Probability and Statistics for Machine Learning

Dr.Danish Lohani

South Asian University,NewDelhi

December 15, 2013

Outline

1 Basic Concepts
- Definition
- Axioms of Probability
- Properties of Probability
- Exercises
- Independence
- Conditional Probability

2 Random Variables
- Classification
- Joint Density Functions
- Distribution functions

3 Inference Methods
- Bayes Rule
- Bayes Decision Theory
- Discriminant Functions

Basic Concepts

Random Experiment

Origin of probability theory lies in observations associated with games of chance.A random experiment is an action or process that leads to one of many possible outcomes.(Tossing coins,Throwing dices,Picking a card from a shuffle..etc)

The list of possible outcomes of a random experiment must be exhaustive and mutually exclusive

Basic Concepts

Sample Space

The collection of all possible outcomes of an experiment is a set Ω called the sample space.

Event

An event is a collection or set of one or more simple events in a sample space.

Definition

Basic Concepts
Definition
Axioms of Probability
Properties of Probability
Exercises
Independence
Conditional Probability
Random Variables
Classification
Joint Density Functions
Distribution functions
Inference Methods
Bayes Rule
Bayes Decision Theory
Discriminant Functions

Classical Approach

The Probability of an event is the number of outcomes favourable to the event,divided by the total number of outcomes,where all outcomes are equal likely.

Example

Roll of a Die $\Omega = \{1, 2, 3, 4, 5, 6\}$
Each number has $\frac{1}{6}$ chance of occuring

Definition

Frequentist Approach

It defines an event's probability as the limit of its relative frequency in a large number of trials.
$\lim_{n_t \to \infty} P(x) = \frac{n_x}{n_t}$

Example

Roll the die 100 times and suppose number of times 1 occurs is 10.
Here, $\Omega = \{1, 2, 3, 4, 5, 6\}$
$A = \{1\}$
$P(A) = \frac{1}{10}$

Axioms of Probability

A probability function P:$\Omega \rightarrow R[0,1]$ is defined on subsets of the sample space Ω to satisfy the following axioms:

1. Non-Negative Probability:
   $P(E) \geq 0$
2. Mutually-Exclusive Events:
   $P(E_1 \cup E_2) = P(E_1) + P(E_2)$
   provided E_1 and E_2 are mutually exclusive.
3. The Universal Set:
   $P(\Omega) = 1$

Some Properties

1. $P(\bar{E}) = 1 - P(E)$
2. $P(\Phi) = 0$
3. If E_1 and E_2 are subsets of Ω such that $E_1 \subset E_2$ then $P(E_1) \leq P(E_2)$
4. For each E $\subset \Omega$,$0 \leq P(E) \leq 1$.
5. Sum Rule $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.
6. $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum\sum_{i<j} P(A_i)P(A_j) \sum \sum_{i<j<k} \sum P(A_i)P(A_j) - \ldots.. + (-1)^{n+1} P(A_1 A_2 \ldots A_n)$.

Problem 1

If a 3-digit number (000 - 999) is chosen at random, find the probability that exactly 1 digit will be ≥ 5.

Problem 2

An urn contains 3 red,8 yellow and 13 green balls; another urn contains 5 red,7 yellow, and 6 green balls. One ball is selected from each urn. Find the probability that both the balls will be of same color.

Problem 3

If a box contains 75 good light bulbs and 25 defective bulbs and 15 bulbs are removed, find the probability that atleast one will be defective.

Problem 4

Eight cards are drawn without replacement from an ordinary deck.Find the probability of obtaining exactly three aces or exactly three kings.

Independent Events

Two events A and B are independent iff $P(A \cap B) = P(A)P(B)$

Remarks

1. P(A$\cap B$) = P(A) P(B) implies $P(A^c \cap B) = P(A^c)P(B)$
2. The condition $P(A_1 \cap .... \cap A_n) = P(A_1)...P(A_n)$ does not imply the analogous condition for smaller events.

