## Describe the general Machine Learning Process? (5m)

Machine learning is a process which **improves** over **task** T, with **respect** to **performance** measure P, **based** on **experience** E. For example, in a speech recognition system, task T is recognising words, with performance P as the accuracy of recognition, and experience E as the voice sample of each word.

Usually, the following steps are followed for an end to end workflow of a machine learning project.

1. Data Collection
2. Data Preparation
3. Choose a Model
4. Train the Model
5. Evaluate the Model
6. Parameter Tuning
7. Make Predictions

## Outline five questions that should be asked when choosing an appropriate Machine Learning technique for a problem. (10m)

1. Type of problem
2. Size of dataset
3. Dimensions of the training set
4. Accuracy expected
5. Computational constraints

## “Machine Learning and A.I are changing the world today”. Discuss (10m)

### Weather predictions and Environmental Protection

The weather predictions were tedious task which was hard and the results was not accurate as desired. The growth of Machine learning techniques has reduced the complexity of mathematical modelling and solutions. The feature engineering for weather models was inaccurate as the data points where not discrete. But the growth of the Artificial Intelligence techniques such as deep learning and unsupervised learning has helped the scientist to derive the factors which may affect the weather. This advancements in weather has also started to help in solving environmental issues using which can derive the causes of pollutions and how effectively we can address them. Evaluating the solutions cost to derive at optimal policy decisions was a major breakthrough in the greener environment movements.

### Undertake Dangerous Tasks

The risk of losing life at high-risk environments like mines, dredges and refineries is greatly reduced by the implementation of autonomous robots which majorly uses the reinforcement learning techniques of Artificial Intelligence. The advancements in the Convolutional Neural Networks used for image processing has been a great factor in the success of robots. This is also supplemented by GPU technologies which is optimal for the CNN calculations.

### Banking and Investments

Previously, the banking and investments sectors hugely depended on highly skilled accounts and even intuitions for making decisions. This made the process like loan approvals, investment choices as biased decisions, even at times a gambling game. But the integration of ML and AI systems made the decisions more data centric.

### Smart Home and IoT

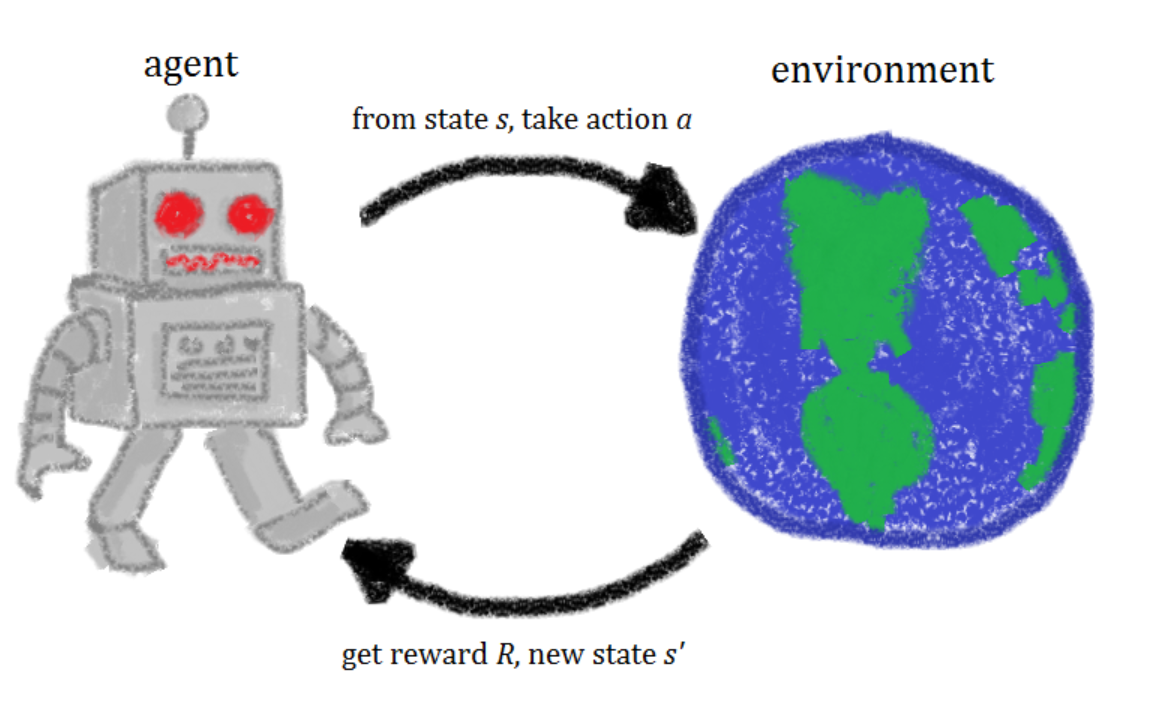
The growth of IoT has created a completely new arena for the applications for AI and ML. This is an example, how the technology can change the life in very atomic element of society, the families. The integration of AI into the day-to-day operations has reduced the human efforts in maintaining a house which reduced the mundane task for people. For example, personal assistants like Alexa, Cortana, Siri etc have created a completely new capabilities for AI systems in home automation. The smart home appliances have reduced power consumptions and optimized the customer experiences using AI tech.

### Self-Driving Cars and Automated Transportation

This is one of the most heavily invested and controversial areas of AI application. Nevertheless, the advancements in this technology is breath-taking and alarming at the same time. The auto-pilot systems have been integrated into the aeroplanes which are using a data centric approach of weather predictions, turbulence control extra. The self-driving cars are tested on road now, which are almost as good as human drivers. This raises same questions as the 1980s when computers started taking over computational task which caused of a panic of losing jobs. But the answer is the same, every era requires people with the skills of the next era.

## Compare and contrast Q Learning with the way humans learn. (10m)

Ans: Q-learning is an off policy reinforcement learning algorithm that seeks to find the best action to take given the current state. It’s considered off-policy because the q-learning function learns from actions that are outside the current policy, like taking random actions, and therefore a policy isn’t needed. More specifically, q-learning seeks to learn a policy that maximizes the total reward.



A rough **framework of reinforcement learning is shown in the above figure.**

Throughout our lives, we perform a number of actions to pursue our dreams. Some of them bring us good rewards and others do not. Along the way, we keep exploring different paths and try to figure out which action might lead to better rewards. We work hard towards our dreams utilizing the feedback we get based on our actions to improve our strategies. They help us determine how close we are to achieving our goals. Our mental states change continuously to representing this closeness.

In that description of how we pursue our goals in daily life, we framed for ourselves a representative analogy of reinforcement learning. Let me summarize the above example reformatting the main points of interest.

Our reality contains environments in which we perform numerous actions. Sometimes we get good or positive rewards for some of these actions in order to achieve goals. During the entire course of life, our mental and physical states evolve. We strengthen our actions in order to get as many rewards as possible.

The key entities of interest are - environment, action, reward and state. Let’s give ourselves a name as well - we are agents in this whole game of life. This whole paradigm of exploring our lives and learning from it through actions, rewards and states establishes the foundation of reinforcement learning. Beyond beating humans in game-playing, there are some other marvellous use cases of reinforcement learning:

* [Optimizing business processes](https://www.technologyreview.com/s/601045/this-factory-robot-learns-a-new-job-overnight/)
* [Minimizing energy costs](https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/)
* [Maximizing revenue shares of a company](https://arxiv.org/pdf/1803.09967) and lots more..

## Discuss the factors that have led to the growth of Machine Learning in the last 10 years. Illustrate your answer with examples and application areas. (10m)

Same as Question 8

## Describe the concept of learning in an ML system in relation to T, P and E? (5m)

Learning is a process which enables a system to perform a task, by using the past experience or actively trying to perform the same. With respect to this, the machine learning algorithm is a system which can automatically improve their performance without being explicitly programming by feeding data about the problem arena. In other words, machine learning is defined as a process which improves over task T, with respect to performance measure P, based on experience E. For example, in a speech recognition system, task T is recognising words, with performance P as accuracy of recognition, and experience E as the voice sample of each words.

## "With the aid of examples and diagrams explain what the difference is between Supervised and Unsupervised Machine Learning (10m)"

### Supervised:

In the supervised learning systems, the previous decisions, categories and relative metrics, are used in order to achieve a model in the form of automated target by providing a different set of values. In other words, these set of values consisting of an input object and a desired or predicted output value.

For e.g.: The training dataset of height, weight and primary sport of male student in a class as shown below

|  |  |  |
| --- | --- | --- |
| Height | Weight | Sport |
| 150 | 60 | Cricket |
| 152 | 65 | Cricket |
| ….. | …. | …. |
| 170 | 70 | Basketball |

We can see a plot of this data,

---------------Plot----------------------

As we can observe, there are cluster-based on height and weight which roughly represent their primary sport.

In this example, let’s assume the primary sport is predicted based on each Student’s height and weight in the class which is called labelled data. This labelled data can be used to create a model, which could predicate a new Student’s primary sport by providing height and weight. In mathematical terms, similar to a function f(x), a Model M (h, w) is created based on the given data which will output a category based on the variables h=height and w=weight.

### Unsupervised:

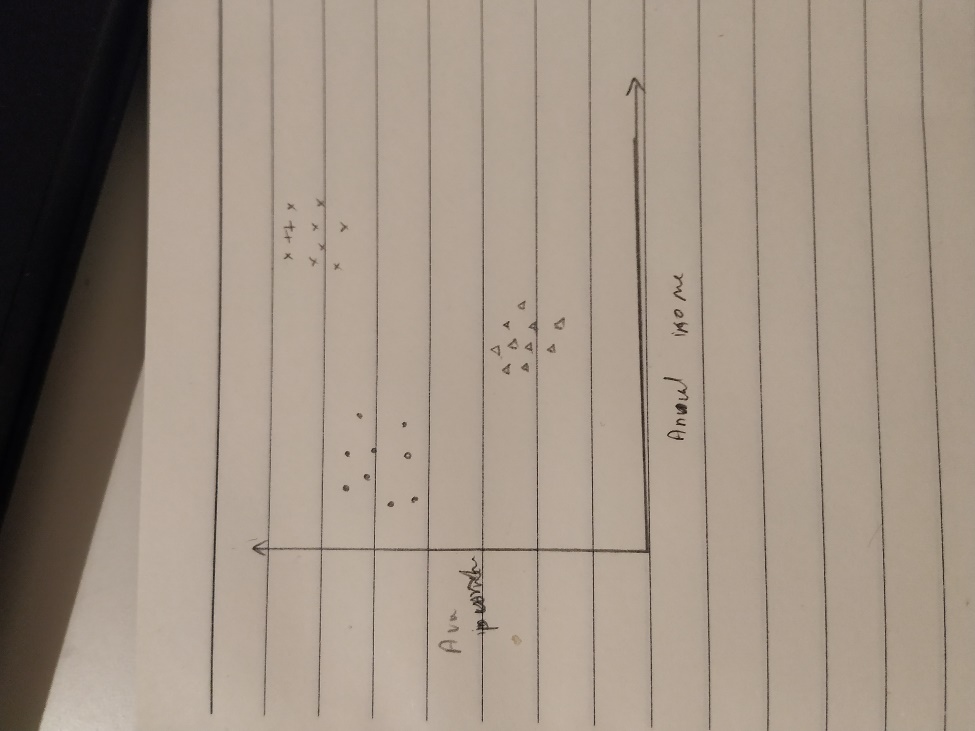
This type of machine learning is used to explore the datasets which have been collected through various sources. This allows for exploring new dimensions in data, unventured correlative paths between variables and clustering of items. One of the main use-cases of this ML type is to find out new categories or to derive new stereotypes.

This method has allowed several companies to re-strategize their segmentation of markets which in turn improved product alignment of customers.

A very good example of this is the customer categorisation of a shopping mall. Below is a sample data of customers,

|  |  |  |
| --- | --- | --- |
| Age | Annual Income | Average Per month Purchase |
| 25 | 20000 | 150 |
| 27 | 30000 | 200 |
| …. |  |  |
| 30 | 45000 | 300 |

As illustrated in the table, the data do not have a predefined segment of customers, rather using the three parameters we have to find new segments. For this example, methods such as K-means can be applied, where an iterative method of finding nearby items by calculating the k-centroids for the defined arbitrary value of k as per the use case. In this example, if we plot the graph for Annual income vs Average per month purchase. We can see some cluster as per the graph below.



As illustrated in the plot, there are 3 clusters emerged, which can be clustered using methods like K-means. This data can be further used for segregation of the display racks, discounts, offers, etc.

Steps:

1. Choose an arbitrary value K.
2. Choose K number of random points.
3. Iterate through each datapoints to find the distance from the selected K-centroids.
4. Assign each point to the nearest centroid which will create initial K clusters.
5. Calculate the new centroid of each cluster.
6. Then re-cluster based on these new centroids.
7. Repeat till the clusters do not change.

Features:

* Works well with Compact cluster
* But sensitive to outliers
* Only numerical values can be calculated.

## "Discuss the reasons for the explosion in the use of Machine Learning in the last five years. (10m)"

There are several reasons for the growth of machine learning in the last 5 years of which some major reasons are discussed below

### Increased computational power and efficiency

The consistent issue faced in the field of machine learning is the huge computational resource utilization. But, in the last decade, the efficiency of the computational systems has increased tremendously especially in terms of energy efficiency. The carbon footprint by the server farms has been reduced which have brought down the cost for the processing of data and training the machine learning models. The high popularity of the cloud-based computational resource pooling has given an optional on-demand scaling which is highly useful for the training purposes as the over-head of maintaining highly configuration dedicated servers is not required.

