WEEK 1 - LECTURE 1

A brief introduction to machine learning

What is Machine Learning?

An agent (Machine / Non Machine) is said to learn from experience with respect to some class of tasks, and a performance measure P, if [the learner's] performance at task in the class, as measured by P, improves with experience.

The three main components are

(i) Defining class of task
(ii) Defining performance measure
(iii) Well defined experience

For eg.

Answering questions in exam = Class of task
~~Class of task~~ = No. of marks you get = Performance Measure
~~Performance Measure~~ Writing more exams = Experience

ML Paradigms

1 Supervised learning

Learn an input and output map

- Classification : categorical output
- Regression : continuous output

2 Unsupervised Learning
Discover patterns in the data
- Clustering : cohesive grouping
- Association : frequent cooccurence

3 Reinforcement Learning
Learning control

Performance measure of different machine learing tasks.

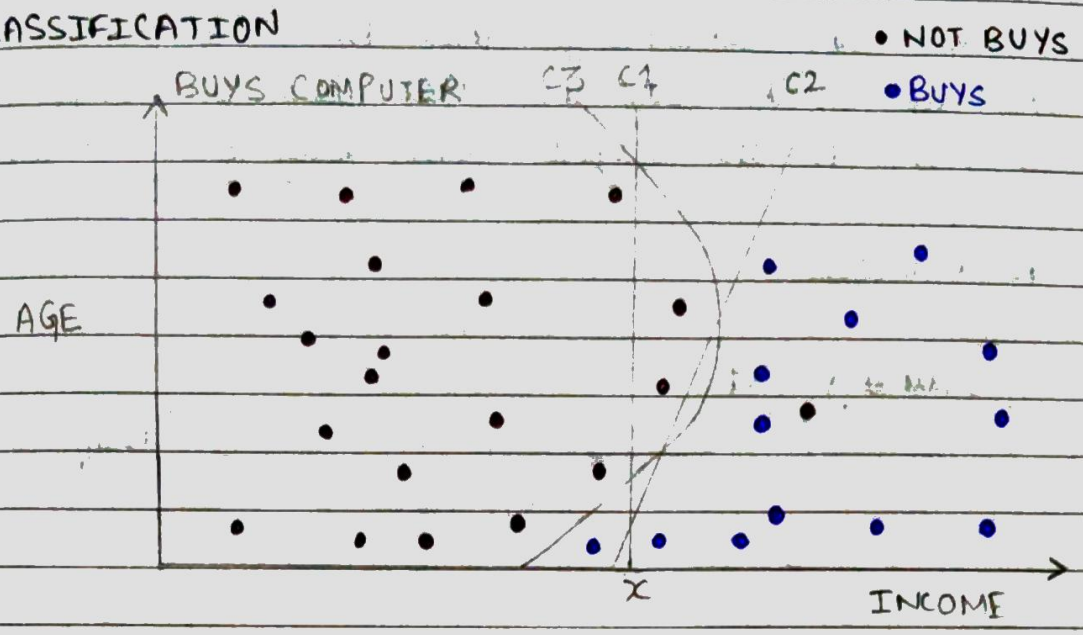
	Task	Performance Measure
1	Classification	Error
2	Regression	Error
3	Clustering	Scatter / Purity
4	Associations	Support / Confidence
5	Reinforcement Learning	Cost / Reward

Challenges faced
- How good is a model?
- How do I choose a model?
- Do I have enough data?
- Is the data of sufficient quality?
  - Errors in data . Eg: Age = 225
  - Missing values
- How confident can I be of the results
- Am I describing the data correctly?
  - How should I represent Age? As a no., or as young, middle age, old?

WEEK 1 - LECTURE 2

Supervised Learning

CLASSIFICATION



Above is the customer data who buys or not buys the computer.
The goal is to come up with a function that takes age and income as input parameters and tells us whether customer will buy the computer or not.

In case of C1, function would be:
If income is less than x, then person will not buy the computer and if it is greater than x, then person will buy the computer.