   It is possible to have P $(A \cap B \cap C) = P(A)P(B)P(C)$ but P(A$\cap B$) $\neq P(A)P(B)$

Problem 5

Let two dice be tossed, and take $\Omega =$ all ordered pairs (i,j),i,j=1,2..6 with each point assigned probability $\frac{1}{36}$.

Let,
A = { first die = 1,2, or 3}
B = {first die = 3,4,or 5}
C = {the sum of two faces is 9}

Verify,
P $(A \cap B \cap C) = P(A)P(B)P(C)$ and P(A$\cap B) = P(A)P(B)$

Independent events..

Bernoulli trials

A sequence of n Bernoulli trails is a sequence of n independent observations each of which may either be a failure or a success.

At each observationthe probability of sucess is p and failure is 1-p

Binomial Probability function

The probability Obtaining exactly k success in n trails is

$$p(k) = \binom{n}{k} p^k q^{n-k}$$

Problem

Basic Concepts
Definition
Axioms of Probability
Properties of Probability
Exercises
Independence
Conditional Probability
Random Variables
Classification
Joint Density Functions
Distribution functions
Inference Methods
Bayes Rule
Bayes Decision Theory
Discriminant Functions

An urn contains t_1 balls of colour C_1, t_2 of colour $C_2, \cdots, t_k$ of colour C_k.If n balls are drawn without replacement, find the probabilty of obtaining n_1 balls of colour C_1, n_2 of colour $C_2, \cdots, n_k$ of colour C_k. Also find probability for the case when balls are drawn independently, with replacement.

Problem

If X is the number of success in n-Bernoulli trials, find the probability that $X \geq 3$ given that $X \geq 1$.

Independent Events..

Multinomial Probability function

The probability that b_1 will occur n_1 times, b_2 will occur n_2 times .. and b_k will occur n_k times is
$p(n_1..n_k) = \frac{n!}{n_1!....n_k!} p_1^{n_1}....p_k^{n_k}$

Problem 6

If 10 balls are tossed independently in to five boxes, with a given ball equal likely to fall in to each box, find the probability that all boxes will have the same number of balls.

Conditional Probability

Basic Concepts
Definition
Axioms of Probability
Properties of Probability
Exercises
Independence
Conditional Probability
Random Variables
Classification
Joint Density Functions
Distribution functions
Inference Methods
Bayes Rule
Bayes Decision Theory
Discriminant Functions

Definition

The conditonal probability of B given A is defined as
$P(B|A) = \frac{P(A \cap B)}{P(A)}$

Product Rule

$P(A \cap B) = P(B|A)P(A)$

Theorem of Total Probability

Let B_1,B_2 ... be a finite or countably infinite family of mutually exclusive and exhaustive events(i.e, the B_i are disjoint and their union is Ω). Uf A is any event, the
$P(A) = \sum_i P(A \cap B_i)$

Problem 7

Two coins are available , one unbiased and the tother two-headed.Choose a coin at random and toss it once; Assume that the unbiased coin is chosen with probability $\frac{3}{4}$. Given that the result is heads, find the probability that the two-headed coin was chosen.

Problem 8

In a certain village 20% of the population has a disease D. A test is administered which has the property that if a person has D, the test will be positive 90% of the time , and if he does not have D , the test will still be positive 30% of the time.All those who test is positive are given a drug which invariably cures the disease, but produces a characteristic rash 25% of the time. Given that a person picked at random has the rash, what is the probability that he actually had D to begin with?

Random Variables

Definition

A random variable is a real-valued function on a sample space.

Example

1. Throw a coin 10 times, and let R be the number of heads. We take $\Omega =$ all sequences of length 10 with components H and T.
   For a sample point A={HHTHTTHHTH}, R(A)=6.
2. Throw two dice. We may take the sample space to be the set of all pairs of integers (x,y).(36 points in all).
   Let,
   $R_1 =$ the result of first toss. Then $R_1(x, y) =$ x
   $R_2 =$ the sum of two faces. Then $R_2(x, y) =$ x+y

Classification of Random Variables

Basic Concepts
Definition
Axioms of Probability
Properties of Probability
Exercises
Independence
Conditional Probability
Random Variables
Classification
Joint Density Functions
Distribution functions
Inference Methods
Bayes Rule
Bayes Decision Theory
Discriminant Functions

Discrete Random Variable

A random variable (R) is said to be discrete iff the set of possible values is finite or countably infinite.