### Introduction of GPU’s

Another key reason for increase in ML and AI, especially deep learning is the availability of cheap commercial grades GPU’s. The GPU’s are specially designed processors for parallel processing for applications like games where high amount of video processing is done in parallel. Coincidently that’s what the deep neural networks also needs, high amount of parallel processing for matrix operations. This has revolutionized the acceptance of deep learning solutions as the computing power are cheaper and available at much more convenience.

### High maturity of the cloud-native machine learning platforms

Most of the prominent cloud service providers like AWS and Azure have cloud-native machine learning platforms has removed the requirement of the environment setup which gave data analytics more independency in terms of technical know-how. This freedom has created more concentrated efforts on developing new concepts. Furthermore, this has also helped in easy experimentation of new algorithms and hypothesis. Additionally, these platforms have provided the opportunity for new people to enter the field as the complexity of the field has reduced which creates a more welcoming environment in previously unapplied areas of ML.

### Growth of social media

The growth of the internet and social media have created a great inflow of data. This has given the technology behemoths a new stream of personal data which was inaccessible before; this has led to higher accuracy for models. This also has made a new platform for the advertisement using the machine learning systems. The social media technology companies like Facebook, LinkedIn are now investing heavily on AI technologies which have boosted the machine learning model development targeted for social media platforms.

### Increased demand for customer-focused marketing

The recommender systems are one of the pioneers in implementing real-world applications of machine learning and efficiency of these applications have led to wider acceptance by the audience. Furthermore, this high popularity has oiled more data flow about people which created demand a new set of applications and algorithms to exploit this new information. For example, Spotify which have used AI algorithms to trace a person’s preferences and mood had gained very popularity which is currently having a lot of targeted advertisements based on customer personality.

### Growth of Personal assistants and IoT

One of the crucial turning points in the popularity in ML techniques especially in vision and natural language processing is the spread of personal assistants like Alexa, Cortana, Siri and Google Assistant. When these devices were initially released, accuracy was not of an acceptable norm. But with increased hardware quality and more data samples. In recent years, the acceptability of these devices has sky-rocketed.

## Explain the concept of Reinforcement Learning and describe where it is applicable. (10m)- same as (4)

## "Describe what each of the following means in relation to Machine Learning: (15m)

#### Variance

The variance gives an estimation of how much “diverse “or “spread-out” is the data. It is calculated as the average of the squared difference of the data from the mean of the data set.

Note: Squaring is done for making the difference from the mean positive.

For example, height of students in a class is given as 10,10,11,11,12,13

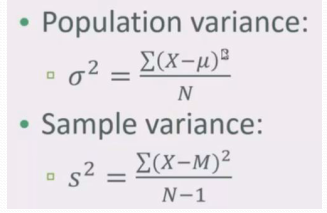
Then mean is calculated as (10+10+11+11+12+13)/6 = 11.5

Sigma square / variance = ((10-11.5) \*\*2 + ….. (13-11) \*\*2) / 6

= ((-1.5) \*\*2 + ….. 2\*\*2) / 6

= (2.25 + .. + 2.25 + 4)/6

=3.25



#### Standard Deviation

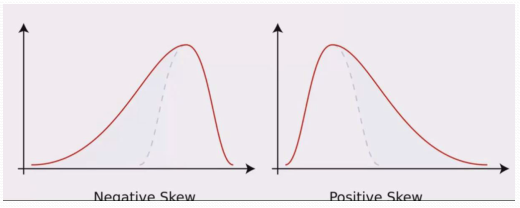
The standard also gives an estimation of how much “diverse “or “spread-out” is the data. It is calculated as the square root of the average of the squared difference of the data from the mean of the data set. It is the square root of Variance; in the example it evaluates to root of 3.5 = 1.79

#### The third moment - Skew

##### Skew

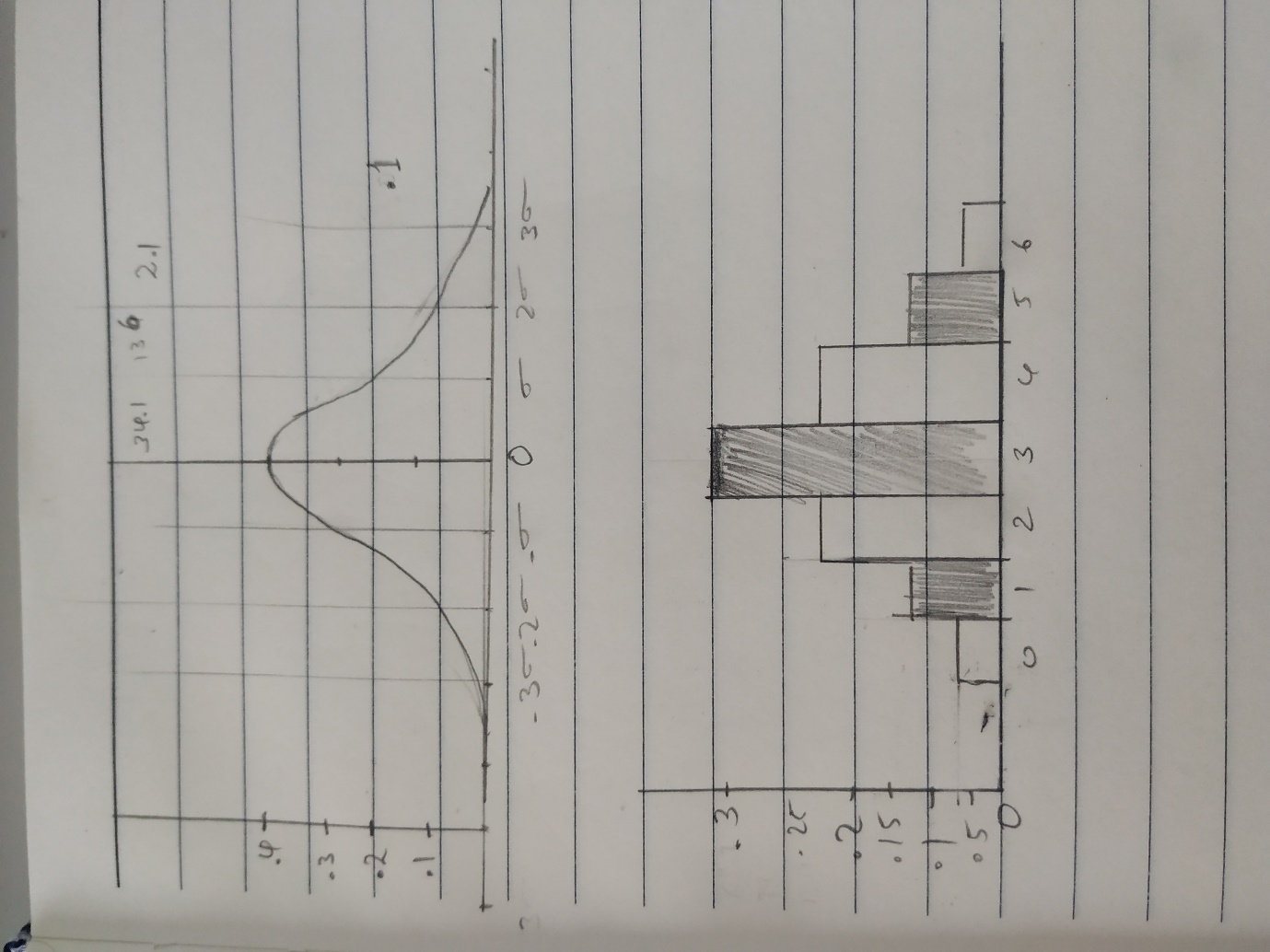
Skew defines the polarity of the data set, if the dataset has more values on the higher spectrum of the data set, it is said to be having a negative skew (left side has a longer tail, fig a). If the values are of the dataset is high in the initial spectrum of the x-axis, it has a positive skew as described in fig b.

**Note:** For a perfect normal distribution, the skew will be 0.



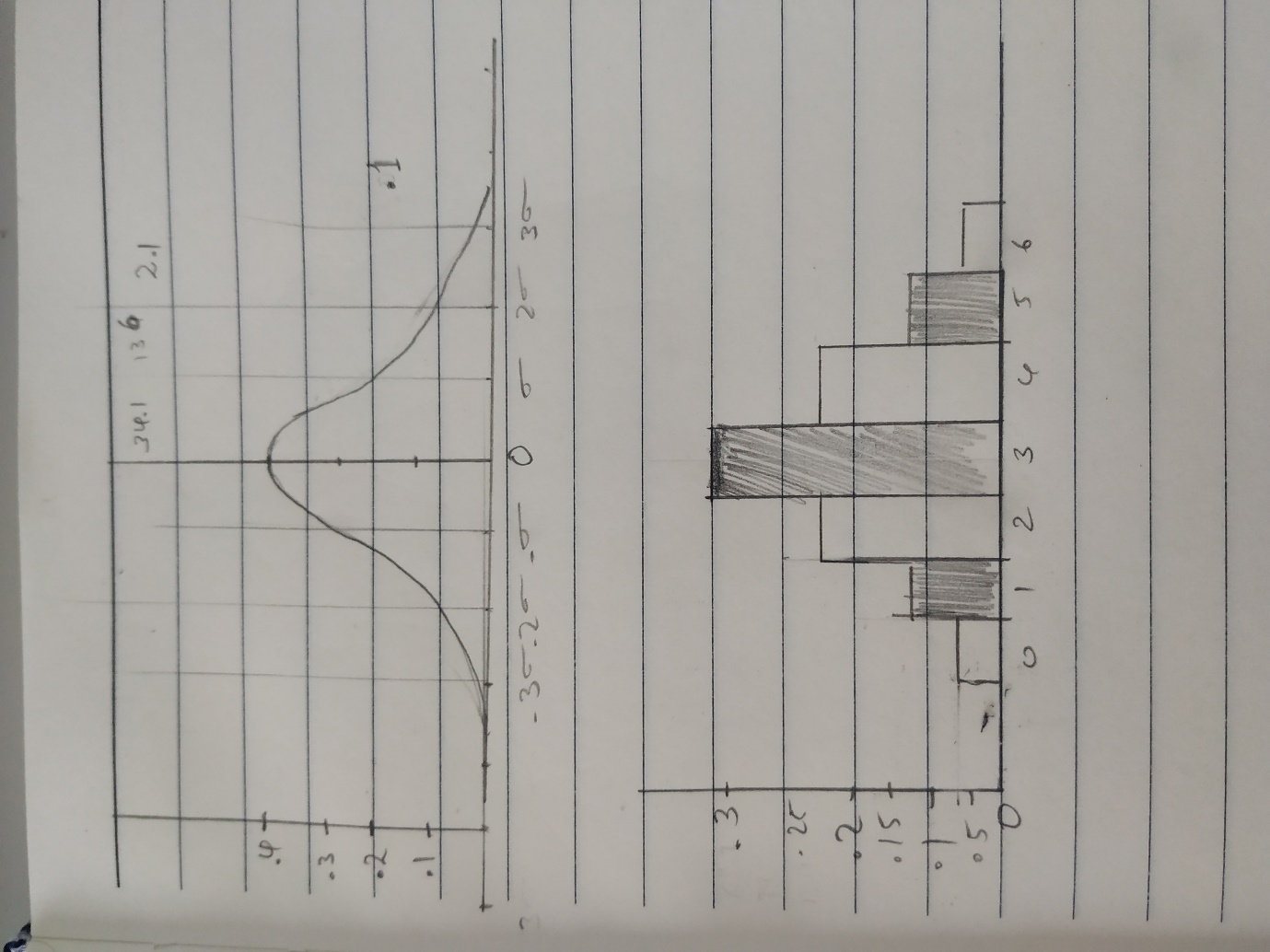
#### PDFs

Probability density function (PDF) is a function that describes the relative likelihood for a random variable to take on a given value w.r.t mean and variance. For a normal distribution, the data will have a have probability of 34.1%+34.1% to be in 1 sigma range of the median, 13.6% +13.6% probability to be in 2 sigma and 1 sigma range, 2.1% + 2.1% between 2 sigma and 3 sigma range and .1% + .1% data outside 3 sigma range.



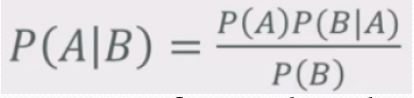
#### PMFs "

Probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value. In below example, the value 3 has a probability of .3 to be present in the data set.



## Describe how you would use Bayes Theorem to create a simple Spam classifier. (10m)

Bayes Theorem is applied for finding the conditional probability of an event A given B using the conditional probability of B given A and independent probabilities of A and B as shown in the equation.



The probability of A given B is the probability of A times the probability of B given A, over the probability of B.

This relation can be used in machine learning because by analysing the training data, we can derive the probabilities A, B and A|B. Hence, using this formula, we can estimate the probability of a new event occurring based on the training data.

For creating a spam classifier. Training data consists of mail text and mail class(ham/spam).

First we need to vectorise the text, i.e. for each mail, find the number of occurrences of each word x, and total number of words. Further, this is used to find the probability of each word in a spam [P(x|spam)] or ham [P(x|ham)] email as well as the absolute probability of a word being present in a mail [P(x)].