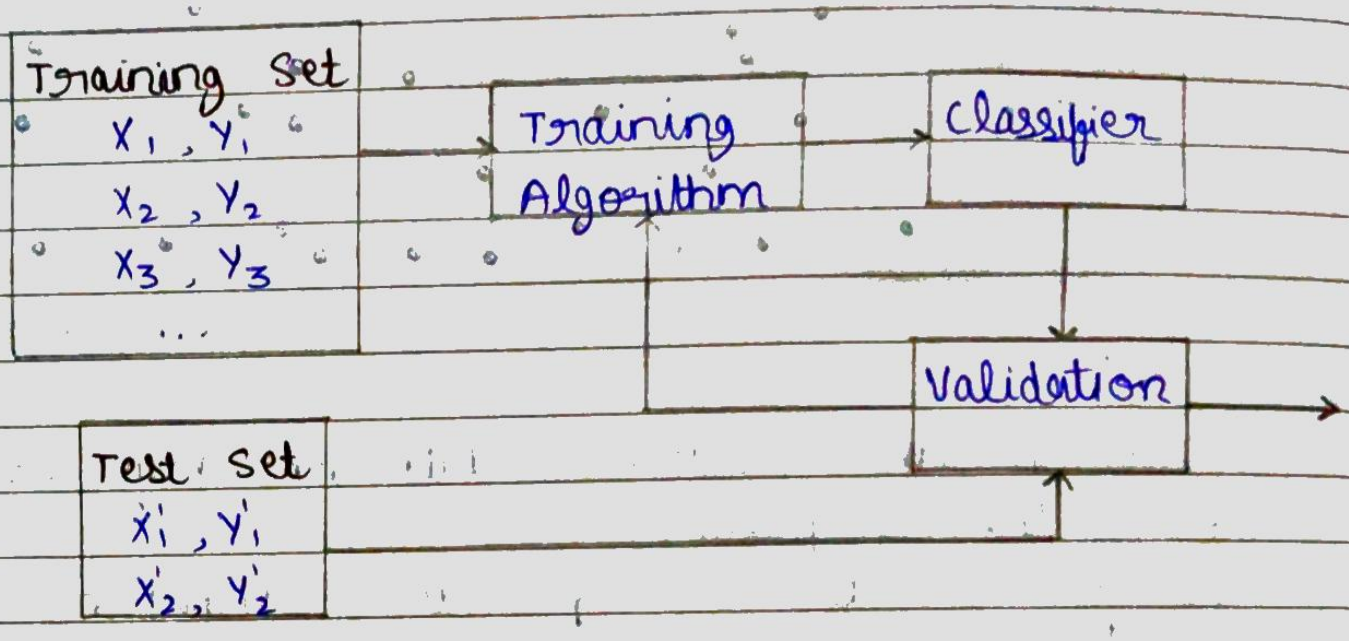
In case of C2, the performance measure improves. The older the person, income threshold is more and younger the person, income threshold is less.

In case of C3, performance measure improves even

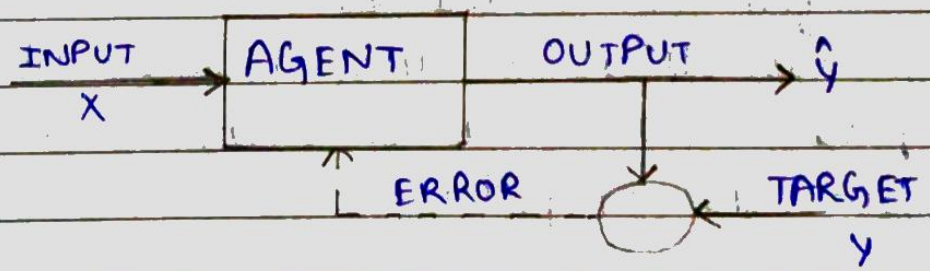
further but at the cost of having a more complex classifier.

So, we can generalize this by saying everything to the left of the line will not buy a computer and everyone to the right will buy the computer.

The Process



What happens inside a training algorithm?

