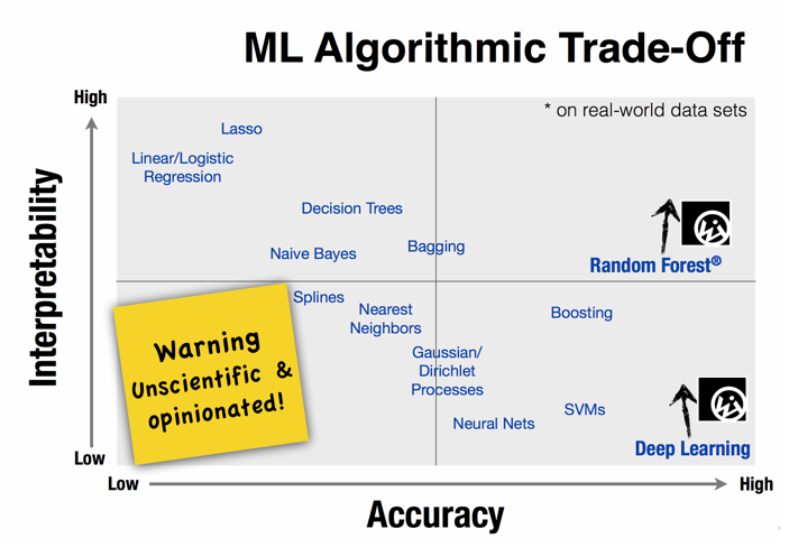
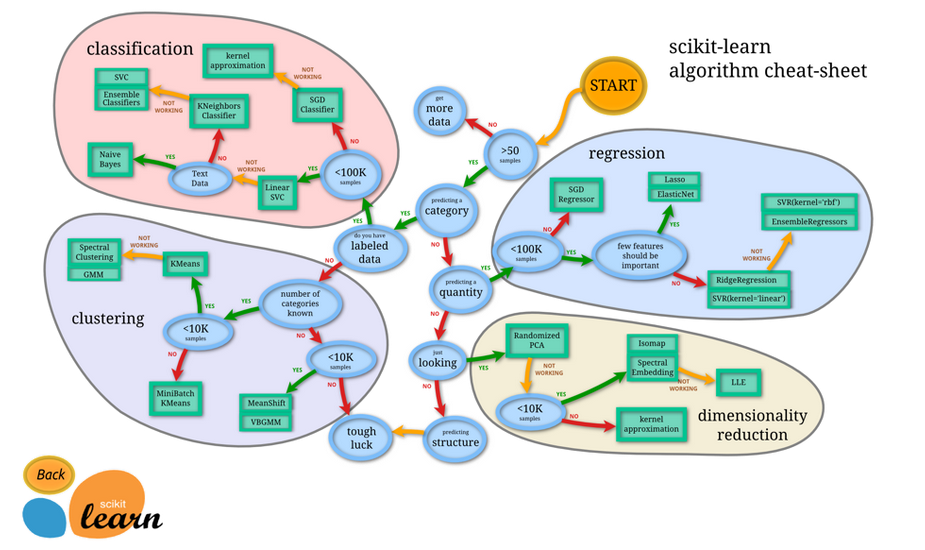
Question 1.