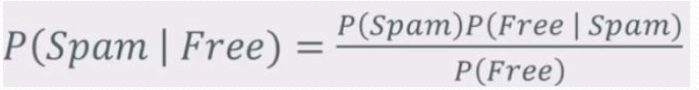
Mathematically the variables are presented as:

P(Spam) = probability of a mail being spam = number spam emails / total number of mails,

P(x) = probability of a mail containing the word “x” = number of emails containing word “x”/ total number of mails,

P(x|Spam) = probability of a mail containing word “x” given the mail is spam = number of spam emails containing words “x” / total number of spam emails,

These three variables have to be calculated for each word in the training data. Once a new mail comes and we can calculate the probability of that email being spam for each word present in that email.



By multiplying the values of P(spam|x), we can conclude if an email is spam based on the confidence level we have set. For example, if the average probability of after calculating the P(spam|x) where represents each word in the new mail. If the probability crosses .5 (arbitrary set value) then we classify it as spam.

## With the aid of examples, differentiate between Numerical, Categorical and Ordinal types of data. (6m)

### Numerical:

The data which have numerical value is called as numerical data. For example, height of a student in cm. In this type, we can do mathematical operations like average of height in a class or logical operations like height of student A > height of student B.

### Categorical:

This type of data does not hold a numerical value, rather a distinguishing feature between the items, for example. Type of building – House, Office, College etc. We cannot do mathematical operations on this. Like, we cannot derive the average type of buildings.

### Ordinal:

There are the data which can some value-based meaning but not direct numerical values. For example, Grades of a student’s -A, B, C and D where A is the highest. In this type of data, we can have comparisons like A is better than B. but we cannot do addition or subtraction.

## How are Standard Deviation and Variance related? Explain with the aid of an example (5m)

The variance gives an estimation of how much “diverse “or “spread-out” is the data. It is calculated as the average of the squared difference of the data from the mean of the data set.

Note: Squaring is done for making the difference from the mean positive.

For example, height of students in a class is given as 10,10,11,11,12,13

Then mean is calculated as (10+10+11+11+12+13)/6 = 11.5

Based on that, the Variance is calculated as following.

Sigma square / variance = ((10-11.5)\*\*2 + ….. (13-11)\*\*2 ) / 6

= ( (-1.5)\*\*2 + ….. 2\*\*2 ) / 6

= (2.25 + + 2.25 + 4 )/6

=3.25

The standard also gives an estimation of how much “diverse “or “spread-out” is the data. It is calculated as the square root of the average of the squared difference of the data from the mean of the data set. It is the square root of Variance; in the example it evaluates to root of 3.5 = 1.87

## Population versus sample datasets. What considerations must be kept in mind when deciding if a Sample dataset is appropriate? (6)

In most cases, it is hard to collect data from the entire population. For e.g. Income of the families in Ireland along with their expenditure data to figure out the spending behaviours. In this example, the amount of data is huge; in these scenarios, extracting a sample dataset would be desirable. But in selecting the sample dataset we have to consider some constraints. Those are:

* 1. The sample should be an almost true representation of the population. For example, when choosing 25000 families from 2 million families in Ireland, it should have approximately right proportion in terms of size of the family, annual income, inherited wealth etc.
  2. The sample size should not be containing same number of outliers as the populations nor discard all of them. For example, if 10 billionaires were there in original population, we should choose only proportional number of billionaires in the sample size, otherwise factors like mean and variance will be greatly affected by these outliers.
  3. True random selection of sampling may seem lucrative but this increases the chances of ending up with sample with very low variance.
  4. When calculating the variance for a sample the squared sum is divide by N-1 where N is the number of observations.

## “Moments are a quantitative measure of the shape of a probability density function”. With the aid of diagrams, describe the four moments (8m)

### Mean

Mean is defined as the average of the data points in a given set. That is, the sum of the samples divided by the number of samples. For example, height of students in a class is given as 10,10,11,11,12,13. Mean will will determine approximate center of the pdf, median will be at exactly the center.

Then mean is calculated as (10+10+11+11+12+13)/6 = 11.5

### Variance

The variance gives an estimation of how much “diverse” or “spread-out” is the data. It is calculated as the average of the squared difference of the data from the mean of the data set.

Note: Squaring is done for making the difference from the mean positive.

For example,

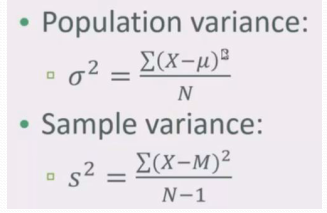
In previous case of heights,

Sigma square / variance = ( (10-11.5)\*\*2 + ….. (13-11)\*\*2 ) / 6

= ( (-1.5)\*\*2 + ….. 2\*\*2 ) / 6

= (2.25 + + 2.25 + 4 )/6

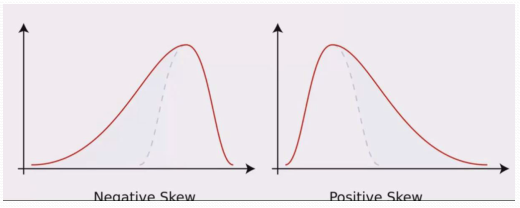
=3.25



### Skew

Skew defines the polarity of the data set, if the dataset has more values on the higher spectrum of the data set, it is said to be having a negative skew (left side has a longer tail, fig a). If the values are of the dataset is high in the initial spectrum of the x-axis, it has a positive skew as described in fig b.

**Note:** For a perfect normal distribution, the skew will be 0.



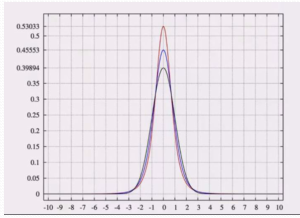
### Kurtosis

The kurtosis is a measurement of the graph sharpness of the data around the median. It is measured with comparison to the normal distribution.

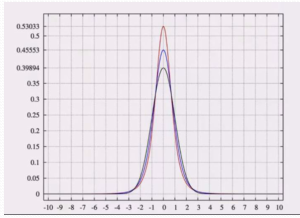
If the distribution has a sharp peak w.r.t to normal distribution it will have a positive kurtosis; higher kurtosis -> shaper peak as shown in fig a

If the distribution has a wider peak w.r.t to normal distribution it will have a negative kurtosis; lower kurtosis -> blunt peak as shown in fig b

High



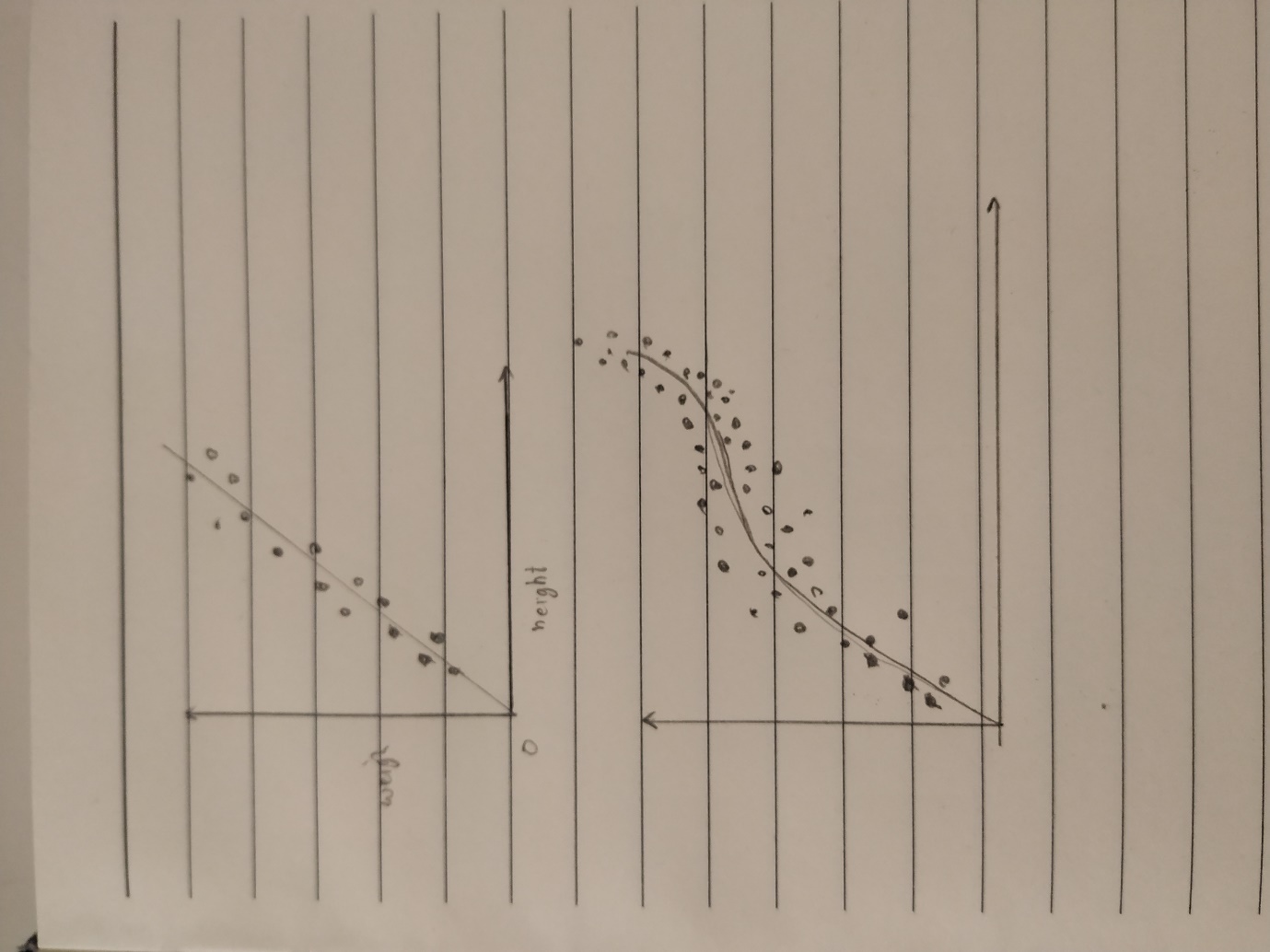
**Low**



## “Linear regression has its limitations. Polynomial Regression and Multivariate Regression address some of these but have their own challenges”. Discuss (10m)

### Linear Regression

Linear Regression can be used to determine a correlation between parameters which in turn can be used to predict a corresponding value of a parameter, given the other parameter. As illustrated in the graph below, the data points are approximately residing in a straight line, this property is called a linear correlation between variables.



Usually, a “least squares error” is used for finding the line of correlation. In this method, a line is selected which has the minimum sum of least squared error from each point. Gradient descent also can be used which finds the line using iterative method.

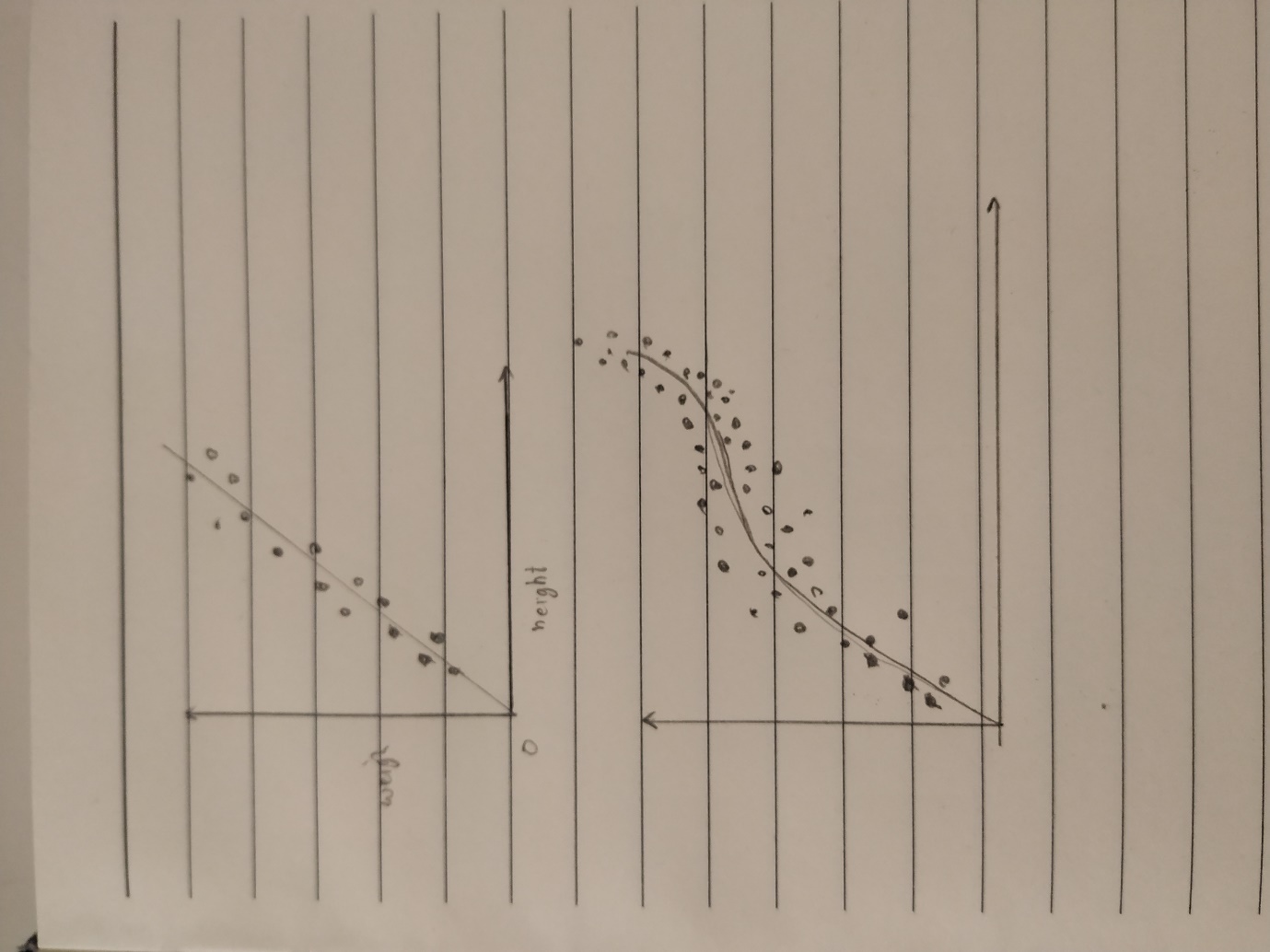
**Limitations**

In real-world data sets, only in a few cases, the parameters exhibit this kind of linear correlation.