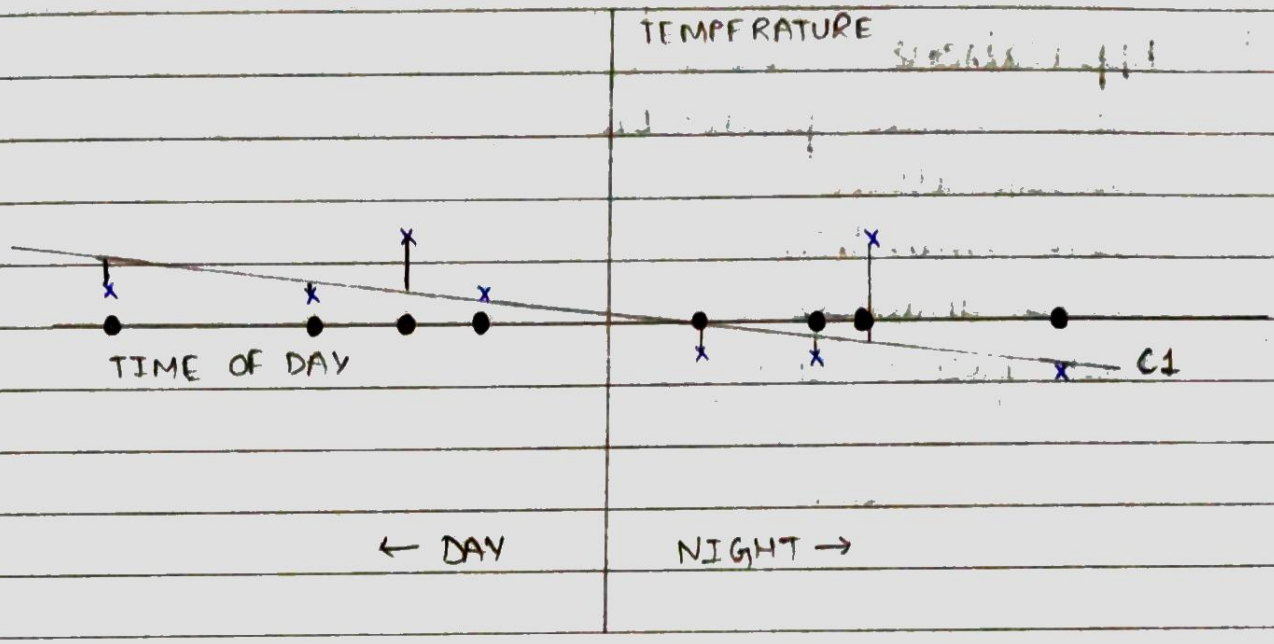


Applications

- Credit card fraud detection
- Sentiment Analysis
- Churn Prediction
- Medical diagnoses

REGRESSION

The output is a continuous value.



The black dots are the input (time of day) and the blue crosses are the output (temperature). The output are continuous values.

C1 is the best fit line as it is nearest to all the blue crosses.

Linear Regression

- Minimize sum squared error
- With sufficient data simple enough
- With many dimensions, challenge is to avoid over fitting
  - Regularization
- Higher order functions?
  - Basic transformations
  - Eg. $(x_1, x_2) \rightarrow (x_1^2, x_2^2, x_1 x_2, x_1, x_2)$

Applications

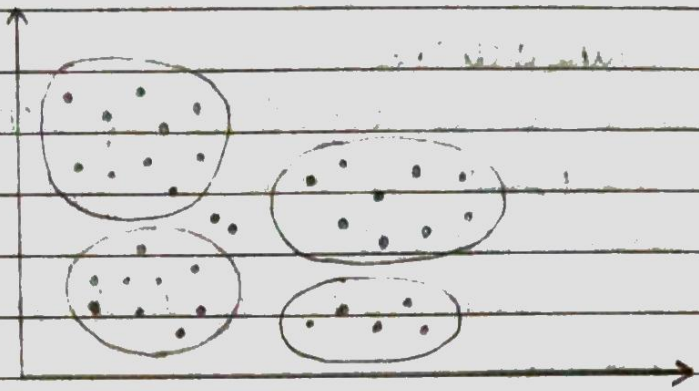
- Time series prediction
- Classification
- Data reduction
- Trend analysis
- Risk factor analysis

WEEK 1 - LECTURE 3

Unsupervised Learning

CLUSTERING

In clustering the goal is to find groups of cohesive data points.



