Probability for machine learning

Axioms of Probability

Axioms ensure that the probabilities assigned in a random experiment can be interpreted as relative frequencies and that the assignments are consistent with our intuitive understanding of relationships among relative frequencies:

1. $0 \leq P(E) \leq 1$. If E1 is an event that cannot possibly occur, then P(E1) = 0. If E2 is sure to occur, P(E2) = 1.
2. S is the sample space containing all possible outcomes, P(S) = 1.
3. If Ei , i = 1,...,n are mutually exclusive (i.e., if they cannot occur at the same time, as in Ei ∩

Ej = , j = i, where is the null event that does not contain any possible outcomes), we have

∅ ∅

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$$

For example, letting Ec denote the complement of E, consisting of all possible outcomes in S that are not in E, we have E ∩ EC = ∅ and

$$P(E \cup E^c) = P(E) + P(E^c)$$

$$P(E^c) = 1 - P(E)$$

If the intersection of E and F is not empty, we have

$$P(E \cup F) = P(E) + P(F) - P$$

Conditional Probability

P(E |F) is the probability of the occurrence of event E given that F occurred and is given as

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Knowing that F occurred reduces the sample space to F, and the part of it where E also occurred is E∩F. Note that equation is well-defined only if P(F) > 0. Because ∩ is commutative, we have

$$P(E \cap F) = P(E|F)P(F) = P(F|E)P(E)$$

which gives us Bayes' formula:

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

When Fi are mutually exclusive and exhaustive

$$E = \bigcup_{i=1}^{n} E \cap F_i$$

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

Bayes' formula allows us to write

$$P(F_i|E) = \frac{P(E \cap F_i)}{P(E)} = \frac{P(E|F_i)P(F_i)}{\sum_j P(E|F_j)P(F_j)}$$

If E and F are independent, we
have P(E|F) = P(E) and thus

$$P(E \cap F) = P(E)P(F)$$

That is, knowledge of whether F has occurred does not change the probability that E occurs.

Random Variables

A random variable is a function that assigns a number to each outcome in the sample space of a random experiment.

Probability Distribution and Density Functions

The probability distribution function F(·) of a random variable X for any real number a is

F(a) = P{X ≤ a} and we have

P{a<X ≤ b} = F(b) − F(a)

$$F(a) = \sum_{\forall x \leq a} P(x)$$

If X is a discrete random variable

where P(·) is the probability mass function defined as P(a) = P{X = a}. If X is a continuous random variable, p(·) is the probability density function such that

$$F(a) = \int_{-\infty}^{a} p(x)dx$$

Joint Distribution and Density Functions

In certain experiments, we may be interested in the relationship between two or more random variables, and we use the joint probability distribution and density functions of X and Y satisfying
F(x, y) = P{X ≤ x, Y ≤ y}

Individual marginal distributions and densities can be computed by marginalizing, namely, summing over the free variable:
FX(x) = P{X ≤ x} = P{X ≤ x, Y ≤ ∞} = F(x, ∞)

In the discrete case, we write

$$P(X = x) = \sum_{j} P(x, y_j)$$

and in the

These can be generalized in a straightforward manner to more than two random variables.

Conditional Distributions

When X and Y are random variables

$$P_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{P\{X = x, Y = y\}}{P\{Y = y\}} = \frac{P(x,y)}{P_Y(y)}$$

Bayes' Rule

When two random variables are jointly distributed with the value of one known, the probability that the other takes a given value can be computed using Bayes' rule:

$$P(y|x) = \frac{P(x|y)P_Y(y)}{P_X(x)} = \frac{P(x|y)P_Y(y)}{\sum_y P(x|y)P_Y(y)}$$

Or, in words

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Expectation

Expectation, expected value, or mean of a random variable X, denoted by E[X], is the average value of X in a large number of experiments:

$$E[X] = \begin{cases} \sum_i x_i P(x_i) & \text{if } X \text{ is discrete} \\ \int x p(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

It is a weighted average where each value is weighted by the probability that X takes that value. It has the following properties (a, b $\in \Re$)

$$E[aX + b] = aE[X] + b$$
$$E[X + Y\ ] = E[X] + E[Y$$

For any real-valued function g(·), the expected value is]

$$E[g(X)] = \begin{cases} \sum_i g(x_i)P(x_i) & \text{if } X \text{ is discrete} \\ \int g(x)p(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

A special g(x) = x^n , called the nth moment of X, is defined as

$$E[X^n] = \begin{cases} \sum_i x_i^n P(x_i) & \text{if } X \text{ is discrete} \\ \int x^n p(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

Mean is the first moment and is denoted by μ

Variance

Variance measures how much X varies around the expected value. If μ ≡ E[X], the variance is defined as

Var(X) = E[(X − μ)2] = E[X2] − μ2

Variance is the second moment minus the square of the first moment. Variance, denoted by σ2, satisfies the following property (a, b ∈ℜ): Var(aX + b) = a2Var(X)

Cov(X, Y) = E [(X − μX)(Y

− μY)] = E[XY] − μXμY

where μX ≡ E[X] and μY ≡

E[Y]. Some other properties

are

Cov(X, Y) = Cov(Y, X)

Cov(X, X) = Var(X)

