equation into mean and covariance context, then we can get: $E[y] = E[Ax + b] = A\mu + b$

where the linearity of expectation satisfies $\mu = E[x]$. Further, let the values of $f(x)$ be scalars, then for the covariance, we have $cov[y] = cov[Ax + b]$ where $\Sigma = cov[x]$ $var[y] = var[a^T x + b] = a^T \Sigma a$

2.7. Chebyshev Inequation

Define a topological space, which included a sequence of elements, and let the sequence be called y_n .

For $y_n \rightarrow y$, there must be a subsequence $y_{n(m_k)}$ converging to y , which is a subsequence of any sequence $y_{n(m)}$ that is a subsequence of y_n . The theorem is proven by a proof by contradiction. Suppose, for sake of contradiction, $y_n \not\rightarrow y$. Then there is an open set G such that $y \in G$, and a subsequence $y_{n(m)} \notin G$ for all m . But obviously there will be no subsequence of $y_{n(m)}$ will converge to y . Let f be a continuous function and $X_n \rightarrow X$ in probability. Then, $f(X_n) = f(X)$ in probability. Further, let f be bounded. Then $Ef(X_n) = Ef(X)$.

Let $X_1, X_2, \dots$ be i.i.d. with $E X_i = \mu$ and $E X_i^4 < \infty$. If $S_n = X_1 + X_2 + \dots + X_n$, then $S_n/n \rightarrow \infty$ a.s.

Set random variable X has its mathematical expectation and variance both existing, then for any constant $\varepsilon > 0$, there is: $P(|X - EX| \geq \varepsilon) \leq \frac{(X)}{\varepsilon^2}$.

The Second Borel-Cantelli Lemma:

Define A_n as some events that are independent. The expression $\sum P(A_n) = \infty$ indicates $P(A_n \text{ i.o.}) = 1$.

A classical application of this Lemma is as follows: Let $X_1, X_2, \dots$ be i.i.d, with $E|X_i| = \infty$, then $P(|X_n| \geq ni \text{ i.o.}) = 1$. Therefore, $S_n = X_1 + \dots + X_n$ $\square P(S_n/n \text{ exists } \in (-\infty, \infty)) = 0$.

Let $A_1, A_2, \dots$, be pairwise independent and $\sum_{n=1}^{\infty} P(A_n) = \infty$.

Then $\sum_{m=1}^n 1_{A_m} / \sum_{m=1}^n P(A_m) \rightarrow 1$ a.s. With the fact that $\sum_{m=1}^n 1/m \sim \log(n)$ and theorem 2.3.8., we have:

Let $R_n = \sum_{m=1}^n 1_{A_m}$ be the number of records at time n . Then as $n \rightarrow \infty, R_n / \log(n) \rightarrow 1$ a.s., which is independent of the distribution F . Strong Law of Large Numbers Let $X_1, X_2, \dots$ be random variables. They must be pairwise independent variables with an identical distribution, and further, $E|X_i| < \infty$. Let $EX_i = \mu$ and $S_n = X_1 + \dots + X_n$. Then $S_n/n \rightarrow \mu$ a.s. as $n \rightarrow \infty$.

Let $Y_k = X_k 1_{|X_k| \leq k}$ and $T_n = Y_1 + \dots + Y_n$. Then $T_n/n \rightarrow \mu$ a.s. $\sum_{k=1}^{\infty} (Y_k) / k^2 \leq 4E|X_1| < \infty$. If $y \geq 0$ then $2y \sum_{k>y} k^{-2} \leq 4$

Let $X_1, X_2, \dots$ be i.i.d. with $E X_i^+ = \infty$ and $E X_i^- < \infty$. If $S_n = X_1 + X_2 + \dots + X_n$, then $S_n / n \rightarrow \infty$ a.s.