In the above data set, we have found 4 clusters. Here the bias is the shape of cluster. Here the shape is ellipse.
All the points need not be the part of clusters.

Applications

- Customer Data (Discover classes of customers)
- Image pixels (Discover regions)
- Words (Synonyms)
- Documents (Topics)

ASSOCIATION RULE MINING

- Mining frequent patterns and rules
- Association rules: conditional dependencies
- Two stages ⇒

1. Find frequent patterns
2. Derive associations ($A \rightarrow B$) from frequent patterns

- Find patterns in
  - Sequences (time series data, fault analysis)
  - Transactions (market basket data)
  - Graphs (social network analysis)

Mining Transactions

- Transaction is a collection of items bought together
  - A (sub) set of items is c/d an itemset
- Find frequent itemsets
- Itemset A $\Rightarrow$ Itemset B, if both A and AUB are frequent itemsets.

Applications

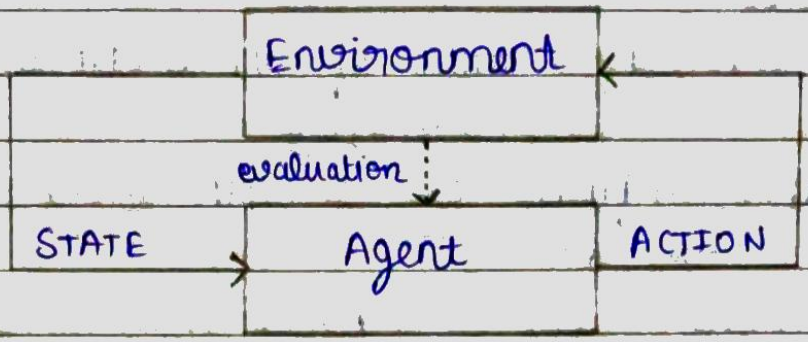
- Predicting co-occurrence
- Market Basket Analysis
- Time Series Analysis (Trigger Events)

WEEK 1 : LECTURE 4

Reinforcement Learning

The type of learning where we learn to control the system through trial and error and the minimal feedback is called reinforcement learning

RL Framework



- Learn from close interaction
- Stochastic environment
- Noisy delayed scalar evaluation
- Learn a policy (Maximize a measure of long term performance)

Applications

- Game playing (Backgammon - world's best player)
- Autonomous agents (Robot Navigation)
- Adaptive control (Helicopter pilot)
- Combinatorial Optimization (VLSI Placement)
- Intelligent Tutoring Systems

WEEK 1 TUTORIAL 1

Probability Basics (1)

Sample Space (Ω)

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

Individual elements are denoted by ω and are termed elementary outcomes.

Examples

- (Finite) A single roll of an ordinary die
  $\Omega = \{1,2,3,4,5,6\}$
- (Countable) Infinite no of coin tosses.
  $\Omega = \{H,T\}^{\infty}$
- (Uncountable) Speed of vehicle measured with infinite precision. $\Omega = R$

Event

Any collection of possible outcomes of an experiment i.e. any subset of Ω

Example

On rolling a die, outcome even (Event $E = \{2,4,6\}$) or odd (Event $O = \{1,3,5\}$)

Set Theory Notations

$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B$

$A = B \Leftrightarrow A \subset B$ and $B \subset A$

$A \cup B = \{x : x \in A \text{ or } x \in B\}$

$A \cap B = \{x : x \in A \text{ and } x \in B\}$

$A^c = \{x : x \notin A\}$

Properties of Set operations

1 Commutativity

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

2 Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

3 Distributivity

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

4 DeMorgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Disjoint Events

Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \phi$

A sequence of events $A_1, A_2, A_3, \cdots$ are pair-wise disjoint if $A_i \cap A_j = \phi$ for all $i \neq j$

Partition

If $A_1, A_2, \cdots$ are pair-wise disjoint and $\bigcup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \cdots$ forms a partition of Ω

Sigma Algebra

Given a sample space Ω, a σ-algebra is a collection $\mathcal{F}$ of subsets of Ω, with the following properties

(a) $\phi \in \mathcal{F}$

(b) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$

(c) If $A_i \in \mathcal{F}$ for every $i \in \mathbb{N}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