If $x_1, x_2...$ are the value of R that belong to B,then
$P\{R \in B\} = P\{R = x_1 \; or \; R = x_2 \; or \; .....\} = \sum_{x \in B} p_R(x)$

where $p_R(x)$ is the probability function of R,defined by,
$p_R(x) = P\{R = x\}$

Remark

Another way of characterizing R is by means of the distribution function defined by
$F_R(x) = P\{R \leq x\}$, x real

Classification of Random Variables

Continuos Random Variables

Continuous random variables can take any value in an interval. They are used to model physical characteristics such as time, length, position, etc.

Definition

X is a continuous random variable if there is a function f(x) so that for any constants a and b, with $-\infty \leq a \leq b \leq \infty$

$P(a \leq X \leq b) = \int_a^b f(x)dx$

where f(x) is called the probability density function.

Cumulative Distribution function

$F_R(x) = \int_{-\infty}^{x} f_R(t)dt$
where $f_R(t)$ is the probability density function.
Differentiating both sides we get $\frac{dF}{dx} = f(x)$

Properties

1. F(x) is non-decreasing; a ¡ b implies F(a) ¡ F(b)
2. $\lim_{x\to\infty} F(x) = 1$
3. $\lim_{x\to-\infty} F(x) = 0$

Joint Density Functions

Two Discrete Random Variables

Joint Probability Mass Function

Let X and Y be two discrete random variables defined on Sample space Ω of an experiment.The joint probability mass function is defined for each pair of numbers (x, y) by,
p(x,y) = P(X=x and Y=Y) , for all $(x, y) \in X * Y$

Marginal Probability mass function

$p_X(x) = \sum_y p(x, y)$
$p_Y(y) = \sum_x p(x, y)$

Joint Density Functions
Two Continuous Random Variables

Joint Probability Density Function

Let X and Y be two continuous random variables.The joint probability density function f(x,y) is a function satisfying $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dydx = 1$.

Then for any two dimensional set A,
P[(X,Y)]$\in A = \int \int_A f(x, y) dydx$

Marginal Probability density function

$p_X(x) = \int_{-\infty}^{\infty} p(x, y) dy$
$p_Y(y) = \int_{-\infty}^{\infty} p(x, y) dx$

Problem

Let $f_{12}(x, y) = 1 \; if \; 0 \leq x \leq 1 \; and 0 \leq y \leq 1$ and 0 elsewhere. Calculate the probability that $\frac{1}{2} \leq X + Y \leq \frac{3}{2}$

Problem

Let R_1 and R_2 be independent and absolutely continuous random variables, each uniformly distributed between 0 and 1. Find the density function of the random variable R_3, where $R_3 = \frac{\max(R_1,R_2)}{\min(R_1,R_2)}$

Problem

Define uniform density function? Let (R_1, R_2) be uniformly distributed over the parallelogram with vertices $(-1, 0), (1, 0), (2, 1), (0, 1)$.
(a) Find the density function of $R_1, \& R_2$.
(b) A new random variable R_3 is defined by $R_3 = R_1 + R_2$. Find its density.

Problem

Let $f_{12}(x, y) = 1$ *if* $0 \leq y \leq x \leq 1$ and 0 elsewhere.

Find $f_1(x)$ and $f_2(y)$

Problem

A man and women agree to meet at a certain place some time between 11 and 12 o'clock. they agree that the one arriving first will wait z hours, $0 \leq z \leq 1$, for the other to arrive. Assuming that the arrival times are independent and uniformly distributed, find the probability that they will meet.