Most datasets consist of multiple features which have an effect on the target parameter.

### Polynomial Regression

When the parameters do not have a linear correlation, we can use the polynomial regression where a polynomial function with a higher degree can be used to represent the correlations. The objective remains the same as the linear regression. Minimize the sum of the squared errors of each point from the polynomial line. As shown in the graph a polynomial function is selected which can accommodate the best line.



The polynomial of degree 3 will be of the form y = ax3 + bx2 + cx + d.

**Pros**

Allows creating more complex curves which can be used to represent non-linear correlations.

**Challenges**

Selecting the best degree for the polynomial is the most challenging task in the polynomial regression

Polynomial regression is prone to overfitting if the degree of the function selected is high.

Computationally expensive as the degree of the polynomials gets higher

### Multivariate Regression

This type of regression is used when multiple variables have an effect on the target parameters. For example, the price of the car could be determined by features like engine capacity, power, torque, seating grade, mileage etc. In these scenarios, we will have to use a multi-variable equation

TargetValue = Alpha + B1\*feature1 = B2 \* feature2 …. Bn\*feature.

**Pros**

Direct applications in real-world applications.

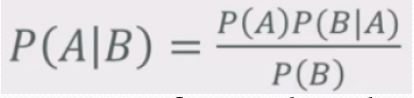
Allows representing target as a function of more than one variable.

**Challenges**

Computationally expensive.   
Complex functions which are harder for human brains to visualize.

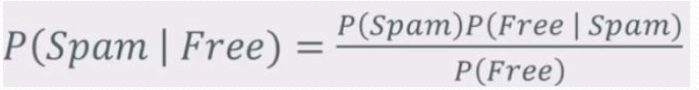
## Explain the operation of Bayes Theorem with the aid of an example. (7m)

Bayes Theorem is applied for finding the conditional probability of an event A given B using the conditional probability of B given A and independent probabilities of A and B as shown in the equation.



The probability of A given B is the probability of A times the probability of B given A over the probability of B.

This relation can be used in machine learning because by analysing the training data, we can derive the probabilities A, B and A|B can be derived. Hence, using this formula, we can estimate the probability of a new event occurring can be derived. For example, an email spam classifier can be considered. Training data consists of texts of the mail and class of that mail. In this example, Bayes theorem can be applied as below for the keyword “Free” in the text.



Where P(Spam) = probability of a mail being spam = number spam emails / total number of mails,

P(Free) = probability of a mail containing the word “Free” = number of emails containing word free/ total number of mails,

P(Free|Spam) = probability of a mail containing word “Free” given the mail is spam = number of spam emails containing words “free” / total number of spam emails,

Once these three variables are calculated using the training data. Once a new mail comes and it contains the word “spam”, we can calculate the probability of that email being spam.

## What is Unsupervised Learning and describe the operation of an unsupervised learning algorithm you are familiar with. (10m)

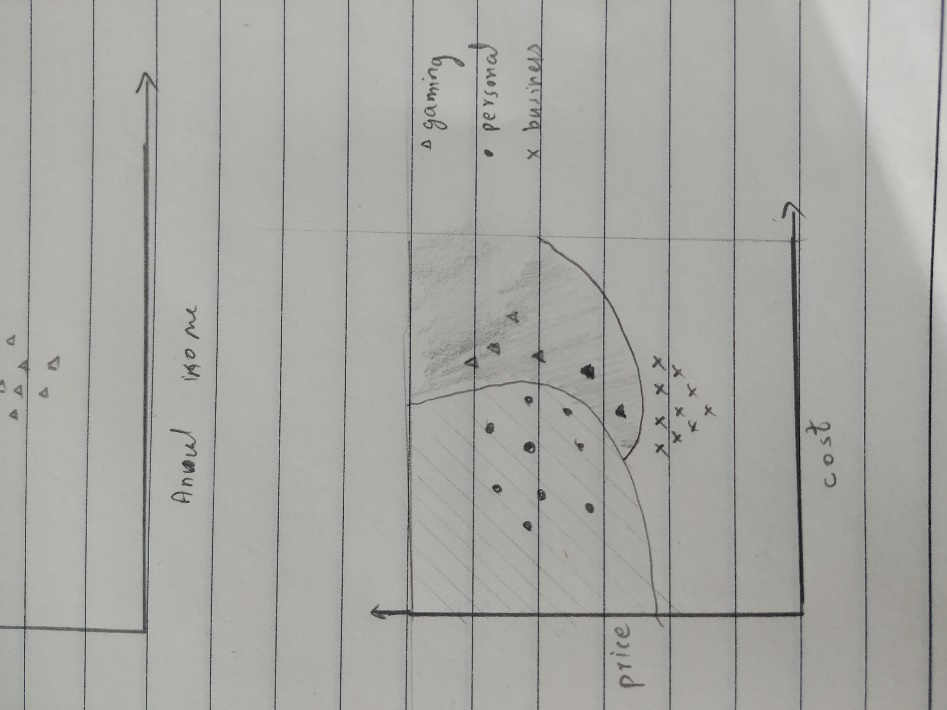
7b

## "Explain the operation of Support Vector Machines and in what cases do they work particularly well? (10m)"

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems.

SVM uses the following theorem. If the number of features is n, we have to use hyperplanes of dimension n-1 to categorise data. A hyperplane is a subspace whose dimension is one less than that of its ambient space. Support vector machine is a mathematical technique which can find the hyperplanes which can be used to categorise the data points in this ambient space(hyperspace).

In the below diagram, ambient space with dimension 2 (a 2D plane) and hyperplanes of dimension 1 (2 polynomials with degree 2) is used divide the plane into 3 areas.



In this example, a polynomial kernel is used which have a polynomial with multiple variables but of the maximum degree of 2 (simple multivariate regression). Other options for kernels are ‘linear’ which will use polynomials with 1 degree (straight lines), ‘sigmoid’ which use tanh function and ‘radial binomial functions’ (rbf/Gaussian) which uses radial degree function.

SVM works well when the numbers of features are high. For example, if the features are 10 and there are non-linear relations between the features. An SVM with kernel poly will have better performance, but a decision tree is selected for this purpose the depth may be too high, which may result in overfitting.

Another scenario where SVM is desired is when the problem statements are used for categorisation of multi-class problems, due to the inherent nature of the planes in dividing the spaces into categories.

## Outline the general operation of Multi-level models (5m)

There are complex decision problems which have correlations at multiple levels of datasets. In Linear, polynomial and multivariate regression problem, we dealt with only flat data, where features and items had one-to-one relationships. In multilevel, there could be one-many relationships in terms of data mapping. For example, the health of a person could depend upon, DNA structure, food habits, family, city etc. Another example, is the wealth of person could depend on people around him like the wealth of parents, grandparents etc. These interdependencies is the key aspect multi-level model tries to address.

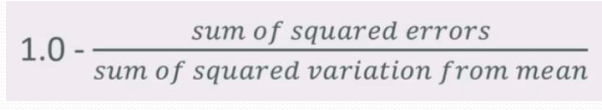
In multi-level model, effect of a factor in a higher level could create a magnifying effect in the lower levels. Furthermore, some factors may have effect on multiple levels. Social science is a field where these kinds of relationships have to be tackled. For example, a GRE exam score could be a factor of his intellectual capabilities, coaching he got which is depended on the financial support he receives which in turn could be dependent on the job of his parents or the quality of the school he attended.

## "Describe the operation and relevance of each of the following in relation to Machine Learning: (6m)

### R-Square error

The fraction of the total variation in Y that is captured by the model

R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.

It is defined as R2 Value = 

A higher value of r2 is generally regarded as a good model because it means the regression function identified have been able to accommodate most of the values along its path. Value of R2 is always between 0 and 1, 0 means none of the variances is captured and 1 means all the variance in the target variable is captured.

### Variance

The variance gives an estimation of how much “diverse“ or “spread-out” is the data. It is calculated as the average of the squared difference of the data from the mean of the data set.

Note: Squaring is done for making the difference from the mean positive.

For example,

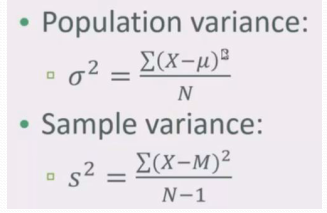
In previous case of heights,

Signma square / variance = ( (10-11.5)\*\*2 + ….. (13-11)\*\*2 ) / 6

= ( (-1.5)\*\*2 + ….. 2\*\*2 ) / 6

= (2.25 + + 2.25 + 4 )/6

=3.25



### K-Fold Cross-Validation

It is an advanced way of testing the accuracy of the model especially when the dataset available is of smaller size. In normal train test distribution strategery, a data point will be present in either train or test data set. If it is in the training dataset, the point has an effect on the model but not while calculating accuracy. And if it is part of the test dataset, the value is not considered for training. K-Fold is designed to overcome this while evaluating the model, to give a chance for each datapoint to contribute both as train and test data.s

Working of K-fold:

1. Split the dataset into K randomly selected segments.
2. Pick the first segment and assign it as validation dataset.
3. Label the rest of the K-1 segments as the training dataset.
4. Train, test and calculate the R2 values
5. Repeat steps 2,3 and 4 for all the segments.
6. Take the average of R2 values of all iterations – this will be the CV score.

## K-Nearest Neighbours and K-Means clustering are often confused. Describe the operation of both (5m)

K-nearest Neighbours is an algorithm which can be used to find categories or to rank similarities based on the distance between data points. The calculation of the distance is usually done an arbitrary function which is specific to the domain and features available in the dataset.

Steps for K-nearest neighbour to find a category is as follows:

1. Define a function for computing distance.

For example, for the height and weight of students in a class, we can set the distance function as weight squared/height squared.

1. For categorisation problem, find the distances between each existing datapoint and new point to be categorised.
2. Sort the points based on ascending order of distances from the new point.
3. Keep only the top K points in the list and drop the remaining items.
4. Select the class which has the highest number of presence on the list.

K-Means

7b

## What is the ‘Curse of Dimensionality’ and using examples, describe an approach to overcome it? (10m)

Same as 32.

## Describe the operation of a Recommender System that you are familiar with. (5m)

Same as 31.

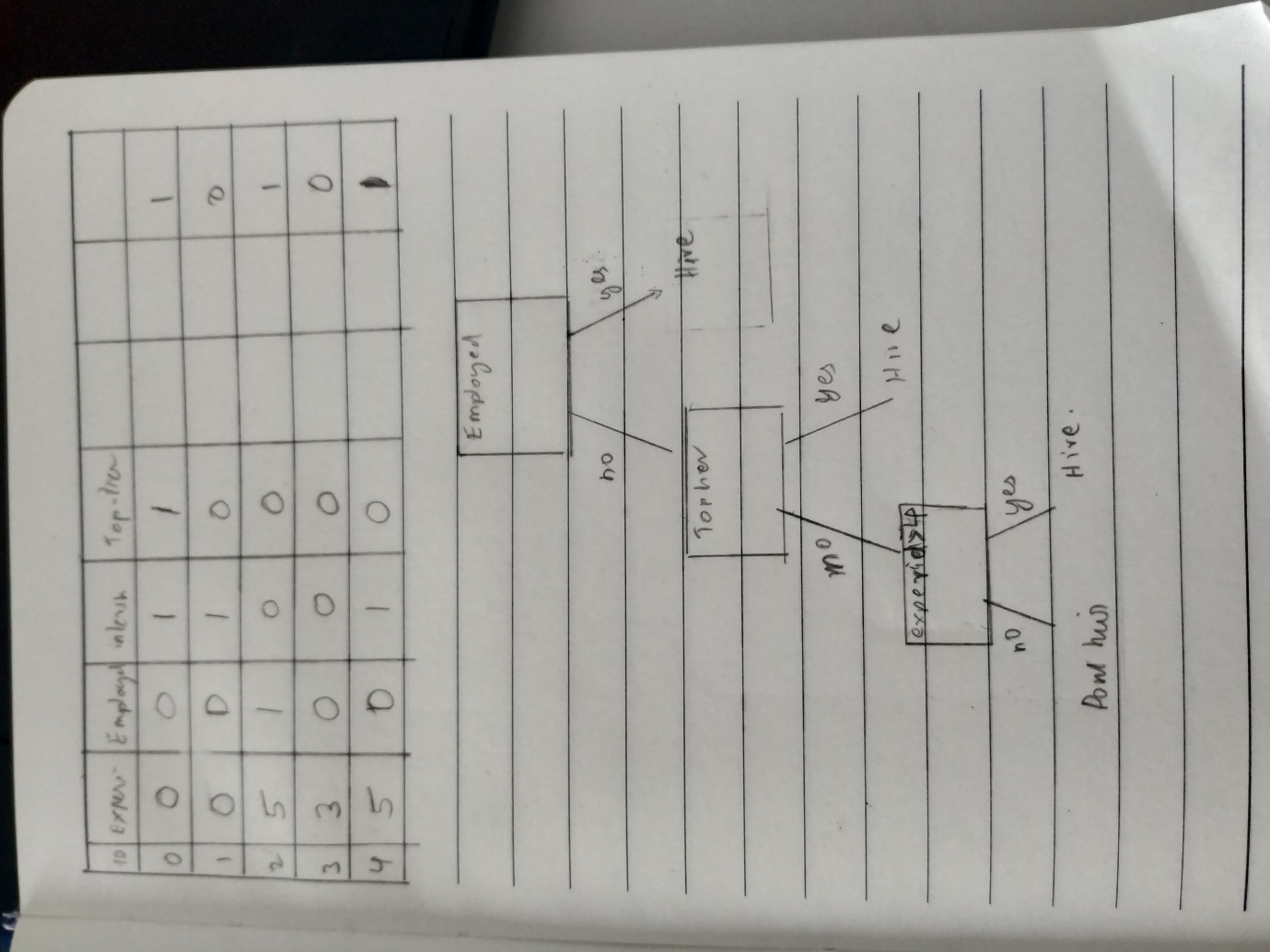
## With the aid of a diagram and an example describe how Decision Trees work? (10m)

A Decision tree is a machine learning algorithm which tries to determine a class by evaluating the effects of each features and its values. Decision tree follows a simple algorithm, pick a feature and a value for the same which reduces the entropy of each spilt.