A set A that belongs to $\mathcal{F}$ is called $\mathcal{F}$-measurable set (event)

Example: Consider $\Omega = \{1,2,3\}$

$\mathcal{F}_1 = \{\phi, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$

$\mathcal{F}_2 = \{\phi, \{1,2,3\}\}$

For any Ω (countable or uncountable) 2^{Ω} is always a σ-algebra

For eg., for $\Omega = \{H,T\}$, a feasible σ-algebra is the power set, i.e. $\mathcal{F} = \{\phi, \{H\}, \{T\}, \{H,T\}\}$

However, if Ω is uncountable, then probabilities cannot be assigned to every subset of 2^{Ω}

Probability Measure and Probability Space

A probability measure P on $(\Omega, \mathcal{F})$ is a function $P: \mathcal{F} \to [0,1]$ satisfying

(a) $P(\phi) = 0 \quad P(\Omega) = 1$

(b) If $A_1, A_2 \cdots$ is a collection of pair-wise disjoint members of $\mathcal{F}$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

The triple $(\Omega, \mathcal{F}, P)$ comprising a set Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω and a probability measure P on $(\Omega, \mathcal{F})$ is called a probability space.

Consider a simple experiment of rolling an ordinary die in which we want to identify whether the outcome results in a prime no or not

$\Omega = \{1,2,3,4,5,6\}$

$\mathcal{F} = \{\phi, \{1,4,6\}, \{2,3,5\}, \{1,2,3,4,5,6\}\}$

$P : \mathcal{F} \to [0,1]$

$P(\phi) = 0$

$P(\{1,4,6\}) = 0.5$

$P(\{2,3,5\}) = 0.5$

$P(\Omega) = 1$

Bonferroni's Inequality

$$P(A \cap B) \geq P(A) + P(B) - 1$$

- It gives a lower bound on the intersection probability which is useful when probability is hard to calculate
- Only useful if the probabilities of individual events are sufficiently large

General form: $P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$

Boole's Inequality

$$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$, for any sets $A_1, A_2 \ldots$

Gives a useful upper bound for the probability of the union of events.

Conditional Probability

Given two events A and B, if $P(B) > 0$, then the conditional probability that A occurs given that B occurs is defined to be

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

Essentially, since event B has occured, it becomes the new sample space.

Conditional probabilities are useful when reasoning in the sense that once we have observed some event, our beleifs or predictions of related events can be updated/improved.

Q A fair coin is tossed twice. What is the probability that both tosses result in heads given that at least one of the tosses resulted in a heads?

$\Omega = \{HH, HT, TH, TT\}$

$P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}$

~~P(HH)~~ $P(HH \mid \text{at least one toss heads})$

$= P(HH \mid HT \cup TH \cup HH)$

$= \frac{P(HH \cap (HT \cup TH \cup HH))}{P(HT \cup TH \cup HH)}$

$= \frac{P(HH)}{P(HT \cup TH \cup HH)} = \boxed{\frac{1}{3}}$

Bayes' Rule

We have

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

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

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{(Bayes' Rule)}$$

It is important as it allows us to compute the conditional probability $P(A|B)$ from the 'inverse' conditional probability $P(B|A)$

Independent Events

Two events A and B, are said to be independent if

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

More generally, a family $A_i : i \in I$ is called independent if,

$$P(\cap_{i \in J} A_i) = \Pi_{i \in J} P(A_i)$$

for all finite subsets J of I

Conditional Independence

Let A, B and C be three events with $P(C) > 0$. The events A and B are called conditionally independent given C if

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

or equivalently,

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

WEEK 1 : TUTORIAL 1

Probability Basics (2)

Random variable

A random variable is a function $X : \Omega \to R$ i.e. it is a function from the sample space to the real numbers.