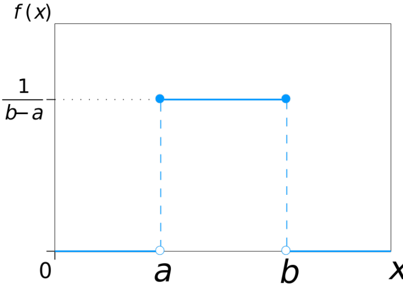
Problem

Two points are to be choosen at random on a line segment whose length is $a > 0$. Find the probability that the three line segments thus formed will be sides of a triangle.

Uniform Distribution function

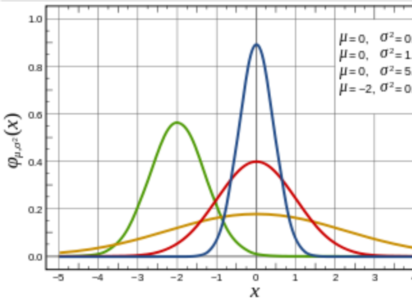
Basic Concepts
Definition
Axioms of Probability
Properties of Probability
Exercises
Independence
Conditional Probability
Random Variables
Classification
Joint Density Functions
Distribution functions
Inference Methods
Bayes Rule
Bayes Decision Theory
Discriminant Functions

- Density $\frac{1}{(b-a)}$, a $\leq$x $\leq$ b
- Parameters a,b real, a $\leq$ b



Normal Distribution function

- Density
  $\frac{1}{\sqrt{2\pi\sigma}} \exp(\frac{-(x-\mu)^2}{2\sigma^2})$
- Parameters
  μ real, $\sigma > 0$

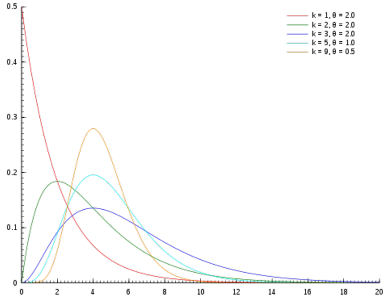


Gamma Distribution function

- Density

$$\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$$

- Parameters
  $\alpha,\ \beta > 0$

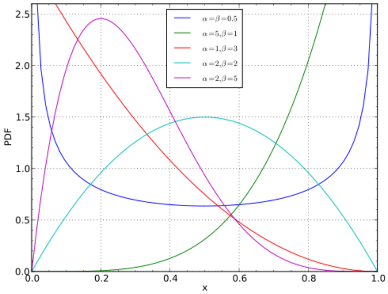


Beta Distribution function

- Density

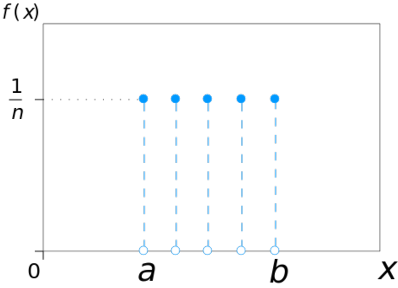
$$\frac{x^{r-1}(1-x)^{s-1}}{\beta(r,s)}$$

- Parameters
r,s > 0



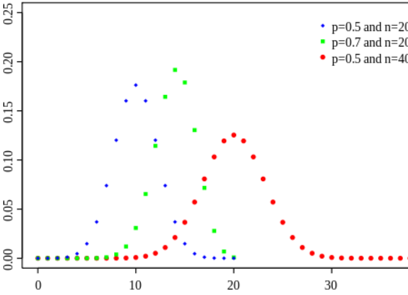
Discrete Uniform Distribution function

- Probability Mass function
  p(k)=$\frac{1}{(N)}$ k=1,2...N
- Parameter N= 1,2...



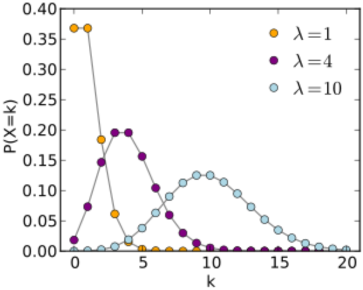
Binomial Distribution

- Probability Mass function
  $$p(k) = \binom{n}{k} p^k q^{n-k}$$
- Parameters
  $0 \ \leq \ p\, 1$, $q = 1 - p$,
  n=1,2...