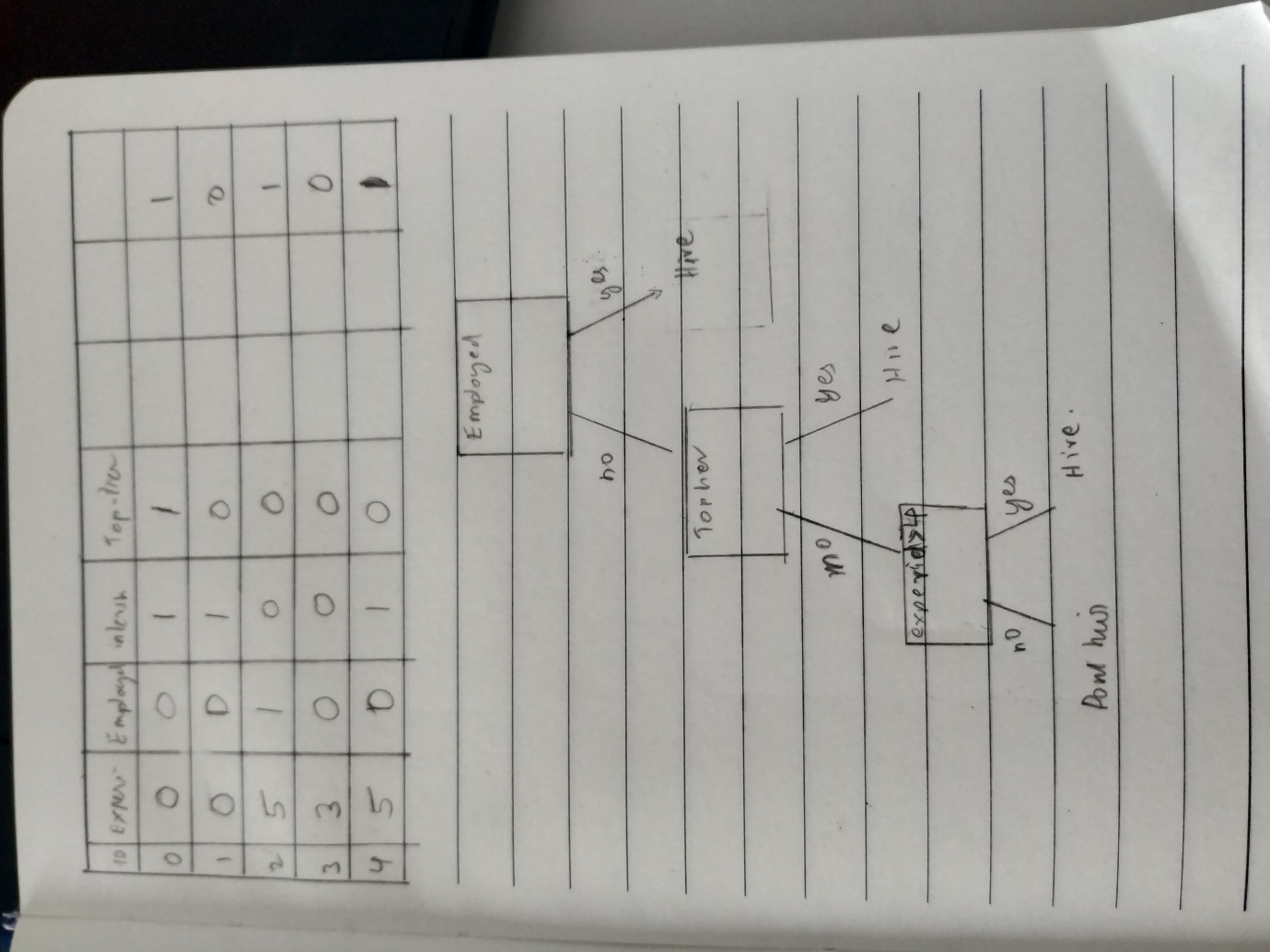
Decision Tree works only on numerical or ordinal data types. If categorical data is included, it has to be converted to sparse matrix with binary representation of each class as true or false.

The degree of Gini index varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes.

A data about a employee is hired or not is given below.



Based on this , we will create a decision tree.



As demonstrated in the diagram above, each parameter is taken and find the values which give entropy in the resulting splits.

Example for creating a decision tree

* 1. Determine the Root of the Tree
  2. Calculate Entropy for The Classes
  3. Calculate Entropy After Split for Each Attribute
  4. Calculate Information Gain for each split
  5. Perform the Split
  6. Perform Further Splits
  7. Complete the Decision Tree

## What is a Random Forest and what are they designed to overcome? Explain in the context of Bagging (5m)

Random forests are an ensemble learning method for classification, regression and other tasks which is usually solved using Decision Trees. Individual decision tree tends to overfit the model. This reduces the ability to deal with variance in data, especially in multi-dimensional correlation. To reduce the risk of overfitting, the random forest method is created. Random forest tree if used in the bagging context uses an independent train and prediction system. The dataset is randomly selected with replacement to create K-sample sub-datasets. K-number of decision trees is independently trained on respective sub-datasets. Afterwards, when a new item is given for processing individual decision trees will process and give a prediction. An aggregation/voting is done to give the final result

## Illustrate the operation of the K-Nearest Neighbours algorithm. (5m)

K-nearest Neighbours is an algorithm which can be used to find categories or to rank similarities based on the distance between data points. The calculation of the distance is usually done an arbitrary function which is specific to the domain and features available in the dataset.

Steps for K-nearest neighbour to find a category is as follows:

1. Define a function for computing distance.

For example, for the height and weight of students in a class, we can set the distance function as weight squared/height squared.

1. For categorisation problem, find the distances between each existing datapoint and new point to be categorised.
2. Sort the points based on ascending order of distances from the new point.
3. Keep only the top K points in the list and drop the remaining items.
4. Select the class which has the highest number of presence on the list.

## Describe the circumstances in which Ensemble Learning and Boosting would be appropriate? (4m)

### Ensemble methods

Ensemble methods combine several decision trees classifiers to produce better predictive performance than a single decision tree classifier. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner, thus increasing the accuracy of the model. When we try to predict the target variable using any machine learning technique, the main causes of the difference in actual and predicted values are noise, variance, and bias. Ensemble helps to reduce these factors (except noise, which is an irreducible error).

### Boosting

Boosting refers to a group of algorithms that utilize weighted averages to make weak learners into stronger learners. Unlike bagging that had each model run independently and then aggregates the outputs at the end without preference to any model. Boosting is all about “teamwork”. Each model that runs, dictates what features the next model will focus on.

## Explain how Train/Test works. What are the benefits and weaknesses of this approach and how may the weaknesses be overcome? (10m)

## Describe Q-Learning (5m)

Ans: Q-learning is an off policy reinforcement learning algorithm that seeks to find the best action to take given the current state. It’s considered off-policy because the q-learning function learns from actions that are outside the current policy, like taking random actions, and therefore a policy isn’t needed. More specifically, q-learning seeks to learn a policy that maximizes the total reward.

You can look ahead more than one step by using a discount factor when computing Q (here s is previous state, s’ is the current state)

Q(s,a) += discount \* (reward (s,a) + max(Q(s’)) – Q(s,a))

Implementation of reinforcement learning:

You have:

* A set of environmental states s
* A set of possible actions for those states a
* A value of each state/action Q
* Start off with Q values of 0
* Explore the space
* As bad things happen after a given state/action, reduce its Q
* As rewards happen after a given state/action, increase its Q

## What is the difference between User-based collaborative filtering and Item-based Collaborative filtering? With the aid of examples explain how does each work? (10m)

Collaborative Filtering:  
It is a method used for finding similar things to suggest for users in the recommender systems. There are two major types of collaborative filtering methods. Those are:

1. User-based collaborative filtering:

The user-based collaborative filtering is a user-centric technique. The similarity metrics between users is calculated based on user personalities. Afterwards, this similarity scores is used to recommend new items to a user based on similar users.

For example, in a shopping site.

* 1. Use the user’s age, income, location, shopping history, the ratings of items bought and viewed items to establish a similarity measurement between each user.
  2. Then evaluate the current state of a user, for example, the person of age 40, with 2 kids with an income of 100000 Euros annually have bought a table.
  3. Check what other users of similar features have bought in recent times, for example, a kitchen sink, portable drying stand etc. and suggest these things to the user.

Drawbacks:

* 1. The users are fickle, their taste changes over time.
  2. The user metrics may not represent the real-world preference, as the psychology of users is hard to be defined in mathematical models.

1. Item-based collaborative filtering:

In item-based collaborative filtering, the similarity is calculated between items rather than users. This helps in two dimensions:

1. Items properties are constant
2. second they are more clearly defined which allows a more precise calculation of similarities.

For example, laptop specs like Ram, Storage is more clearly distinguishable than users preferences which can change like colour, shape etc and usage like for gaming or work which can change over time.

In the example of a shopping site, steps are as follows,

* 1. Use the item’s specifications like size, category, usage, price etc to determine the similarity indices.
  2. Find the items which are similar or bought together and create a similarity index matrix between them.
  3. When a user buys an item show an item which is bought together or when a user views a potential purchase, show the similar items which can help him to compare the features and make a decision. (Usually, an expensive item with better features are shown in order to increase the sales)

Advantages:

* 1. The similarity indices are more or less constant.
  2. It is harder to manipulate the system behaviour by artificial behaviours.

## Describe the operation of Principal Component Analysis and outline how the technique is related to the curse of dimensionality. (10m)

Principal Component Analysis is a concept of reducing the dimensions/features of the dataset while maintaining the maximum variance as possible. It is achieved using Eigenvectors of features.

The steps to perform a principal component analysis for a data set with features four features a,b,c,d which need to be reduced to 2 dimensions.

1. Calculate the covariance matrix X of data points.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Va |  |  |  | Ca,e |
|  | Vb |  | Cb,d |  |
|  |  | Vc |  |  |
|  | Cb,d |  | Vd |  |
| Ca,e |  |  |  | Ve |