Example:

- Sum of outcomes on rolling 3 dice
- No. of heads observed when tossing a fair coin 3 times

Induced Probability Function

Consider the previous example of tossing a fair coin 3 times. Let X be the no of heads obtained in the three tosses. Enumerating the elementary outcomes, we observe the value of X as

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

Instead of using the probability measure defined on the elementary outcomes or events, we should ideally like to measure the probability of the random variable taking on values in its range

X	0	1	2	3
$P_X(X = x)$	1/8	3/8	3/8	1/8

Let $\Omega = \{\omega_1, \omega_2, \dots\}$ be a sample space and P be a probability measure (function)

Let X be a random variable with range $X = \{x_1, x_2 \cdots x_m\}$

We define the induced probability function P_x on X as

$$P_x(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$

Cumulative Distribution Function

The cdf of a random variable X, denoted by $F_x(x)$ is defined by

$$F_x(x) = P_x(X \leq x), \text{ for all } x$$

Example

x	$(-\infty, 0]$	$(-\infty, 1]$	$(-\infty, 2]$	$(-\infty, 3]$	$(-\infty, \infty)$
$F_x(x)$	1/8	1/2	7/8	1	1

Properties of cdf

A function $F_x(x)$ is a cdf iff the following three conditions hold:

- (Monotonicity) If $x \leq y$, then $F_x(x) \leq F_x(y)$
- (Limiting Values) $\lim_{x \to -\infty} F_x(x) = 0$ and $\lim_{x \to \infty} F_x(x) = 1$
- (Right-continuity) For every x, we have $\lim_{y \downarrow x} F_x(y) = F_x(x)$

Continuous and Discrete Random Variable

Random variable X is continuous if $F_x(x)$ is a continuous function of x

Random variable X is discrete if $F_x(x)$ is a step function of x

Probability Mass Function (pmf)

The pmf of a discrete random variable X is given by

$$f_X(x) = P(X = x), \text{ for all } x$$

Example:

For a geometric random variable X with parameter p,

$$f_X(x) = \begin{cases} (1-p)^{x-1} p & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Properties:

- $f_X(x) \geq 0$, for all x
- $\sum_x f_X(x) = 1$

Probability Density Function (pdf)

The pdf of a continuous random variable is the function $f_X(x)$ which satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \text{ for all } x$$

Properties:

- $f_X(x) \geq 0$, for all x
- $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$

Expectation

The expected value or mean of a r.v. X, denoted by E[X], is given by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx \quad \text{(continuous RV)}$$

$$E[X] = \sum_{x : P(x)>0} x f_X(x)$$

$$= \sum_{x : P(x)>0} x\, P(X=x) \quad \text{(discrete RV)}$$

Properties of Expectation

Let X be a r.v. and let a, b, c be constants. Then, for functions $g_1(x)$ and $g_2(x)$ whose expectations exists

- $E(a g_1(X) + b g_2(X) + c) = a E g_1(X) + b E g_2(X) + c$
- If $g_1(X) \geq 0$ for all X, then $E g_1(X) \geq 0$
- If $g_1(X) \geq g_2(X)$ for all X, then $E g_1(X) \geq E g_2(X)$
- If $a \leq g_1(X) \leq b$ for all X, then $a \leq E g_1(X) \leq b$

Moments

For each integer n, the n^{th} moment of X is

$$\mu_n' = E X^n$$

The n^{th} central moment of X is

$$\mu_n = E(X-\mu)^n$$

Variance

The variance of a r.v. X is its second central moment

$$\text{Var } X = E(X-\mu)^2 = E(X-EX)^2 = EX^2 - (EX)^2$$

The +ve square root of $\text{Var } X$ is the standard deviation of X

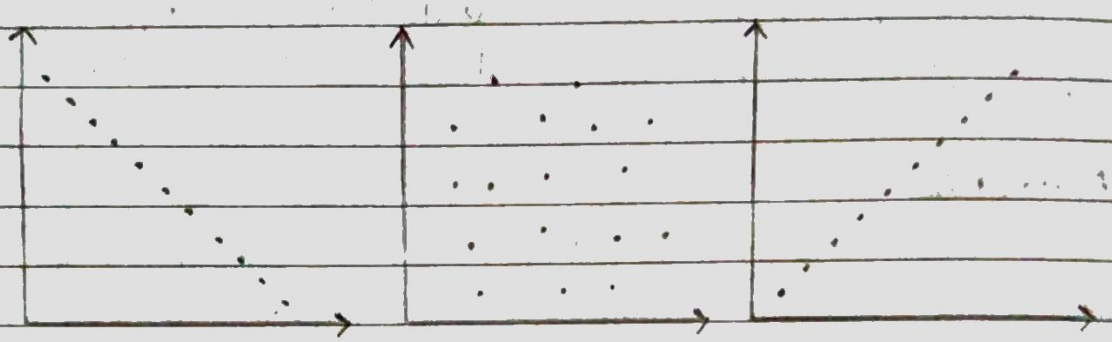
Note: $\text{Var}(aX+b) = a^2 \text{Var}\,X$
where a, b are constants

