Tecnológico de Monterrey

Bayes Theorem

CS5056 Data Analytics

Francisco J. Cantú, Héctor Ceballos
Tecnológico de Monterrey
April 7, 2021

Februrary-June, 2021

Probability Theory

- Probability theory is the study of uncertainty

- Concepts from probability theory are used for deriving machine learning algorithms.

- The mathematical theory of probability delves into a branch of analysis known as Measure Theory

Sample Space

- Sample space Ω: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.

- Set of events (or event space) F: A set whose elements $A \in F$ (called events) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)[1]

Probability Axioms

Probability measure: A function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties,

- $P(A) \geq 0$, for all $A \in \mathcal{F}$
- $P(\Omega) = 1$
- If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

Axioms of Probability Functions

- If $A \subseteq B \implies P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$.
- (Union Bound) $P(A \cup B) \leq P(A) + P(B)$.
- $P(\Omega \setminus A) = 1 - P(A)$.
- (Law of Total Probability) If $A_1, \ldots, A_k$ are a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_k) = 1$.

Conditional Probability

Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as,

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

- P (A|B) is the probability measure of the event A after observing the occurrence of event B
- Two events are called independent if and only if $P(A \cap B) = P(A)P(B)$ (or equivalently, $P(A|B) = P(A)$)
- Independence is equivalent to saying that observing B does not have any effect on the probability of A.

Random Variable

- Thus, A random variable X is a function $X : \Omega \dashrightarrow R$.
- Typically, we will denote random variables using upper case letters $X(\omega)$ or more simply X (where the dependence on the random outcome ω is implied)
- We will denote the value that a random variable may take on using lower case letters x.

Discrete Random Variable

- Suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω
- Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values, so it is known as a **discrete random variable**
- The probability of the set associated with a random variable X taking on some specific value k is:

$$P(X = k) := P(\{\omega : X(\omega) = k\}).$$

Continuous Random Variable

- Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay
- In this case, $X(\omega)$ takes on an infinite number of possible values, so it is called a **continuous random variable**.
- We denote the probability that X takes on a value between two real constants a and b (where $a < b$) as

$$P(a \le X \le b) := P(\{\omega : a \le X(\omega) \le b\}).$$

Cumulative Distribution Function

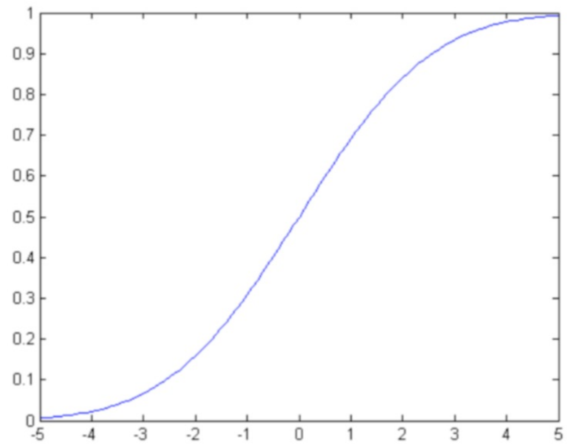
- In order to specify the probability measures used when dealing with random variables, it is often convenient to specify alternative functions from which the probability measure governing an experiment immediately follows.
- These include Cumulative Distribution Functions (CDF), Probability Density Functions (PDF), and Probability Mass Functions (PMF).
- **A Cumulative Distribution Function (CDF)** is a function $F_X : R \to [0, 1]$ which specifies a probability measure as:

$$F_X(x) = P(X \leq x)$$

- By using this function one can calculate the probability of any event in F

Cumulative Distribution Function

Properties



- $0 \leq F_X(x) \leq 1.$
- $\lim_{x \to -\infty} F_X(x) = 0.$
- $\lim_{x \to \infty} F_X(x) = 1.$
- $x \leq y \Longrightarrow F_X(x) \leq F_X(y).$

Probability Mass Function

- When a random variable X takes on a finite set of possible values (i.e., X is a discrete random variable), a simpler way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume
- In particular, a *probability mass function (PMF)* is a function $pX : \Omega \to R$ such that :

$$p_X(x) = P(X = x)$$

- In the case of discrete random variable, we use the notation Val(X) for the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $Val(X) = \{0, 1, 2, \ldots, 10\}$.

Probability Mass Function

Properties

- $0 \leq p_X(x) \leq 1$.
- $\sum_{x \in Val(X)} p_X(x) = 1$.
- $\sum_{x \in A} p_X(x) = P(X \in A)$.

Probability Density Function

- For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. In these cases, we define the **Probability Density Function** (PDF) as the derivative of the CDF, i.e.,

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

- The PDF for a continuous random variable may not always exist (i.e., if $F_X(x)$ is not differentiable everywhere)

Probability Density Function

Properties

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x) = 1.$
- $\int_{x \in A} f_X(x)dx = P(X \in A).$

- Suppose that X is a discrete random variable with Probability Mass Function MF $p_X(x)$ and $g : R \rightarrow R$ is an arbitrary function
- g(X) is a random variable, and we define the expectation or **expected value** of g(X) as:

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x) p_X(x)$$

Expected Value

- If X is a continuous random variable with PDF $f_X(x)$, then the **expected value** of g(X) is defined as:

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

- Intuitively, the expectation of g(X) can be thought of as a "weighted average" of the values that g(x) can taken on for different values of x, where the weights are given by $p_X(x)$ or $f_X(x)$
- As a special case of the above, note that the expectation, E[X] of a random variable itself is found by letting g(x) = x; this is also known as the mean of the random variable X.

