Beta Distribution

listenzcc

March 9, 2020

Abstract

Bayesian learning is an active field of machine learning. It estimates parameters of known distribution rather than calculates coefficients directly. This article is a shallow instance of Bayesian learning. It shows how to estimate success rate with strong prior knowledge in Bernoulli experiment. Inevitably, it involves Beta distribution, Bernoulli distribution and their relationship of being conjugate prior pair. Additionally, it contains the calculation of the integral of Beta function using Γ function, which provides an easy path to calculate the expectation of mean and variance of Beta distribution.

Contents

1 Beta distribution

The *Beta distribution* can be used to a conjugate prior[1] to *Bernoulli distribution*.

[1] Conjugate prior. Suppose we have data with likelihood function $f(x|\theta)$ depending on a hypothesized parameter. Also suppose the prior distribution for θ is one of a family of parametrized distributions. If the posterior distribution for θ is in this family then we say the the prior is a conjugate prior for the likelihood.

1.1 Definition

The probability function $P(x)$ of Beta distribution is defined as proportional to the function of $x \in (-\infty, \infty)$

$$P(x) \propto x^{\alpha-1}(1-x)^{\beta-1} \tag{1}$$

there are two prior parameters, α and β controlling the shape of the distribution.

To guarantee that the integral of $P(x)$ is 1, and we have

$$P(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{2}$$

where $B(\alpha,\beta)$ is Beta function. The guarantee is based on a theorem, we will prove it in later.

Thus, the probability density function of Beta distribution can be expressed as

$$Beta(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{3}$$

The overlook of Beta distribution among different parameters combinations is drawn as Fig.1.

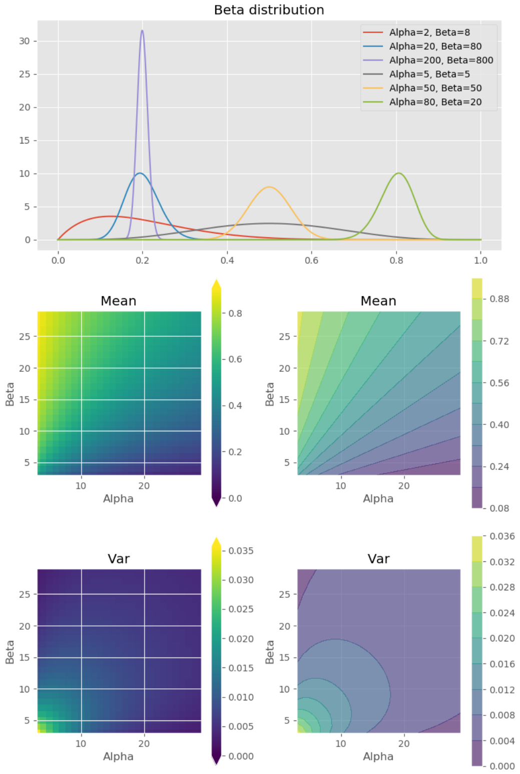

Figure 1: Beta distribution among different parameters combinations

2 Beta function

The *Beta function*[2] is important in calculus and analysis due to its close connection to the Γ function, which is itself a generalization of the factorial function. Many complex integrals can be reduced to expressions involving the beta function.

2.1 Definition

The Beta function, denoted by $B(x, y)$, is defined as

$$B(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \tag{4}$$

it is definite integrals.

2.2 Theorem

Theorem 2.1. *Beta function is symmetry*

$$B(\alpha, \beta) = B(\beta, \alpha) \tag{5}$$

Proof. Because of the convergent property of definite integrals

$$\int_0^a f(t)dt = \int_0^a f(a-t)dt$$

so we can rewrite the integral (4) as

$$B(\alpha, \beta) = \int_0^1 t^{\beta-1}(1-t)^{\alpha-1} dt$$

Thus, we get that Beta function is symmetric, $B(\alpha, \beta) = B(\beta, \alpha)$. □

Theorem 2.2. *The integral of Beta function can be calculated using* Γ *function.*

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{6}$$

where Γ *function is defined as*

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \tag{7}$$

Proof. One can write

$$\Gamma(m)\Gamma(n) = \int_0^\infty x^{m-1} e^{-x} dx \int_0^\infty y^{n-1} e^{-y} dy$$

Then rewrite it as a double integral

$$\Gamma(m)\Gamma(n) = \int_0^\infty \int_0^\infty x^{m-1} y^{n-1} e^{-x-y} dx dy$$

Applying the substitution $x = vt$ and $y = v(1-t)$, we have

$$\Gamma(m)\Gamma(n) = \int_0^1 t^{m-1}(1-t)^{n-1} dt \int_0^\infty v^{m+n-1} e^{-v} dv$$

Using the definitions of Γ and Beta functions, we have

$$\Gamma(m)\Gamma(n) = B(m, n)\Gamma(m+n)$$

Hence proved. □

[2] Beta function is also known as Euler's integral of the first kind

Lemma 2.1. *For positive integer α and β, Beta function can be expressed as*

$$B(\alpha,\beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} \tag{8}$$

since $\Gamma(\alpha) = (\alpha-1)!$ *when* $\alpha \in [1, 2, \ldots]$

3 Bernoulli distribution

3.1 Definition

The *Bernoulli distribution* is a discrete distribution having two possible outcomes labelled by $y = (0, 1)$ in which $y = 0$ ("success") occurs with probability p and $n = 0$ ("failure") occurs with probability $q = 1 - p$, where $0 \leq p \leq 1$. It therefore has probability density function.