**Ans:**

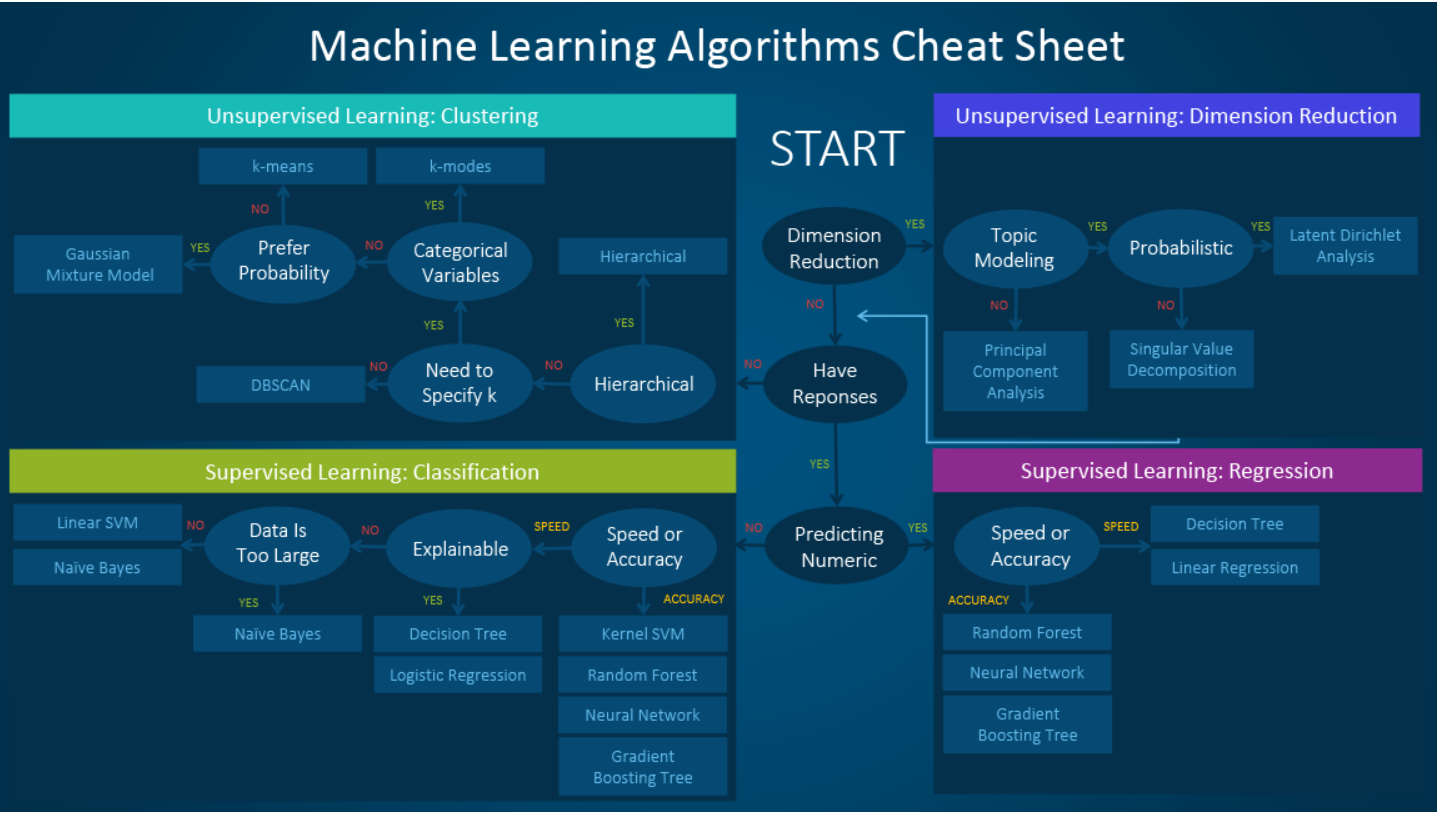
* Logistic Regression
  + Care about interpretation more than accuracy
* Support Vector Machines
  + Care only about accuracy and not about interpretation
  + Scoring throughput is high
* Bayesian classifier
  + Care about interpretation more than accuracy
  + Want to build a robust and deterministic model
* K-Nearest Neighbour
  + Care about interpretation more than accuracy
  + Want to visualize data that has no class labels
* Principal Component Analysis
  + Don’t want to build a model at all
  + It is a Feature selection technique
* Decision Trees
  + Care about interpretation more than accuracy
  + Each class has potentially multiple sub-classes



**Fig**: ML algorithms and their interpretability and accuracy



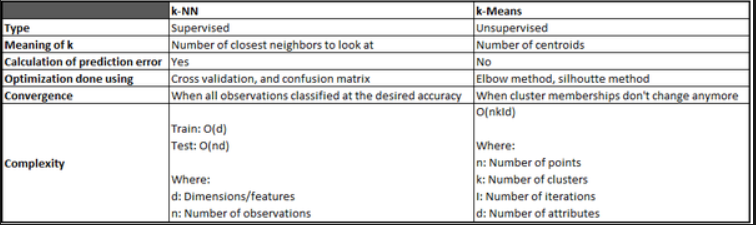
**Fig:** Process of solving problems in ML

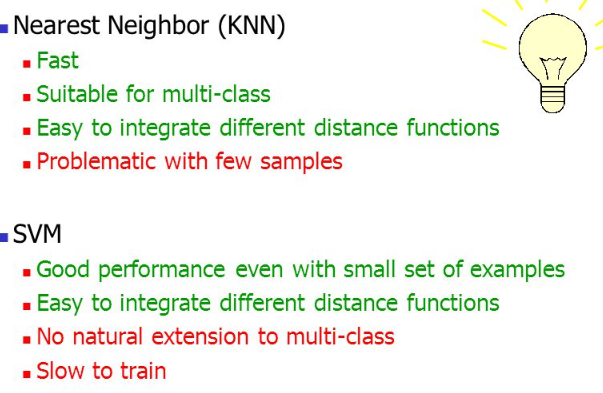


**Fig:** ML algorithms cheat sheet

|  |  |
| --- | --- |
| Classification of Models based on | What to use? |
| I/P →  Labelled data →  Unlabelled data →  Optimize objective function → | Supervised Learning  Unsupervised Learning  Reinforcement Learning |
| O/P →  Number →  Class →  Set of input groups →  Detect anomaly → | Regression Problem  Classification Problem  Clustering Problem  Anomaly detection |

