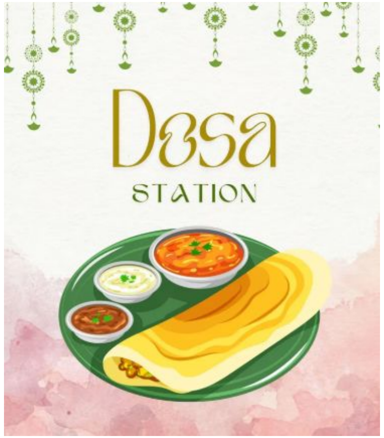
Data Analytics

Mrinal Das

Probability: Theory of Chance

Matter of Chance



Options Lead to Chances

There are 2 machines for you to do the same task.

Machine-1 is 2 times faster than Machine-2

But Machine-1 can fail randomly

How will you choose a machine if you have to complete task within a deadline ?

Experiments and Outcomes

- Any natural or artificial activity that produces some outcomes are experiments
- Some experiments are random if the outcome is not certain
- There is a set of possible outcomes – sample space
- Each outcome has a chance of occurring – uncertainty
- Probability is the tool to measure this uncertainty

Experiments and Outcomes

Choosing a machine to do the task is an experiment

Possible outcomes are task complete or incomplete

Why is this uncertain ?

Machine-1 can fail arbitrarily

Machine-2 can delay arbitrarily

Probability

Let's study only one machine at a time

Let $X=0$ if machine fails, task incomplete

$X=1$ if machine finishes the task in time, task complete

There is some randomness in X taking 0 or 1

Probability is the tool to measure this randomness

What is the probability that task will be completed ? $P(X=1)=?$

Random Variables

Outcomes of an experiment need not be numeric, like completion of task

But for the sake of numerical calculations, we map them to numbers

Actually we map events to numbers, not outcomes

An event is a collection of outcomes

For example, task completion is an event

Task not completed is another event, but this may be due to machine failure or machine delay

Construction of Probability

Task completed and not-completed are only two options

So either one of them should occur, but both can not occur

So probability of these to events sum to 1

Why 1 ? This is a design that helps, nothing else

So probability of neither of them occurring is 0

Probability of any of them occurring is between 0 and 1

$P(X=1) + P(X=0)=1$

Measuring the probability

To make the decision, we need to know the values,

What is $P(X=1)$? Is it higher than $P(X=0)$?

We need to estimate this. Then only we can take informed decision.

Let's go back to data collection and analysis.

We run this experiment n times, and the task got completed m times

So, $P(X=1) = m/n$, and $P(X=0)=(n-m)/n$

Clearly $P(X=1)+p(X=0)=1$

Types of Random Variables

As X can take only 2 values, X is binary

If X takes k values then X is not binary but it is still discrete

So X is a discrete random variable.

Let Z is another random variable that captures percentage of completion of task

Z can take any value between 0 to 100, like 1, 5.9, 99.99999 etc

The possible values are not known, only the range is known

Z is called continuous random variable.

Distributions

For binary X, we use Bernoulli distribution

P(X=1)=p, p is a parameter of Bernoulli distribution

We say X ~ Bernoulli(p)

Distributions help in calculating the probabilities.

There are many other discrete distributions.

Similarly there are continuous distributions.

Gaussian or Normal distribution is the most common continuous distribution.

Probability Mass Function (pmf)

Let X is a discrete random variable

X takes values in {1,2,...,k}

Probability of X taking these values is measured by pmf

Distributions help us to compute pmf

Probability Density Function (pdf)

If X is continuous random variable

Then X takes a value in a range which is measured by pdf

Distributions help us to calculate the pdf

Expectation

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

X is continuous RV

$$\mathrm{E}[X] = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

X is discrete RV

Expectations are measured by sample mean and is the central tendency of the sample.

Variance

$$\mathrm{Var}(X) = \mathbf{E}\left[(X-\mu)^2\right]$$

Variance is related dispersion or spread of the sample.

Covariance and Correlation

$$\mathrm{cov}_{XY} = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

$$\rho_{XY} = \sigma_{XY} / (\sigma_X \sigma_Y)$$

Covariance is related to association.

Expectation, Variance, Covariance

These are very important quantities

They are heavily used in data analytics and AI

Expectation and Variance summarizes data – descriptive analytics

They provide central tendency and spread of data, also outliers

Covariance analyzes association between factors

Heavily used in diagnostic and predictive analytics

Association

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3

Mean(water)=76.6, Mean(QI)=4.3.

Var(water)=16.6, Var(QI)=3.1

Cov(water,QI) = -5.56, Corr = -0.93

Mean(Material)=15, Mean(QI)=4.3.

Var(Material)=0, Var(QI)=3.1

Cov(Material,QI) = 0, Corr = NA

Association

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3

Mean(water)=76.6, Mean(QI)=4.3.

Var(water)=16.6, Var(QI)=3.1

Cov(water,QI) = -5.56, Corr = -0.93

Mean(Fragrance)=8.3, Mean(QI)=4.3.

Var(Fragrance)=16.6, Var(QI)=3.1

Cov(Fragrance,QI) = 5.56, Corr = 0.93

Association

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3

Mean(water)=76.6, Mean(QI)=4.3.

Var(water)=16.6, Var(QI)=3.1

Cov(water,QI) = -5.56, Corr = -0.93

Mean(Time)=19.3, Mean(QI)=4.3.

Var(Time)=1.1, Var(QI)=3.1

Cov(Time,QI) = 0.55, Corr = 0.36

Diagnostic Analytics

Covariance and correlation provides diagnostic analytics

As Material is not correlated to QI, we can remove in preprocessing

Predictive Analytics

How can we do predictive analytics ?

The question when QI will be bad (>4) ? Can we forecast ?

P(QI is bad) = ?

P(QI is bad | features) = ?

Conditional Probability

$P(\, QI > 4 \mid Water > 75) \;= ?$

What is the probability that QI is bad given the fact that water is more than 75 ?

$P(A \mid B)$ is a general form, B is in condition that means B has happened, we know B, but do not know A

$P(B \mid A)$ we know A but do not know B

Rules to Remember

$P(A \mid B)\,P(B) = P(A, B)$ – product rule

$P(A, B)$ means probability of A and B

$P(A \mid B) = [\, P(B \mid A)\, P(A) \,] \,/\, [\, P(B) \,]$ – Bayes rule

Predictive Analytics

$P(QI > 4.5 \mid Water < 78) = ?$

We can not find this.

$P(QI > 4.5 \mid Water < 78) = P(QI>4.5, Water<78)/P(Water<78)$

Water	Quality Index
75	5
70	7
80	2
80	4
75	5
80	3

$P(QI>4.5, Water<78) = 3/6$

$P(Water<78)=3/6$

$P(QI > 4.5 \mid Water < 78) = 1$

Predictive Analytics

$$P(QI < 4.5 \mid Water < 78) = ?$$

$$P(QI < 4.5 \mid Water < 78) = P(QI<4.5, Water<78)/P(Water<78)$$

Water	Quality Index
75	5
70	7
80	2
80	4
75	5
80	3

$$P(QI<4.5, Water<78) = 0/6$$

$$P(Water<78)=3/6$$

$$P(QI < 4.5 \mid Water < 78) = 0$$

Predictive Analytics

Water	Quality Index
75	5
70	7
80	2
80	4
75	5
80	3
77	?
82	?

Predictive vs Diagnostic

Water	Quality Index
75	5
70	7
80	2
80	4
75	5
80	3

P(QI | Water) vs P(Water | QI) ?