In an experiment, the probability of obtaining output of Y is

$$P(y) = p^y (1-p)^{1-y} \tag{9}$$

In N experiments, the probability of obtaining n successes is

$$P(N,n) = \binom{N}{n} p^n (1-p)^{N-n} \tag{10}$$

To estimate the value of p based on N and n, we can apply MLE [3] method, in this case MLE is to solve $p = \hat{p}$ when it maximizes $P(N, n)$.

$$\hat{p} = \arg\max_p P(N,n) \tag{11}$$

The solution can be obtained with $\frac{\partial}{\partial \hat{p}} P(N,n) = 0$ yields $\hat{p} = \frac{n}{N}$. We will explain why in the following.

4 Beta and Bernoulli distributions

4.1 Bayes' Theorem and Conditional Probability

Bayes' theorem is a formula that describes how to update the probabilities of hypotheses when given evidence. It follows simply from the axioms of conditional probability, but can be used to powerfully reason about a wide range of problems involving belief updates.

Given a hypothesis H and evidence E, Bayes' theorem states that the relationship between the probability of the hypothesis before getting the evidence $P(H)$ and the probability of the hypothesis after getting the evidence $P(H|E)$ is

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{12}$$

where $P(H|E)$ is the posterior probability, conditional probability of the hypothesis given the evidence, and $P(E|H)$ is the posterior probability of the evidence being obtained when applying the hypothesis. The $P(E)$ is the universal probability of the evidence, which is usually difficult to know the exact value.

[3] MLE: Maximum Likelihood Estimation

Many modern *machine learning* techniques rely on Bayes' theorem. For instance, spam filters use Bayesian updating to determine whether an email is real or spam, given the words in the email. Additionally, many specific techniques in statistics, such as calculating *p-values* or *interpreting medical results*, are best described in terms of how they contribute to updating hypotheses using Bayes' theorem.

4.2 Conjugate prior

In a Bernoulli experiment, if we repeat the experiment N times, and obtain n successes. The estimation of p value is usually the first important problem to solve. Traditional method can give an estimation as $\hat{p} = \frac{n}{N}$, however it seems too arbitrary when N is small, and has ignored the useful of prior probability.

Here we try to estimate the value of p using Bayes' theorem. Firstly, we have an assumption that p fits Beta distribution with parameters of α and β

$$Beta(p;\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \tag{13}$$

The expectation of mean and variance of p is obvious when α and β are positive integers.

$$E(p) = \frac{\alpha}{\alpha+\beta} \tag{14}$$

$$D(p) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \tag{15}$$

Proof. Start from the definition of expectation

$$E(P) = \int_0^1 pBeta(p;\alpha,\beta)dp$$

it is easy to calculate the integral of the numerator is $B(\alpha+1,\beta)$. Thus the exception can be calculated by

$$E(P) = \frac{B(\alpha+1,\beta)}{B(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}$$

One can obtain the variance following

$$D(P) = E(P^2) - E^2(P)$$

where $E^2(P)$ is obvious and $E(P^2)$ can be calculated as

$$E(P^2) = \int_0^1 p^2 Beta(p;\alpha,\beta)dp = \frac{B(\alpha+2,\beta)}{B(\alpha+\beta)} = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$$

thus, we have

$$D(p) = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha\alpha}{(\alpha+\beta)(\alpha+\beta)}$$

□

Suppose we have a prior assumption that probability of success is p. The posterior probability of the n successes outcome is

$$P(E|H) = \binom{N}{n} p^n (1-p)^{N-n} \tag{16}$$

where H refers the assumption that probability of success is p, E refers obtaining n successes out of N experiments.

We can write prior probability as following

$$P(H) \propto p^{\alpha-1}(1-p)^{\beta-1} \tag{17}$$

Applying Bayes' theorem we have

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = p^{n}(1-p)^{N-n} \cdot p^{\alpha-1}(1-p)^{\beta-1} \cdot \mathcal{C} \tag{18}$$

where the denominator is simplified as a constant $\mathcal{C}$. The simplify is reasonable since the $P(E)$ is invariant with respect to p, despite its difficulty to solve.

It turns out that $P(H|E)$ is proportional to Beta function family as following

$$P(H|E) \propto p^{n+\alpha-1}(1-p)^{N-n+\beta-1} \tag{19}$$

applying substitution of $\alpha^{'} = n+\alpha$ and $\beta^{'} = N-n+\beta$, and denominator of Beta function, we have

$$P(H|E) = \frac{p^{\alpha^{'}-1}(1-p)^{\beta^{'}-1}}{B(\alpha^{'}, \beta^{'})} \tag{20}$$

we can see that the posterior probability $P(H|E)$ and prior probability $P(H)$ are from the same distribution family.

4.3 Bayesian learning

The way we apply Bayes' theorem using conjugate prior is the idea of Bayesian learning. We can not say which is better, Bayesian learning or MLE. However, we can say the Bayesian learning method is more stable since it yields smaller variance. The smaller variance is because Bayesian learning method uses the prior knowledge (17). In prior, we assumed that p is variable and centered on the point of $\frac{\alpha}{\alpha+\beta}$. After we obtained the evidence that n out of N successes, we formulate a closer estimation in (20). It can be found that the new estimation has much lower variance (15), since $\alpha^{'} > n$ and $\alpha^{'} + \beta^{'} > N$.

This explains another different between MLE and Bayesian learning method. MLE is more like a pure data driven method, it estimate coefficients based on only observed data. Bayesian learning uses data as adjustment for prior, the more data obtained the larger adjustment can be made. The adjustment strategy may provide a robust estimation against arbitrary noisy.