|  |  |
| --- | --- |
| Models / Methods | When to use? |
| Linear Regression | * To predict continuous values(numeric values) |
| Logistic Regression | * When o/p variable is discrete/categorical * To perform binary classification * When we don’t want to worry about whether the feaures are correlated or not. * Want to have probabilistic interpretation and easily update the model when taking new data * To understand the contributing factors behind the prediction. |
| Decision Tree | * Features are non-parametric * Don’t want to worry about outliers or if data is linearly seperable   Not to use   * When new data comes, need to build the tree again. |
| K-Means | * When don’t know any labels and need to label the data based on features * To divide large group of data and need to divide them in groups * To organized / grouping / concatenating on particular groups |
| PCA | * To do dimensionlaity reduction * When having wide range of features, which are higly correlated between each other * When model can easily overfit on a huge amount of data * To find the a synchronized low-dimensional representation of the variables. * Synchronized sample and variable representation → visually find variables that are of a group of samples. * Remove noise * When the problem has too many features and too few training sample. For that PCA is done as feature selection technique. |
| SVM | * pattern recognition and classification problems * Want to have high accuracy   Disad   * memory-intensive, hard to interpret, and difficult to tune. |
| Naive Bayes | * Fast and easy * Very large datasets * When CPU memory resources are a limiting factor |





**Question 2.**

**Ans:**

i).

|  |  |
| --- | --- |
| Domains | Types of Problem |
| news.google.com | Classifier Problem(types of news, news based on the area or country of a person, in that language)  Recomendation Problem(Types of news personalized for the user) |

**Explanation:**

Google news contains various classes of news, which are clasified. Then they also send news recomendation which is personalized according to the users. Then, they show the news related to the user, top five or 8 news.

ii).

|  |  |
| --- | --- |
| Domains | Types of Problem |
| youtube.com | Recommendation Problem  Classification problem  Retrieval Problem |

**Explanation:**

It uses recomendation system to give recomendation to their users according to their locations, dialect and country. As we can see that the youtube show us the videos which are related to the domain where a person is living. Another problem solved by youtube is that they Retrieve information about the user and give suggestions to the users, based on his/her past history of watched videos.

It also uses Classification problem which will help to classfies the videos according to their types, such as songs, movies, songs, etc.

iii).

|  |  |
| --- | --- |
| Domains | Types of Problem |
| linkedin.com | Recommendation Problem  Classification Problem  Retrieval Problem |

**Explanation:**

This uses solve retrieval problem by collecting information about the candidate and then solve classification problem to sort out the candidate needed by the company. So, in future when their is demand, then HR can sort out all eligible candidates from a large pool of people.

LinkedIn also solves the probem of recommendation as they also give recommendation to the employees about the new / upcoming jobs or qualified candidated for hiring.

iv).

|  |  |
| --- | --- |
| Domains | Types of Problem |
| ola.com / uber.com | Regression Problem  Retrieval Problem |

**Explanation:**

These company heavily rely on the retrieval of information as it doesn’t want to ask user about the different features to book a ride. So, when we book a ride, it send alert to all the drivers near to the pick-up location. They also do solve regression problem as they predict the cost of riding from one place to other depending on the timing, weather and traffic.

v).

|  |  |
| --- | --- |
| Domains | Types of Problem |
| 99acres.com / magicbricks.com | Regression Problem |

**Explanation:**

As these are dependants on predicting the value of the location of the plot. Based on the location of the problem, the cost of the plot will vary. So, they rely on predicting the cost of plot based on the various factors which helps to predict the cost of the plot.