Let $X_1, X_2, \dots$ be i.i.d. with $0 < X_i < \infty$. Define T_n as the total time elapsed to reach the n^{th} occurrence of some event, and let $T_n = X_1 + X_2 + \dots + X_n$, where X_i represents the time between the $(i-1)^{th}$ event and the i^{th} event. A good example will be a janitor who replaces a lightbulb once it burns out, and X_i therefore denotes the lifetime of the i^{th} bulb. $N_t = \sup\{n : T_n \leq t\}$ denotes the amount of bulbs that burned out at time t . $EX_1 = \mu \leq \infty$, then $N_t/t \rightarrow 1/\mu$ a.s. when t approaches ∞ . If we take X to be the successful times (Event A) of n test, then X can be taken from $0, 1, \dots$ to n , if we take p to be the probability of Event A to happen in every trial, then we have:

$P(A) = p; P(\underline{A}) = 1 - p$. Since the result of this n test can be taken as: $\omega = (\omega_1, \omega_2, \dots, \omega_n)$.

In the set, ω_i can either be A or $\underline{A}$, and there are $2^n \omega$, they form the sample space Ω

Now when we want to know the probability of event $\{X = k\}: P(\omega) = p^k(1-p)^{n-k}$. Meanwhile, in event $\{X = k\}$, we have $P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$, $k = 0, 1, \dots, n$. This distribution we call it binomial distribution, note as $X \sim b(n, p)$ Suppose X to be a random variable, and k to be a positive integer, if the following mathematical expectations exist: $\mu_k = E(x^k)$.

3. Conclusion

The probability measure is frequently assigned a value between 0 and 1, and the sample space is a collection of likely events, and probability is described in terms of a probability space. The results of an event are sometimes referred to as the event's outcomes. In general, probability theory tries to use mathematical abstractions to represent uncertain, or non-deterministic, processes. The mathematical study of uncertainty is known as probability theory. Because probabilistic data assumptions are frequently employed in the building of learning algorithms, it's critical in machine learning. We'll look at machine learning from a probabilistic perspective in the future.

References

- [1] Rick Durrett. Probability: theory and examples. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. 2010.
- [2] Peter Morters and Yuval Peres. "Brownian motion. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. 2010.
- [3] R.B. Ash and C.A. Doléans-Dade. Probability and Measure Theory, 2nd Ed. 2000.
- [4] Chow, Y. S. and H. Teicher Probability. Addison-Wesley, Reading, Mass. 1997.
- [5] Chung, K. L. Probability Theory: Independence, Interchangeability, Martingales, 3rd Ed. Springer Verlag, New York. 2001.
- [6] Dudley, R. M. A Course in Probability Theory, 3rd Ed. Academic Press, San Diego. 1989.
- [7] Durrett, Richard. Real Analysis and Probability. Wadsworth, Pacific Grove, California. 1996.
- [8] Holden-Day, San Francisco. Stoyanov, J. M. Probability: Theory and Examples 2nd Ed. Duxbury, Belmont, California. Neveu, J. 1965.
- [9] Taylor, J. Probability for Analysts. Chapman-Hall, New York. 1996
- [10] Murphy, Kevin P. Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press, 2012.
- [11] Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
- [12] Chinese Institute of Certified Public Accountants. "Corporate Strategy and Risk Management".
- [13] Harrison, Peter. Machine Learning in Action. Manning Publications Co, Shelter Island. Download from Wow! Ebook <www.wowebook.com>
- [14] J.Ross Quilan. C4.5 programs for machine learning. Morgan Kaufmann Publishers, San Mateo, California.
- [15] Sullivan, William. Machine learning for beginners: Guide Algorithms. Healthy Pragmatic Solutions Inc. Copyright 2017.
- [16] Larose, Chantal, Daniel (2014). Discovering Knowledge in Data. Hoboken, NJ: John Wiley & Sons
- [17] Utgoff, P. E. (1989). Incremental induction of decision trees. Machine learning, 4(2), 161–186.
- [18] R.Quinlan, "Learning efficient classification procedures", Machine Learning: an artificial intelligence approach, Michalski, Carbonell & Mitchell (eds.), Morgan Kaufmann, 1983, p. 463–482
- [19] Deng, H.; Runger, G.; Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).
- [20] Plapinger, Thomas "What is a Decision Tree?". Towards Data Science. Retrieved 5. December 2021.