Expected Value

Properties

- $E[a] = a$ for any constant $a \in \mathbb{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.
- (Linearity of Expectation) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
- For a discrete random variable X, $E[1\{X = k\}] = P(X = k)$.

Variance

- The variance of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean.
- Formally, the variance of a random variable X is defined as:

$$Var[X] \triangleq E[(X - E(X))^2]$$

- Using the properties Expected Value, we can derive an alternate expression for the variance:

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2, \end{aligned}$$

where the second equality follows from linearity of expectations and the fact that E[X] is actually a constant with respect to the outer expectation.

Variance

Properties

- $Var[a] = 0$ for any constant $a \in \mathbb{R}$.
- $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

Summary of PMF or PDF

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Bayes Theorem

$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}}$$

where A and B are events and $P(B) \neq 0$.

- $P(A \mid B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
- $P(B \mid A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$ and $P(B)$ are the probabilities of observing A and B respectively; they are known as the marginal probability.

https://en.wikipedia.org/wiki/Bayes%27_theorem

Bayes Theorem

- Derive Bayes Theorem:

Likelihood Function

- Measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters.
- It is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only, thus treating the random variables as fixed at the observed values
- The procedure for obtaining these arguments of the maximum of the likelihood function is known as **maximum likelihood estimation**, which for computational convenience is usually done using the natural logarithm of the likelihood, known as the log-likelihood function

Likelihood Function

Let X be a discrete random variable with probability mass function p depending on a parameter θ. Then the function

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x),$$

considered as a function of θ, is the *likelihood function*, given the outcome x of the random variable X.

Example

Consider a simple statistical model of a coin flip: a single parameter p_{H} that expresses the "fairness" of the coin. The parameter is the probability that a coin lands heads up ("H") when tossed. p_{H} can take on any value within the range 0.0 to 1.0. For a perfectly fair coin, $p_{\mathrm{H}} = 0.5$.

Imagine flipping a fair coin twice, and observing the following data: two heads in two tosses ("HH"). Assuming that each successive coin flip is i.i.d., then the probability of observing HH is

$$P(\mathrm{HH} \mid p_{\mathrm{H}} = 0.5) = 0.5^2 = 0.25.$$

Hence, given the observed data HH, the *likelihood* that the model parameter p_{H} equals 0.5 is 0.25. Mathematically, this is written as

$$\mathcal{L}(p_{\mathrm{H}} = 0.5 \mid \mathrm{HH}) = 0.25.$$

This is not the same as saying that the probability that $p_{\mathrm{H}} = 0.5$, given the observation HH, is 0.25. (For that, we could apply Bayes' theorem, which implies that the posterior probability is proportional to the likelihood times the prior probability.)