**Question 3.**

**Ans.**

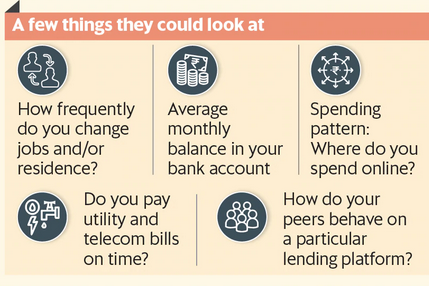
|  |  |  |
| --- | --- | --- |
| Model / Algorithms | Similarity | Dissimalirity |
| K-nearest Neighbour and K-Means | Both methods include calculating the distances in input space and assigning data points to a set of nearest 'prototype points'. | KNN is a supervised learning model, while K-Means is unsupervised learning model.  K-Means is clustering algorithm, whereas KNN is clasification algorithm.  In KNN, k means number of closeset neighbour to look at. In K-Means, k means number of centroids.  Complexity:  KNN: O(nd) [TEST]  , where n= no. of observation, d = dimensions  K-Means: O(nkid)  , where n = no.of points, k = number of clusters, i = number of iteration,  d = number of attributes |
| PCA and MDS | Techniques of Dimensionality Reduction | PCA → Linear data,  retains m important dimensions, is a method, as part of mapping PCA is particular case of MDS,it is a data reduction technique which is more than only mapping.  MDS → Non-Linear data, it is a class, is only mapping technique |
| Classification and Regression | Both are supervised Learning techniques | Classification→ data is descrete/ categorical,  Regression → data is continuous |
| Bag-of-words and market Basket Analysis | In both we take count of the words appearing in the input | Bag of words is a feature extraction procedures for sentence and documents.  Market Basket analysis is the analysis of how likely a product is being given we few items in our basket. |
| Naïve Bayes and Decision Trees | Both can deal with noisy data.  Both works on numerical and categorical data types.  Both are simple to understand and build. | Decision tree will select the features to build a model, whereas in Naive Bayes we have to give selective features.  Decision trees work better with lots of data compared to Naive Bayes. |

**Question 4.**

**a).**You are a new Fintech start up in town. You build credit models based on people’s SMS data (financial related SMS’s only). What kind of features will you engineer that will help you predict credit score of your customers?

**Ans.**

In my opinion following things will help me to predict credit card score of customers:

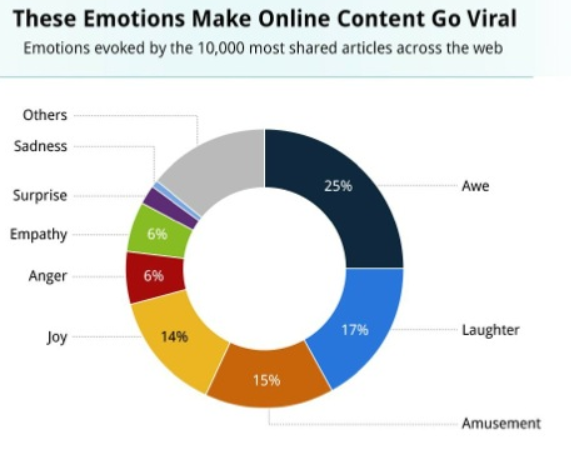


**Fig:** Things needed to determine informal credit card score

1. If user responding to the new offer rolled out.
   1. It will help to find previously invisible customers.
   2. If s/he buy the offer or not, it will create a data that will be helpfull to analyze the pattern / behaviour of the person.
2. Spending pattern of the customer
   1. It will be helpful to find out the spending behaviour of the customer, which will helpfull to give a credit score.
3. E-commerce spends of the person
   1. If a person is buying anything online, s/he leaves digital footprints, that can be analysed and helpful in find the person spending behaviour or pattern.
   2. This will also be helpful to determine the credit card score of the person.
4. Checking information from social media platforms, travelling history
   1. This is very helpful to find out the informal credit card score, as to get trace of person activities on social media and travelling history it’s easy to predict the credit score of that person.
   2. For example, if a person posts of travelling to a new place every weekend/month, then it means that the person is rich and pretty much s/he is employed.

**b).**Let’s say WhatsApp hires you to build the following model. WhatsApp wants to find the most “popular” people in its customer base for viral marketing. What features on each person you will you use to come up with a popularity score of a person on WhatsApp. (Note: A person might be part of multiple groups)

**Ans.**



**Fig:** This figure shows that contribution of emotions in viral marketting.

Following are the features on each person you will you use to come up with a popularity score of a person on WhatsApp:

1. The person who created past viral videos
   1. It will be helpful to dig deep in the history of last viral video maker
   2. Dig out the common characteristics and finding out similar people using WhatsApp.
2. A person whose videos, gets reply as emoji of awe, laughter, amusement and joy.
   1. Such person, who gets most of these emotions are easily determined as most effective person. Such person connects to wider audience even if they don’t live near them.
   2. Emotions are very strong side of every human being. We connect to other person through our emotions and empathy. So, this is important feature in determining influential person rather than be sure about being them as very popular.
3. A person who is knowned for his charisma
   1. Finding such person can be possible by looking for the characteristics like whether a person mentioned more in shared videos, or a person who makes small, interactive videos regularly.
4. A person who is part of many groups in a particular area.
   1. Finding such person can be possible is by looking for the person who has maximum number of people in the group and they are very active on the WhatsApp.