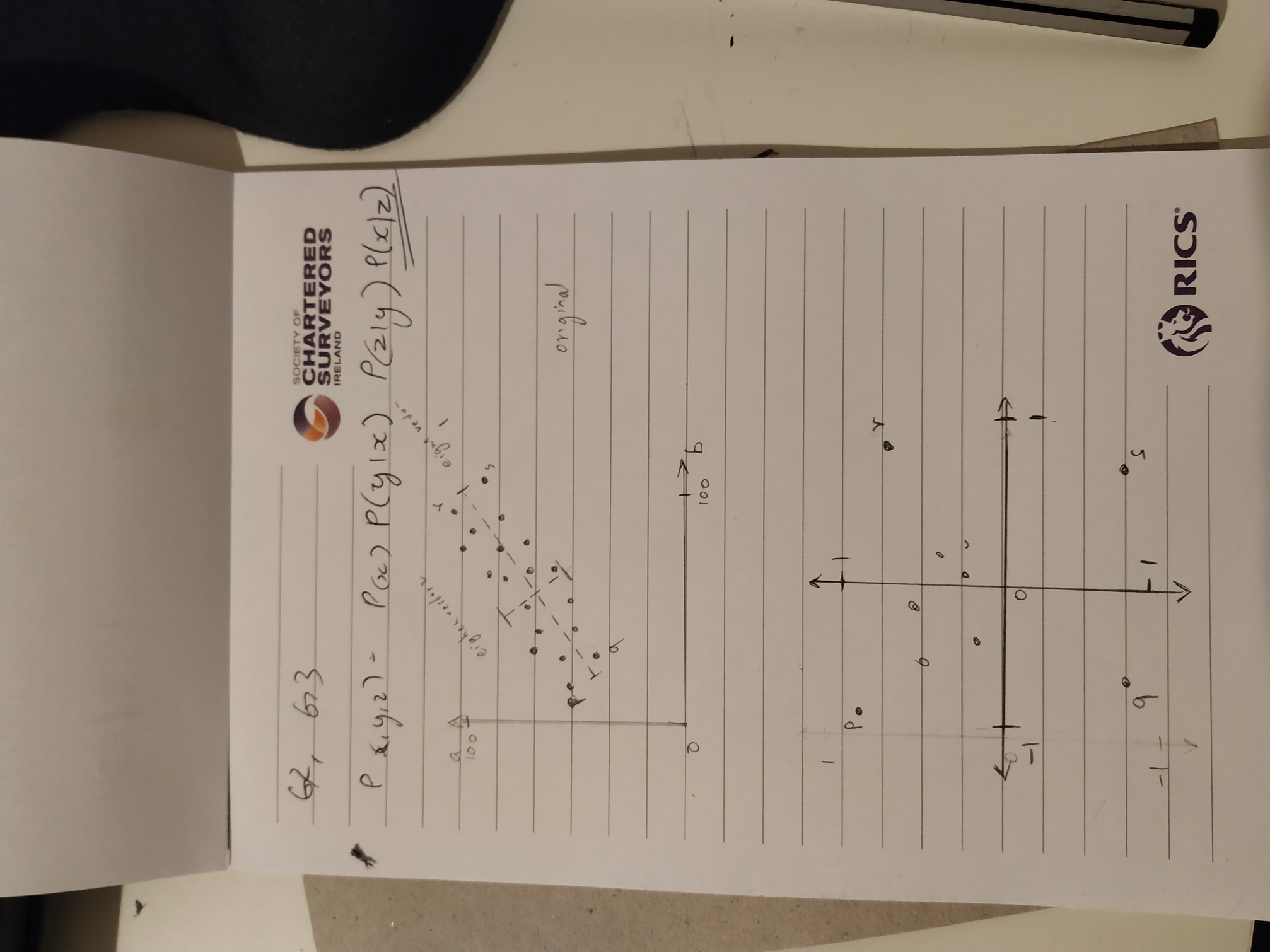
Where C a,b = C b,a

1. Calculate eigen vectors and corresponding eigen values.

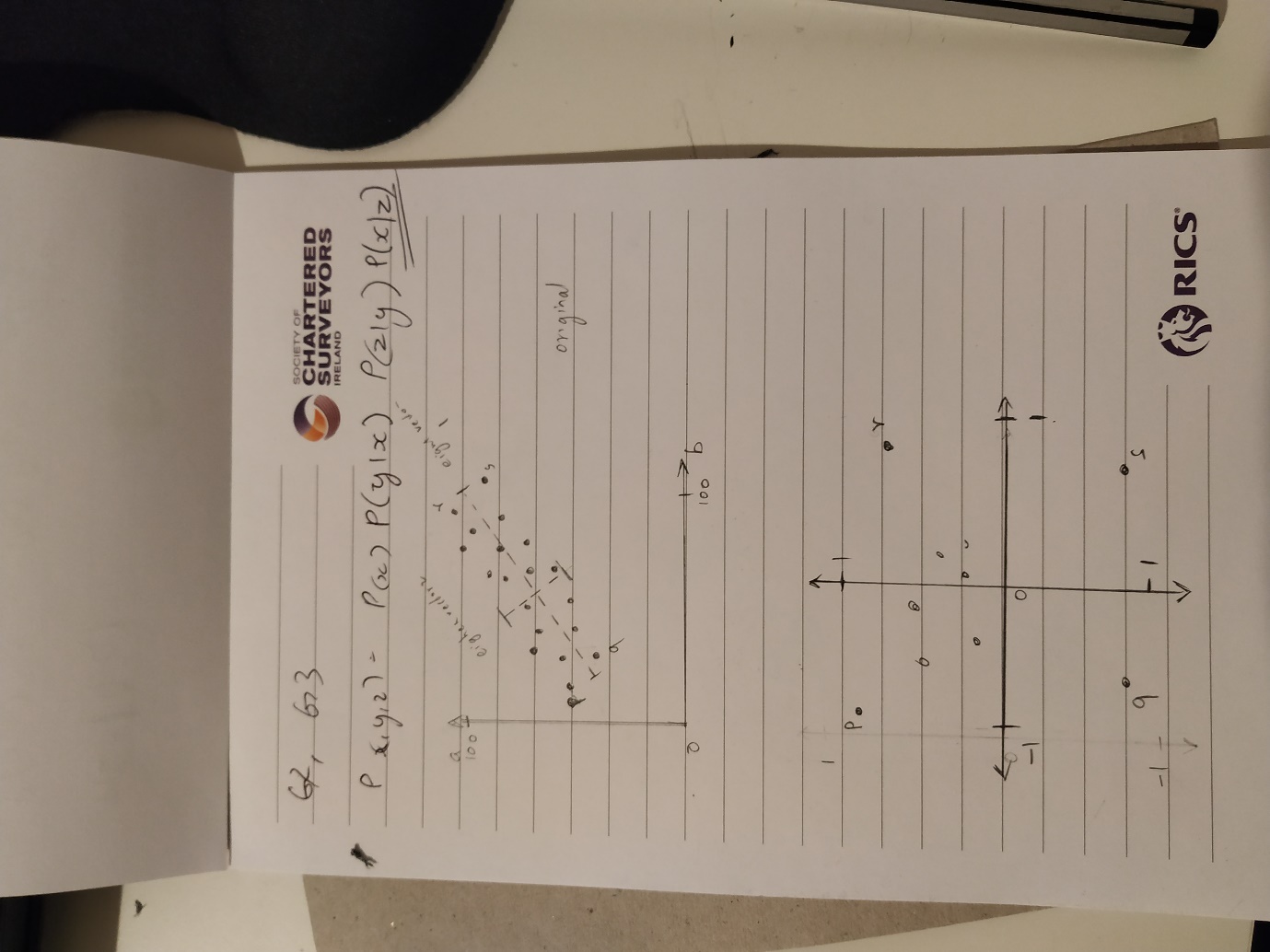
Eigenvectors with the highest variance will have the highest eigenvalues.

1. Sort the eigenvectors according to their eigenvalues in decreasing order.
2. Choose the first 2 eigenvectors and that will be the new 2 dimensions of the new dataset.
3. Transform the original 4-dimensional data points into 2 dimensions.

Original dataset plotted along the features a and b.



After transformation into 2 new eigenvectors, notice the position points p,q,r and s in the new projection for understanding the transformation.

1. 

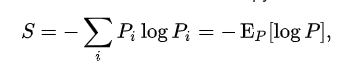
After this, the dataset is ready for training and prediction. Note: new values which come in for categorisation or prediction should be converted into the new dimensions set before giving to the model trained in the PCA plane.

Curse of Dimensionality.

Most of the data we received will be generated by star scheme-based Database systems where data structures highly relay on dim tables to find the meaning of the fact table. This often creates a large set of dimension data which may be having only a slight dependency. But this creates a stress on the machine learning algorithms which are computationally expensive to train and predict with those many features. The principal component analysis is a method which can be used to reduce the dimensions but can be used to preserve variance using the transformations of the data into Eigen Vectors.

## What is Entropy and how is it useful in Machine Learning? (5m)

Entropy measure of a data set’s disorder. That is, how same or different it is. If we classify a data set into N classes, if all element in a class is same then it has zero entropy. For example, in a data set of animal attributes and their species, we select a class for example, monkey. If all features of all members of the class are same like same height, weight etc. then it has zero entropy and vice versa.

Entropy is defined as 

P of i represents the proportion of the data labelled for each class

Entropy is crucial in terms of machine learning modelling. In terms of training, models will learn to predict or categorise diverse types of data. But this will come at cost of higher complex models with more resource-savvy computations. But this reduce the overall accuracy in some scenarios. On the other hands, if entropy is very low, then the machine learning task becomes too mundane, which could have solved using conventional programming paradigms.

## What is K-fold cross-validation? Explain how it works (5m)

Same as 21 c

## With the use of examples to illustrate your answer, describe Boosting and Bagging? (10m)

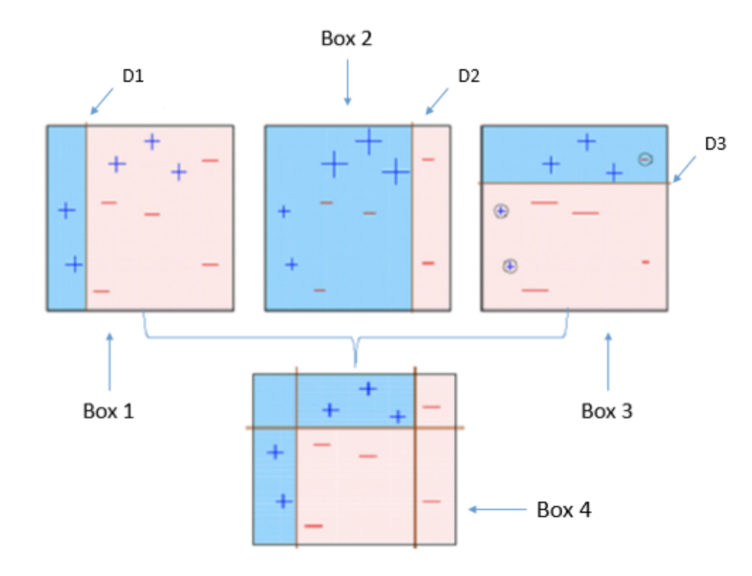
### Bagging

Bagging is a type of ensemble learning where multiple models are trained independently, typically used in decision trees. The following steps are followed for implementing bagging.

* 1. The dataset is randomly selected with replacement to create K-sample sub-datasets (bootstrapping)
  2. K-number of models is independently trained on respective sub-datasets.
  3. When a new item is given for evaluation, individual models will process and give a prediction.
  4. Then, an aggregation/voting is done to give the final result.

### Boosting

Boosting refers to a group of algorithms that utilize weighted averages to make weak learners into stronger learners. Unlike bagging that had each model run independently and then aggregates the outputs at the end without preference to any model. Boosting is all about “teamwork”. Each model that runs, dictates what features the next model will focus on.



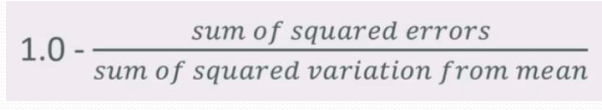
<https://becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e>

Bagging and boosting both give better stability. Bagging reduces the risk of overfitting and boosting increases the bias which gives more specificity to the model which reduces the error margin.

## What does the r-squared measure? What is the formula used to calculate r-squared and is an r-squared value of 0.08 good or bad? Explain (5m)

The fraction of the total variation in Y that is captured by the model

R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.

It is defined as R2 Value = 

A higher value of r2 is generally regarded as a good model because it means the regression function identified have been able to accommodate most of the values along its path. Value of R2 is always between 0 and 1, 0 means none of the variances is captured and 1 means all the variance in the target variable is captured.

Hence, 0.08 is a poor R2 score for a regression line.

## How is Covariance related to Correlation? What does a small Covariance (close to 0) and a large Covariance (far from 0) mean? (5m)

Think of the data sets for the two variables as high-dimensional vectors

Convert these to vectors of variances from the mean

Take the dot product (cosine of the angle between them) of the two vectors

Divide by the sample size

We know a small covariance, close to 0, means there isn’t much correlation between the two variables. And large covariance – that is, far from 0 (could be negative for inverse relationships) mean there is a correlation. But how large is “large”?

Just divide the covariance by the standard deviations of both variables, and that normalizes things.

So, a correlation of -1 means a perfect inverse correlation

Correlation of 0: no correlation

Correlation 1: perfect correlation

## In what way is the concept of Least Squares related to Linear regression (5m)

The least square is the minimum Euclidian distance between a point and a line.

In terms of linear regression, the line with the least sum of the distances from each point is selected as the regression line.

## Where would you need to use Ensemble Learning and Boosting? (5m)

### Ensemble methods

## Ensemble methods combine several decision trees classifiers to produce better predictive performance than a single decision tree classifier. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner, thus increasing the accuracy of the model. When we try to predict the target variable using any machine learning technique, the main causes of the difference in actual and predicted values are noise, variance, and bias. Ensemble helps to reduce these factors (except noise, which is an irreducible error).

### Boosting

Boosting refers to a group of algorithms that utilize weighted averages to make weak learners into stronger learners. Unlike bagging that had each model run independently and then aggregates the outputs at the end without preference to any model. Boosting is all about “teamwork”. Each model that runs, dictates what features the next model will focus on.

## With the aid of diagrams, explain the operation of a modern deep neural network. In your answer discuss Artificial neurons, LTUs, Perceptron, Multi-Layer Perceptron and Softmax. (15m)

Same as 44

## Outline the steps that could be taken to avoid overfitting in a neural network (5m)

Same as 51

## What is TensorFlow and why is it used for Deep Learning? (5m)

TensorFlow is an end-to-end platform that makes it easy for you to build and deploy ML models.

TensorFlow has always provided a direct path to production. Whether it’s on servers, edge devices, or the web, TensorFlow lets you train and deploy your model easily, no matter what language or platform you use.

A tensor is an algebraic object that describes a linear mapping from one set of algebraic objects to another.

TensorFlow is a framework which is developed in highly efficient C++ Code with easy to use Python API for matrices calculations. It helps in parallel execution of arithmetic calculations which required for deep learning and facilitated by the GPU’s processing architecture. But they were not as per se designed as deep learning end to end platform. With the easy to use advanced high-level API’s in Keras has facilitated the deep learning in TensorFlow much easier.