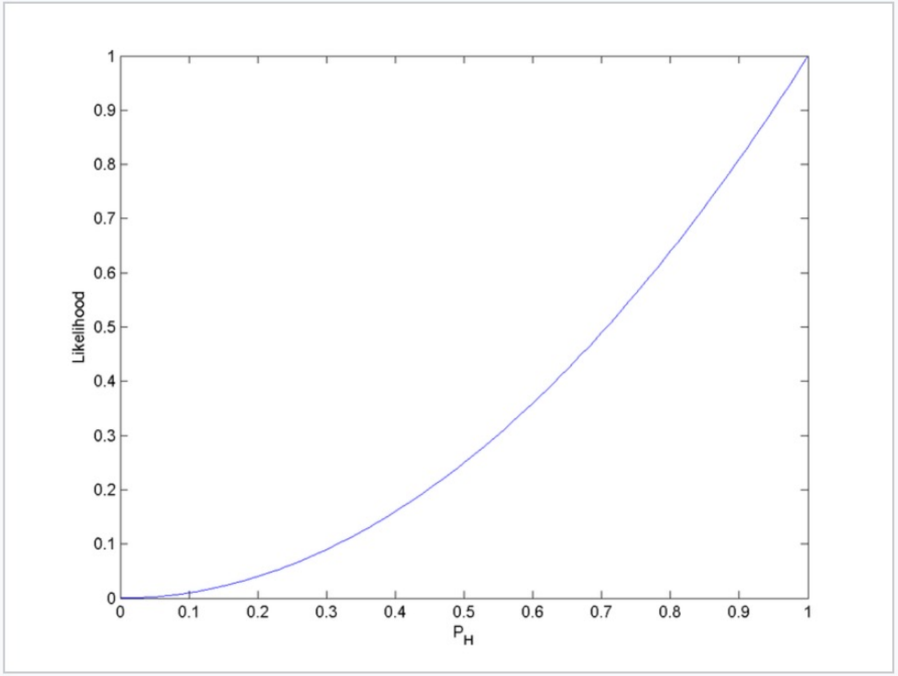
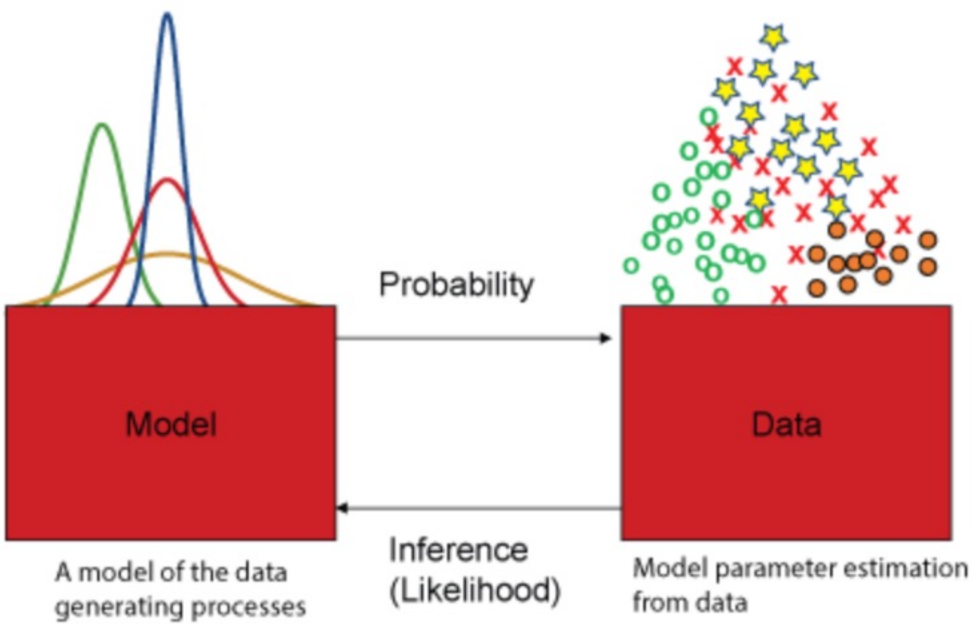


Figure 1. The likelihood function (p_{H}^2) for the probability of a coin landing heads-up (without prior knowledge of the coin's fairness), given that we have observed HH.

Likelihood Function

The Likelihood Function finds the *best* model given the data.

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$



Expectation Maximization

- Maximum likelihood estimation is an approach to density estimation for a dataset by searching across probability distributions and their parameters.
- It is a general and effective approach that underlies many machine learning algorithms, although it requires that the training dataset is complete, e.g. all relevant interacting random variables are present.
- Maximum likelihood becomes intractable if there are variables that interact with those in the dataset but were hidden or not observed, so-called latent variables.

https://machinelearningmastery.com/expectation-maximization-em-algorithm/

Expectation Maximization

- The expectation-maximization algorithm is an approach for performing maximum likelihood estimation in the presence of latent variables.
- It does this by first estimating the values for the latent variables, then optimizing the model, then repeating these two steps until convergence.
- It is an effective and general approach and is most commonly used for density estimation with missing data, such as clustering algorithms like the Gaussian Mixture Model.

https://machinelearningmastery.com/expectation-maximization-em-algorithm/

Expectation Maximization

- Maximum likelihood estimation is challenging on data in the presence of latent variables.
- Expectation maximization provides an iterative solution to maximum likelihood estimation with latent variables.
- Gaussian mixture models are an approach to density estimation where the parameters of the distributions are fit using the expectation-maximization algorithm.

https://machinelearningmastery.com/expectation-maximization-em-algorithm/

Parameter Learning

Score-based Metrics Approach

Maximize the likelihood of a set of observed data , which can be computed as the product of the probability of each observation

- Maximum Likelihood Estimation (MLE)
- Bayesian Information Criterion (BIC)
- Chow-Liu Algorithm
- Search Algorithms: Greedy, Hill-climbing, Local search, Simulated Annealing, Tabu search, Genetic Algorithms

Constraint-based Approach

Reflect the dependence and independence relations in the data that match the empirical distribution

- PC Algorithm
- Incremental Association Markov Blanket Algorithm

Francisco J. Cantú-Ortiz, PhD

Professor of Computer Science and Artificial Intelligence
Tecnológico de Monterrey
Enago-Academy Advisor for Strategic Alliances

E-mail: fcantu@itesm.mx, fjcantor@gmail.com
Cel: +52 81 1050 8294, SNI-2 CVU: 9804
Personal Page: http://semtech.mty.itesm.mx/fcantu/
Facebook: fcantu; Twitter: @fjcantor; Skype: fjcantor
Orcid: 0000-0002-2015-0562
Scopus ID:6701563520
Researcher ID: B-8457-2009
https://www.researchgate.net/profile/Francisco_Cantu-Ortiz
https://scholar.google.com.mx/citations?hl=es&user=45-uuK4AAAAJ
https://itesm.academia.edu/FranciscoJavierCantuOrtiz
Ave. Eugenio Garza Sada No. 2501, Monterrey N.L., C.P. 64849, México