Poisson Distribution

- Probability Mass function
  $p(k) = \frac{e^{-\lambda}\lambda^k}{k!}$
- Parameters
  $\lambda > 0$



Expectation

Expectation (also called as mean or average value) of a random variable is defined as

$\mathrm{E(R)} = \sum_{x} xP(R = x)$ when R is discrete

$\mathrm{E(R)} = \int_{-\infty}^{\infty} xf_R(x)$ when R is continuous

If R is identically constant,say R $=$ c , then E(R) $=$ c

Moments

If R is a random variable on a given probability space, the k^{th} moment of R (k >0), is defined by

$$\alpha_k = \sum_x^k xP(R = x) \text{ when R is discrete}$$

$$\alpha_k = \int_{-\infty}^{\infty} x^k f_R(x) \text{ when R is continuous}$$

The kth central moment of R is defined by, $\beta_k = E[(R - m)^k]$
β_2 is variance of R denoted by σ^2 , σ is called standard deviation.

For R_1 and R_2 , the joint central moment is
$\beta_{jk} = E[(R_1 - m_1)^j (R_2 - m)^k]$ β_{11} is called Co-Variance of R_1 and R_2.

Bayes Rule

Basic Concepts
Definition
Axioms of Probability
Properties of Probability
Exercises
Independence
Conditional Probability
Random Variables
Classification
Joint Density Functions
Distribution functions
Inference Methods
Bayes Rule
Bayes Decision Theory
Discriminant Functions

Bayes Formula

$$P(\frac{\omega_j}{x}) = \frac{p(x/\omega_j)P(\omega_j)}{p(x)}$$

Bayes rule show how observing the value of x changes a Prior probability $P(\omega_j)$ to a Posterior probability $P(\omega_j/x)$.

The term $p(x/\omega_j)$ is called Likelihood.

Basic Concepts
- Definition
- Axioms of Probability
- Properties of Probability
- Exercises
- Independence
- Conditional Probability

Random Variables
- Classification
- Joint Density Functions
- Distribution functions

Inference Methods
- Bayes Rule
- Bayes Decision Theory
- Discriminant Functions

Let ω denote the state of natue of a random variable with two possible states ω_1 and ω_2.

Let $P(\omega_1)$ and $P(\omega_2)$ be prior probabilities.The problem is to assign a value to ω after a observation x.

Decide $\omega = \omega_1$ if p(ω_1 /x) >p(ω_2 /x).
Decide $\omega = \omega_2$ if p(ω_2 /x) >p(ω_1 /x).

Error

P(error/x) = P(ω_1/x) if we decide ω_2
P(error/x) = P(ω_2/x) if we decide ω_2

Suppose we observe x and that we contemplate taking decision α_i. If true state of nature is ω_j we incur the loss $\lambda(\alpha_i/\omega_j)$.

Conditional Risk

$$R(\alpha_i/x) = \sum_{j=1}^{s} \lambda(\alpha_i/\omega_j) P(\omega_j/x)$$

Select the action for which R(α_i/x) is minimum.

Basic Concepts
Definition
Axioms of Probability
Properties of Probability
Exercises
Independence
Conditional Probability
Random Variables
Classification
Joint Density Functions
Distribution functions
Inference Methods
Bayes Rule
Bayes Decision Theory
Discriminant Functions

Let , $\lambda_{ij} = \lambda\,(\alpha_i/\omega_j)$
Conditional risk for two category classification is given by
$\mathrm{R}(\alpha_1/\mathrm{x}) = \lambda_{11} P(\omega_1/x) + \lambda_{12} P(\omega_2/x)$
$\mathrm{R}(\alpha_2/\mathrm{x}) = \lambda_{21} P(\omega_1/x) + \lambda_{22} P(\omega_2/x)$
In terms of Posterior Probabilities we decide ω_1 if

$$(\lambda_{21} - \lambda_{11}) P(\omega_1/x) > (\lambda_{12} - \lambda_{22}) P(\omega_2/x)$$