Covariance

The covariance of two r.v. X and Y is

$$\text{cov}(X,Y) = E[(X-EX)(Y-EY)]$$

It is a measure of how much two r.v. change together.



Large -ve c.v. Neare zero c.v. Large +ve c.v.

Correlation

The correlation of two r.v. X and Y is

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

Note:

- For correlation to be defined, individual variances must be non-zero and finite.
- $\rho(X,Y)$ lies b/w -1 and $+1$

Probability Distributions

Consider two variables X and Y, and suppose we know the corresponding pmf f_X and f_Y

Can we answer the following question:

$$P(X = x \text{ and } Y = y) = ?$$

Joint Distributions

To capture the properties of two r.v X and Y, we use the joint PMF

$f_{X,Y} : R^2 \to [0, 1]$, defined by
$f_{X,Y}(x, y) = P(X = x, Y = y)$

Marginal Distributions

Suppose we have the joint PMF

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

From this joint PMF, we can obtain the PMF's of the two r.v.

$f_X = \sum_y f_{X,Y}(x, y)$ (marginal PMF of R.V. X)
$f_Y = \sum_x f_{X,Y}(x, y)$ (marginal PMF of R.V. Y)

Conditional Distributions

Like joint distributions, we can also consider conditional distributions

$$f_{X|Y}(x|y) = P(X = x | Y = y)$$

Using conditional probability definition, we have

$$f_{X|Y}(x|y) = f_{X,Y}(x,y) / f_Y(y)$$

Note that the above conditional probability is undefined if $f_Y(y) = 0$

Bernoulli Distribution

Consider a R.V. X taking one of two possible values (either 0 or 1). Let the PMF of X be given by

$$f_X(0) = P(X=0) = 1-p \qquad (0 \le p \le 1)$$
$$f_X(1) = P(X=1) = p$$

This describes a Bernoulli distribution

$$E[X] = p \qquad Var[X] = p(1-p)$$

Binomial Distribution

Consider the situation where we perform n independent Bernoulli trials where

- probability of success (for each trial) = p
- probability of failure = 1-p

Let X be the no. of success in the n trials, then we have

$$P(X = x|n,p) = \binom{n}{x} p^x (1-p)^{n-x} \quad ; 0 \le x \le n$$

$$E[X] = np$$
$$Var(X) = np(1-p)$$

Geometric Distribution

Suppose we perform a series of independent Bernoulli trials, each with a probability p of success. Let x represent the no of trials before the first success, then we have

$$P(X = x|p) = (1-p)^{x-1} p \qquad x = 1, 2, 3 \ldots$$

$$E[X] = \frac{1}{p} \qquad Var(X) = \frac{(1-p)}{p^2}$$

Uniform Distribution

A continuous R.V X is said to be uniformly distributed on an interval [a, b] if its PDF is given by

$$f_X(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

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

Normal Distribution

A continuous RV X is said to be normally distributed with parameters μ and σ^2 if the density of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

Importance of Normal Distribution

Roughly, the central limit theorem states that the distribution of the sum (or average) of a large no of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

Multivariate Normal Distribution

$$N(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

where,

- μ is the D-dimensional mean vector
- Σ is the $D \times D$ covariance matrix
- $|\Sigma|$ is the determinant of the covariance matrix.

Beta Distribution

The pdf of the beta distribution in the range $[0, 1]$, with shape parameters α, β is given by

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where the gamma function is an extension of the factorial function

$$E[x] = \frac{\alpha}{(\alpha+\beta)} \qquad Var(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$