## "Give a brief description of each of the following: (6m)

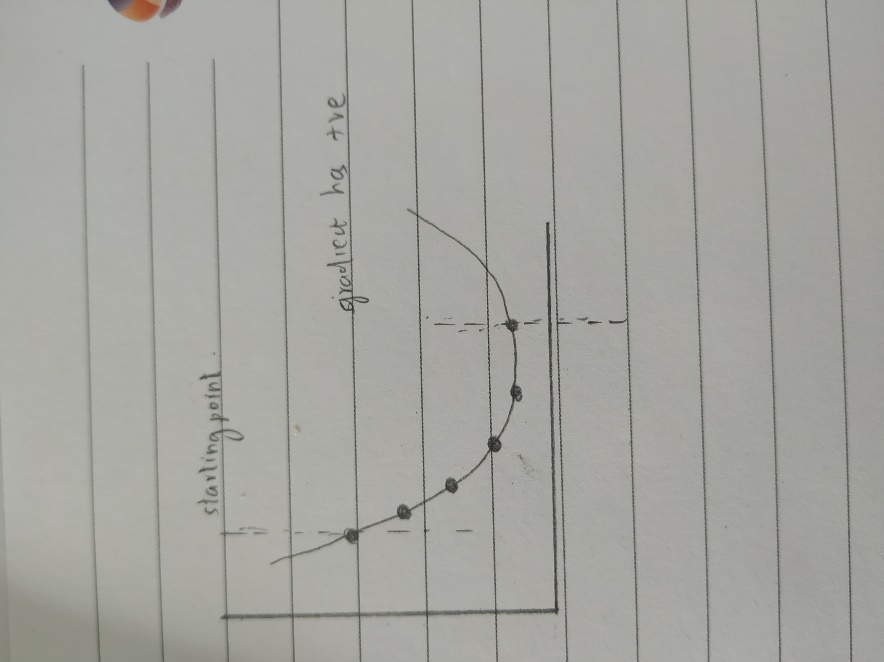
### Gradient descent

### Autodiff

### Softmax "

### Gradient descent:

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.



### Autodiff:

Autodiff is a technique to numerically evaluate the derivative of a function specified by a computer program using simply chaining the basic arithmetic functions like addition, subtraction, multiplication and division.

Autodiff is a calculus trick for finding the gradients in gradient descent

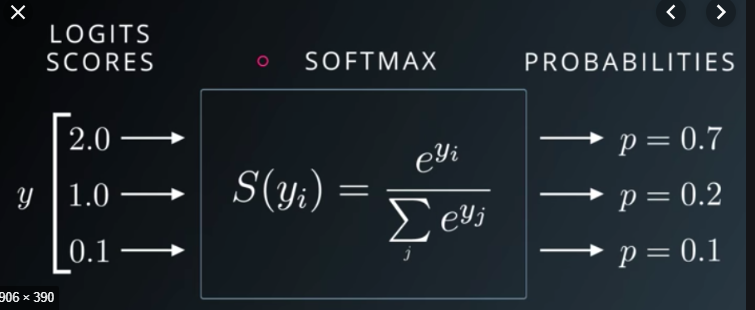
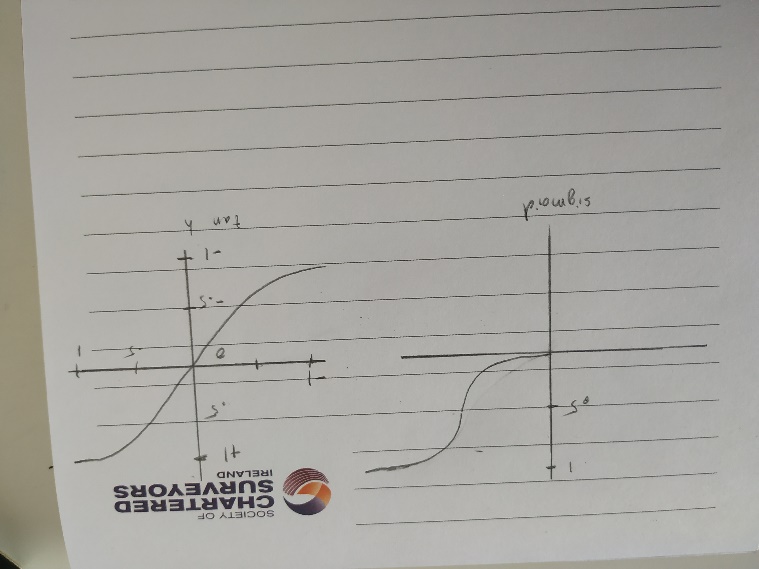
As the Calculus is hard to compute using logical gates, this is a huge advantage

The operations are aligned with the structure of the neural network making it ideal for training.

Optimized for many inputs and few outputs like a neuron.

### Softmax:

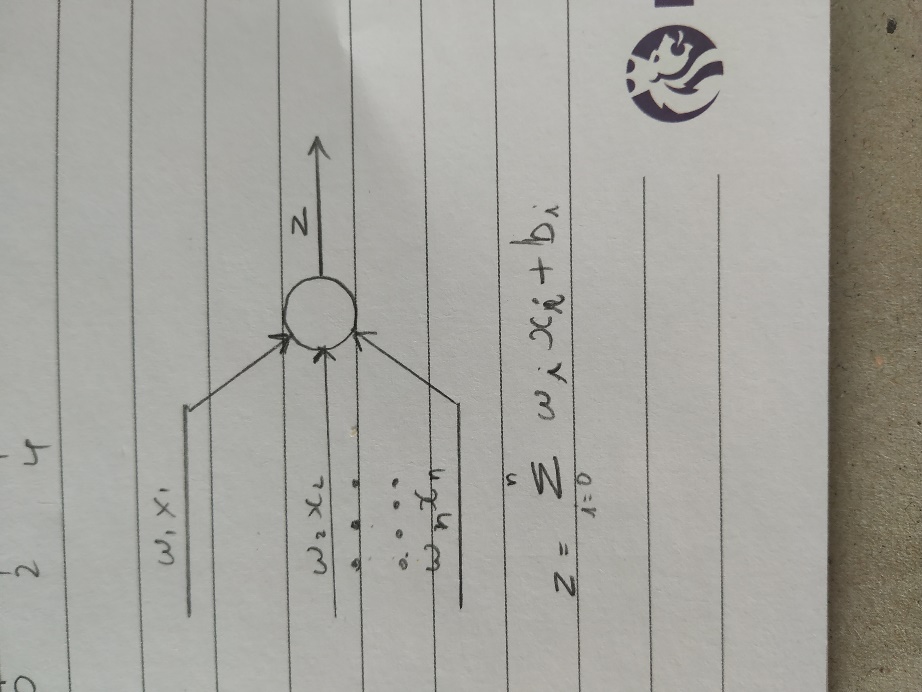
Softmax is a function for choosing the most probable classification given several input values. The output is a value between 0 and 1 inclusive. It is used in multi class problems where it returns probability of each class, with the target class having the highest probability.



## With the aid of diagrams, explain the operation of a modern deep neural network. (9m)

### Perceptron:

A perceptron is a bio-inspired atomic structure of a neural network, which takes multiple inputs and gives out one output.



The activation function z = w\*x +b determines the output values.

B is bias which is also a learnable parameter.

### Activation functions

Various activation functions are:

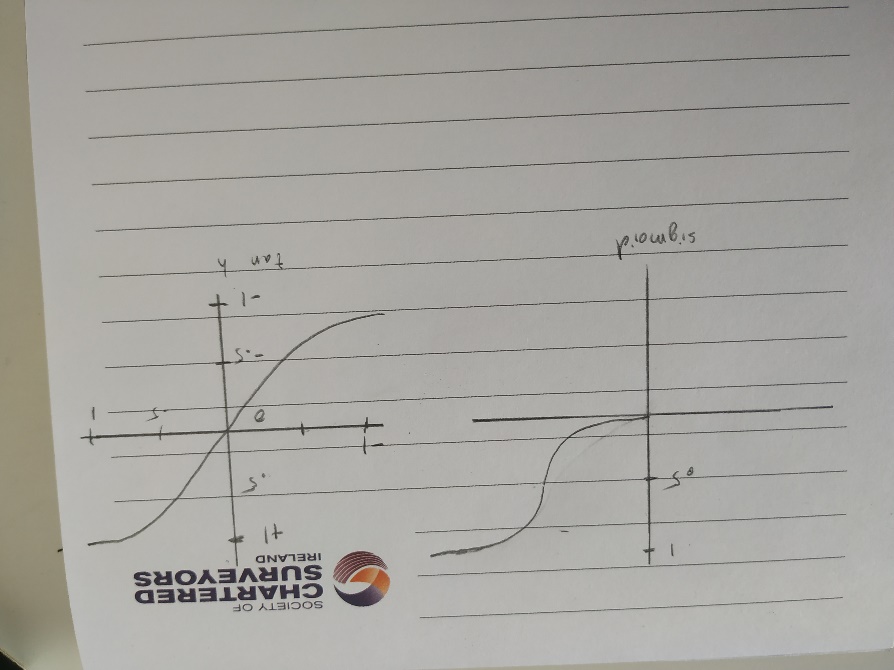
Linear Neuron / Step function

F(x) = 1 if sum of w\*x is > threshold.

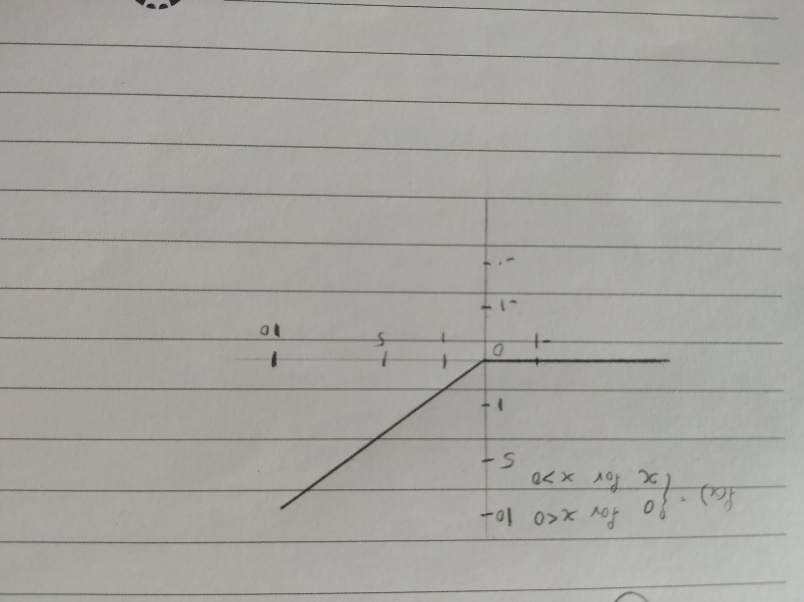
0 if the sum of w\*x is < threshold.

Sigmoid Functions

Converts the value of x into the range of 0 to 1 with a smoothening. But saturates at the when nearing 0 or 1.

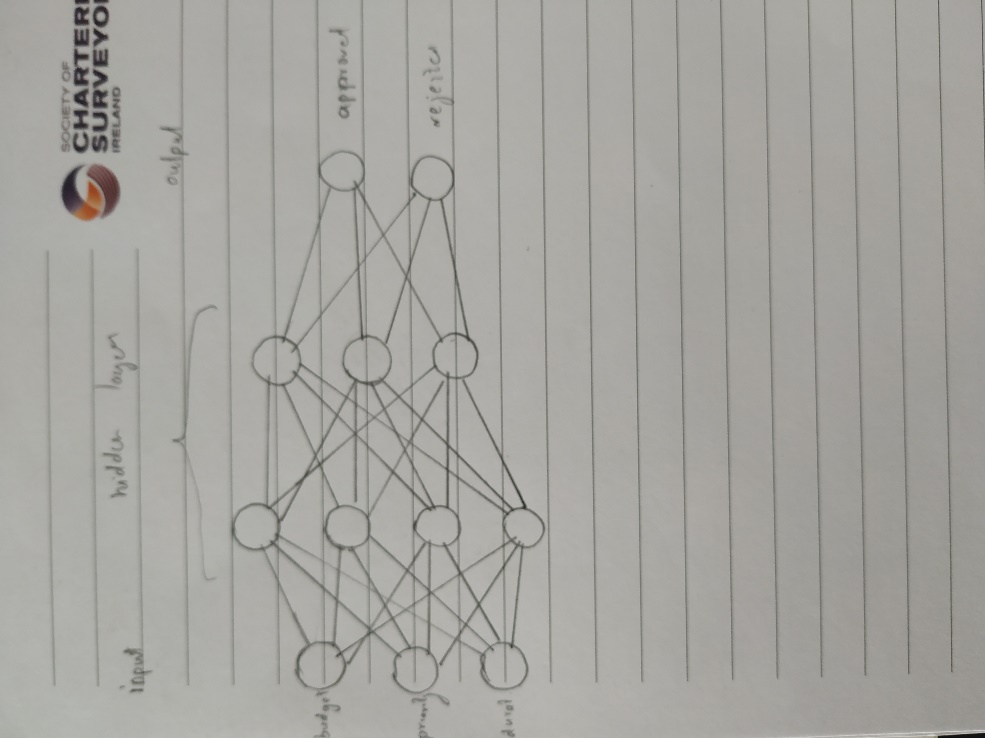


Rectified Linear Unit (ReLu):



### Deep Neural Networks:

The most basic structure of a neural network is as shown in the diagram below.



It consists of input layers of which inputs like pixels of the image, features in a dataset can be fed into. It has an output layer which gives out the output classes for categorisation problems. There are one or more hidden layers between the input and output layers. When a neural network has more than one hidden layer it is called deep neural networks and process of training such a network is called deep learning.