Likelihood Ratio

$\frac{p(x/\omega_1)}{p(x/\omega_2)} > \frac{\lambda_{12}-\lambda_{22})P(\omega_2)}{(\lambda_{21}-\lambda_{11})P(\omega_1)}$ when $\omega = \omega_1$

Minimum Error Rate Classification

Zero-One Loss Function

$\lambda(\alpha_i/\omega_j) = 0$ if i=j and 1 if i $\neq$j

Conditional Risk is $R(\alpha_i/x) = \sum_{j=1}^{c} \lambda(\alpha_i/\omega_j) P(\omega_j/x)$

$=\sum_{j \neq i} P(\omega_j/x)$

$=1-P(\omega_i/x)$

Decision Rule

Decide ω_i if $P(\omega_i/x) > P(\omega_j/x)$ for all j $\neq$ i

Discriminant Functions

The classifer is said to assign a feature vector x to class ω_i if $g_i(x) > g_j(x)$ for all $j \neq i$.
For minimum error rate classification, any of the following choices give identical classificiation results, but some can be much simpler to compute than others.

1. $g_i(x) = P(\omega_i/x)$
2. $g_i(x) = \frac{p(x/\omega_j)P(\omega_i)}{\sum_{j=1}^{c} p(x/\omega_j)P(\omega_j)}$
3. $g_i(x) = p(x/\omega_i)\ \mathrm{P}(\omega_i)$
4. $g_i(x) = \log\ p(x/\omega_i) + \log\ \mathrm{P}(\omega_i)$

Problem 11

Consider the following decision rule for a two-category one dimensional problem. Decide ω_1 if a $>$0; otherwise decide ω_2. Show that the probability of error for this rule is given by, $\mathrm{P(error)} = \mathrm{P}(\omega_1)\int_{-\infty}^{\theta} p(x/\omega_1)dx + \mathrm{P}(\omega_2)\int_{0}^{\infty} p(x/\omega_2)dx$.

By differentiating , show that a necessary condition to minimize P(error) is that to satisfy,
$p(\theta/\omega_1)P(\omega_1) = p(\theta/\omega_2)P(\omega_2)$

Problem

In many pattern classification problems one has the option either to assign the pattern to one of c classes or to *reject* it as being unrecognizable. If the cost for rejects is not high, rejection may be the desiarable action. Let

$$\lambda\left(\frac{\alpha_i}{w_j}\right) = \begin{cases} 0 & \text{if } i = j \quad i, j = 1, \cdots, c; \\ \lambda_r & \text{if } i = c+1; \\ \lambda_s & \text{otherwise.} \end{cases}$$

where λ_r is the loss incurred for choosing the $(c+1)$th action of rejection , and λ_s is the loss incurred for making a substitution error. Show that the minimum risk is obtained if we decide w_i if $p\left(\frac{w_i}{x}\right) \geq p\left(\frac{w_j}{x}\right)$ for all j and if $p\left(\frac{w_i}{x}\right) \geq 1 - \frac{\lambda_r}{\lambda_s}$ and reject otherwise. What happens if $\lambda_r = 0$? what happens if $\lambda_r > \lambda_s$?

Problem

If a and b are non negative numbers, where $\min(a, b) \leq \sqrt{ab}$. Use this to show that the error rate for a two category Bayes classifer must satisfy $p(error) \leq \sqrt{p(w_1)p(w_2)}\rho \leq \frac{1}{2}\rho$, where

$$\rho = \int \left[p\left(\frac{x}{w_1}\right) p\left(\frac{x}{w_2}\right) \right]^{\frac{1}{2}} dx$$

Problem

Let the conditional densities for a two category one-dimensional problem be given by the cauchy distribution $p\left(\frac{x}{w_i}\right) = \frac{1}{\pi b} \frac{1}{1+\frac{(x-a_i)^2}{b^2}}$, $i = 1, 2$. If $p(w_1) = p(w_2)$, show that $p\left(\frac{w_1}{x}\right) = p\left(\frac{w_2}{x}\right)$ if $x = \frac{1}{2}(a_1 + a_2)$.