Cov(X + Z,Y) = Cov(X, Y) + Cov(Z, Y)

$$\mathrm{Cov}\left(\sum_i X_i, Y\right) = \sum_i \mathrm{Cov}(X_i, Y)$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i) + \sum_i \sum_{j \neq i} \mathrm{Cov}(X_i, X_j)$$

If X and Y are independent, E[XY] = E[X]E[Y] = μXμY and Cov(X, Y) = 0. Thus if Xi are independent

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$$

Correlation is a normalized, dimensionless quantity that is always between −1 and 1:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Weak Law of Large Numbers

Let X={Xt}Nt=1 be a set of independent and identically distributed (iid) random variables each having to mean μ and a finite variance σ^2. Then for any ǫ > 0

$$P\left\{\left|\frac{\sum_t X^t}{N} - \mu\right| > \epsilon\right\} \rightarrow 0 \text{ as } N \rightarrow \infty$$

That is, the average of N trials converge to the mean as N increases.

Special Random Variables

Bernoulli Distribution

A trial is performed whose outcome is either a “success” or a “failure.” The random variable X is a 0/1 indicator variable and takes the value 1 for a successful outcome and is 0 otherwise. P is the probability that the result of trial the is a success. Then P{X = 1} = p and P{X = 0} = 1 − p which can equivalently be written as

P{X = i} = pi(1 − p)1−i , i = 0, 1

If X is Bernoulli, its expected value and variance are

E[X] = p, Var(X) = p(1 − p)

Binomial Distribution

If N identical independent Bernoulli trials are made, the random variable X that represents the number of successes that occurs in N trials is binomial distributed. The probability that there are i successes is

$$P\{X=i\} = \binom{N}{i} p^i (1-p)^{N-i}, i = 0 \ldots N$$

If X is binomial, its expected value and variance are
E[X] = Np, Var(X) = Np(1 − p)

Multinomial Distribution

Consider a generalization of Bernoulli where instead of two states, the outcome of a random event is one of K mutually exclusive and exhaustive states

$$P(N_1, N_2, \dots, N_K) = N! \prod_{i=1}^{K} \frac{p_i^{N_i}}{N_i!}$$

A special case is when N = 1; only one trial is made. Then Ni is a 0/1 indicator variable of which only one of them is 1 and all others are 0.

$$P(N_1, N_2, \dots, N_K) = \prod_{i=1}^{K} p_i^{N_i}$$

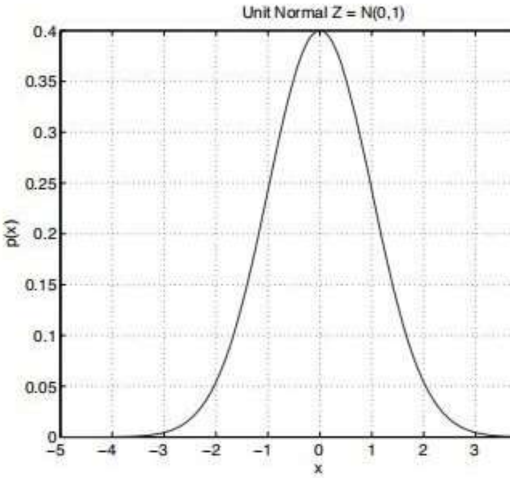
Uniform Distribution

X is uniformly distributed over the interval [a, b] if its density function is given by

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

If X is uniform, its expected value and variance are

$$E[X] = \frac{a+b}{2}, \ \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$



Normal (Gaussian) Distribution

X is normal or Gaussian distributed with mean μ and variance σ2 denoted as N (μ, σ2) if its density function is

If X ~ N (μ, σ2) and Y

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$$

$$p_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$

~

= aX + b, then Y

N(aμ + b, a2σ2). The sum of independent normal variables is also normal with μ = i μi and σ2 = i σi2 . If X is N (μ, σ2), then

$$\frac{X-\mu}{\sigma} \sim Z$$

This is called z-normalization.

C

hi-Square Distribution If Zi are

independent unit normal random

variables, then

$$X = Z_1^2 + Z_2^2 + \ldots + Z_n^2$$

is chi-square with n degrees of freedom, namely, X ~X2n , with E[X] = n, Var(X) = 2n When Xt~ N(μ, σ2), the estimated sample variance is

$$S^2 = \frac{\sum_t (X^t - m)^2}{N-1}$$

and we have

$$(N-1)\frac{S^2}{\sigma^2} \sim X^2_{N-1}$$

It is also known that m and S2 are independent.

t

Distribution If Z $\sim$ Z

and X $\sim$ X^2n are

independent, then

$$T_n = \frac{Z}{\sqrt{X/n}}$$

is t-distributed with n degrees of freedom with

$$E[T_n] = 0, n > 1,\ \mathrm{Var}(T_n) = \frac{n}{n-2}, n > 2$$

F Distribution

If X1 $\sim$ X2n and X2 $\sim$ X2m are independent chi-square random variables with n and m degrees of freedom respectively,

$$F_{n,m} = \frac{X_1/n}{X_2/m}$$

is F-distributed with n
and m degrees of
freedom with

$$E[F_{n,m}] = \frac{m}{m-2}, m > 2, \ \mathrm{Var}(F_{n,m}) = \frac{m^2(2m+2n-4)}{n(m-2)^2(m-4)}, m > 4$$