### Training:

The NN can be trained using the following steps for one epoch:

1. Initially, random weights and bias are assigned to neuron network.
2. Divide the data set into n batch with a size of s.
3. Pass each set of features one by one to the input layers for all items in a batch.
4. Compute the output error for each row.
5. Compute how much each neuron in the previously hidden layer contributed
6. Back-propagate that error in a reverse pass
7. Tweak weights to reduce the error using gradient descent
8. Repeat the steps 3,4,5,6,7 for all batches.

These steps constitute one epoch, depends on the precision after each epoch and desired accuracy, we decide the number epochs we should train the network.

### Hyperparameters:

Depending on the use case, dataset size, features and computing constraints. The following hyperparameters should be decided by the NN Engineer.

#### Batch size:

This determines the number of data points to be passed in a single cycle of feed forwarding, backpropagation and adjustment of weights and bias. If the size is very less, it may lead to overfitting.

#### Learning Rate:

This hyperparameter determines the rate at which the adjustments should be done to the network, in other words, the step size in the gradient descent method for reaching minima. If it is more, the accuracy will fluctuate around the optimal value. If it is small, then the learning will be too slow, which will increase the number of epochs required to reach the minima. Subsequentially, higher computing cost.

#### Epochs:

After completing the training of the neural network with the whole dataset (all batches) once, the NN may not be optimal yet. It takes several passes dataset to achieve the desired accuracy

## “Ethical Challenges exist with the implementation of Deep Learning”, Discuss (10m)

As our technology becomes more powerful, the potential harms from new technologies will become larger. Artificial intelligence, in particular, raises many ethical challenges, here are some:

### Biases in Algorithms

### Transparency of Algorithms

### Supremacy of Algorithm

### Fake News and Fake Videos

### **Lethal Autonomous Weapon Systems**

### Self-driving cars

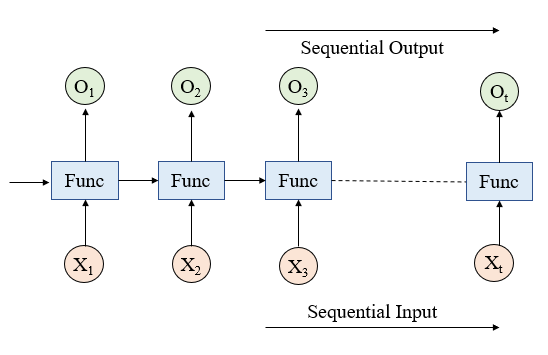
### Privacy vs Surveillance

<https://towardsdatascience.com/7-short-term-ai-ethics-questions-32791956a6ad>

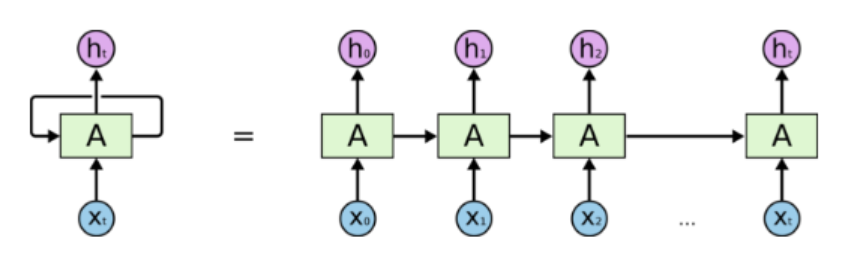
## Describe the operation of a Recurrent Neural Network (RNN) and its application areas. What are the challenges associated with the use of RNNs? (10m)

A **recurrent neural network** (RNN) is a type of artificial **neural network** commonly **used in** speech recognition and natural language processing (NLP). RNNs are designed to recognize data's sequential characteristics and **use** patterns to predict the next likely scenario. A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed graph along a sequence. This allows it to exhibit dynamic temporal behaviour for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

For sequential data feeding, this type of DNNs is used in order to resolve the issue of time-series [8] as represented in Fig. 5. Internal memory is used in each neuron cells in order to retain the previous state of the cell. These are used in natural language processing and language translation services where contextual meaning is critical.



Operation of a Recurrent Neural Network (RNN):



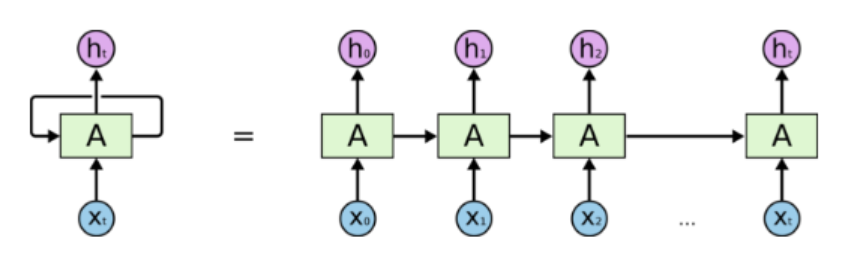
In the above diagram, a chunk of a neural network, A, looks at some input x-t and outputs a value h-t. A loop allows information to be passed from one step of the network to the next. A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor as shown on the right side of the above figure.

Applications of Recurrent Neural Networks include:

* Machine Translation.
* Robot control.
* Time series prediction.
* Speech recognition.
* Speech synthesis.
* Time series anomaly detection.
* Next word prediction.
* Music composition.
* Image captioning
* Stock market prediction

Challenges associated with the use of RNNs:

See the horizontal arrow in the diagram below:



1. This arrow means that long-term information has to sequentially travel through all cells before getting to the present processing cell. This means it can be easily corrupted by being multiplied much time by small numbers < 0. This is the cause of [vanishing gradients](http://www.wildml.com/2015/10/recurrent-neural-networks-tutorial-part-3-backpropagation-through-time-and-vanishing-gradients/).
2. It’s really hard to train RNN’s as they are very sensitive to topologies, choice of hyperparameters.
3. Very resource-intensive
4. A wrong choice can lead to an RNN that doesn’t converge at all.

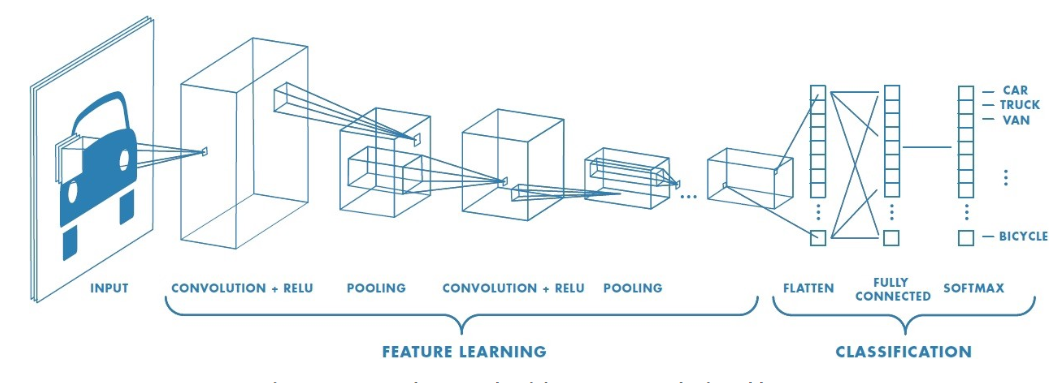
## “Deep Learning is potentially unethical”, Discuss (10m)- same as 45

## Briefly describe the ethical challenges of Deep Learning. (5m) same as 45

## Describe the operation of a Convolutional Neural Network (CNN) and its application areas. What are the challenges associated with the use of CNNs? (10m)

Convolutional neural network (CNN) is one of the main algorithms to do images recognition, images classifications. Objects detections, recognition faces etc., are some of the areas where CNNs are widely used. CNN image classifications take an input image, process it and classify it under certain categories (E.g., Dog, Cat, Tiger, Lion). Computers see an input image as an array of pixels and it depends on the image resolution. Based on the image resolution, it will see h x w x d(h = Height, w = Width, d = Dimension ).

In Deep learning CNN models, to train and test, each input image will pass it through a series of convolution layers with filters, pooling layers, fully connected layers (FC) and finally applies a Softmax function to classify an object with probabilistic values between 0 and 1. The below figure is a complete flow of CNN to process an input image and classifies the objects based on values.



**Convolution Layer:** Convolution is the first layer to extract features from an input image. Convolution of an image with different filters can perform operations such as edge detection, blur and sharpen by applying filters.

ReLU: stands for Rectified Linear Unit for a non-linear operation. The output is **ƒ(x) = max(0,x).** ReLU’s purpose is to introduce non-linearity in our ConvNet.

**Pooling Layer:** Pooling layers section would reduce the number of parameters when the images are too large.

**Fully Connected Layer:** we flattened our matrix into a vector and feed it into a fully connected layer like a neural network. With the fully connected layers, we combined these features together to create a model. Finally, we have an activation function such as Softmax or sigmoid to classify the outputs as a cat, dog, car, truck etc

**Steps in order.**

* Provide input image into convolution layer
* Choose parameters, apply filters with strides, padding if requires. Perform convolution on the image and apply ReLu activation to the matrix.
* Perform pooling to reduce dimensionality size
* Add as many convolutional layers until satisfied
* Flatten the output and feed into a fully connected layer (FC Layer)
* Output the class using an activation function (Logistic Regression with cost functions) and classifies images.

CNN application areas include:

* Decoding Facial Recognition
* Analysing Documents- handwriting analysis
* Understanding Climate- understanding the reasons why we see such drastic changes
* Object detection for self-driving cars
* Image analysis in healthcare

Challenges associated with the use of CNN:

* Very resource-intensive (CPU, GPU, and RAM)
* Lots of hyperparameters (Kernel sizes, many layers with different numbers of units, amount of pooling… in addition to the usual stuff like number of layers, choice of optimizer)
* Getting the training data is often the hardest part

## Outline what the Keras library is used for and its relationship to TensorFlow. (5m)

A tensor is an algebraic object that describes a linear mapping from one set of algebraic objects to another.

TensorFlow is a framework which is developed in highly efficient C++ Code with easy to use Python API for matrices calculations. It helps in parallel execution of arithmetic calculations which required for deep learning and facilitated by the GPU’s processing architecture. But they were not as per se designed as deep learning end to end platform

Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano.

Advantages of Keras are:

1. Allows for easy and fast prototyping (through user-friendliness, modularity, and extensibility).
2. Supports both convolutional networks and recurrent networks, as well as combinations of the two.
3. Runs seamlessly on CPU and GPU.
4. Reduces the size of the codebase for deep learning projects.

Answer 2-

1. Keras is one of the leading high-level neural networks APIs.

2. It is written in Python and supports multiple back-end neural network computation engines.

3. Keras was created to be user friendly, modular, easy to extend, and to work with Python.

4. Keras has support for [convolutional](https://en.wikipedia.org/wiki/Convolutional_neural_networks) and [recurrent neural networks](https://en.wikipedia.org/wiki/Recurrent_neural_networks). It supports other common utility layers like dropout, batch normalization, and pooling.

5. [Keras follows best practices for reducing cognitive load](https://blog.keras.io/user-experience-design-for-apis.html)

6. This makes Keras easy to learn and easy to use. As a Keras user, you are more productive, allowing you to try more ideas than your competition, faster -- which in turn [helps you win machine learning competitions](https://www.quora.com/Why-has-Keras-been-so-successful-lately-at-Kaggle-competitions).

Relationship to Tensorflow- There are several differences between these two frameworks. Keras is a neural network library while TensorFlow is the open source library for a number of various tasks in machine learning. TensorFlow provides both high-level and low-level APIs while Keras provides only high-level APIs.

## Outline the steps that could be taken to avoid overfitting in a neural network. (5m)

### Dropout

Randomly drop units (along with their connections) during training

Each unit retained with fixed probability p, independent of other units

Hyper-parameter p to be chosen (tuned)

### L2 = weight decay

Regularization term that penalizes big weights, added to the objective

Weight decay value determines how dominant regularization is during gradient computation

Big weight decay coefficient → big penalty for big weights

### Early-stopping

Continuous checking of validation error. Use validation error to decide when to stop training.

Stop when monitored quantity has not improved after n subsequent epoch.

Answer 2-

Ans: Training a deep neural network that can generalize well to new data is a challenging problem. A model with too little capacity cannot learn the problem, whereas a model with too much capacity can learn it too well and overfit the training dataset. Both cases result in a model that does not generalize well.

There are two ways to approach an overfit model:

1. Reduce overfitting by training the network on more examples.
2. Reduce overfitting by changing the complexity of the network.

* Regularization terms added to cost function during training
* Dropout – ignore say 50% of all neurons randomly at each training step
* Trial & error is one way

- Evaluate a smaller network with less neurons in the hidden layers

- Evaluate a larger network with more layers

- Try reducing the size of each layer as you progress – form a funnel

## What approaches may be taken to tune the topology of a deep learning neural network. (5m)

There are several ways to derive and tune the topology for a deep learning problem.

* The easiest way is to use “model zoos”, based on the domain and class of problem we can check the previously experimented models and their success rate to decide what may be good topology for the current problem.
* Evaluate a smaller network with fewer neurons in the hidden layers.
* Evaluate a larger network with more layers.
* Reducing the size of each layer as you progress to form a funnel, this creates a better feature extraction in initial layers.
* More layers can yield faster learning just
* use more layers and neurons than required and use early stopping.

## Explain how backpropagation works to train MLP’s weights. (5m)

Gradient Descent using reverse-mode auto diff!

For each training step:

1. Compute the output error
2. Compute how much each neuron in the previously hidden layer contributed
3. Back-propagate that error in a reverse pass
4. Tweak weights to reduce the error using gradient descent

## “Deep Learning raises a number of ethical challenges” Discuss (10m)- same as 45