Assignment Title **Machine Learning**

Submitted To. **Dr. Asim Noor**

Submitted By. **Afzaal Ahmed Chughtai**

Registration **FA19-RCS-007**

Date **12-01-2021**

Department **Computer Science**

Comsats University Islamabad
Pakistan

Machine Learning

Q. Nr. 1: In your own words explain machine learning. How it works. What are limitation of machine learning, what are main challenges is machine learning and how those are solved?

Machine learning is a type of artificial intelligence that enables self-learning from data and then applies that learning without the need for human intervention. Its **algorithms allows computer programs to automatically improve through experience to generate more accurate results.** Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches:

Supervised learning: In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified. **Unsupervised learning:** This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. Both the data algorithms train on and the predictions or recommendations they output are predetermined. **Semi-supervised learning:** This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set. **Reinforcement learning:** Reinforcement learning is typically used to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

Machine learning algorithms learn, but it's often hard to find a precise meaning for the term learning because different ways exist to extract information from data, depending on how the machine learning algorithm is built. Generally, the learning process requires huge amounts of data that provides an expected response given particular inputs. Each input/response pair represents an example and more examples make it easier for the algorithm to learn. That's because each input/response pair fits within a line, cluster, or other statistical representation that defines a problem domain. Machine learning is the act of optimizing a model, which is a mathematical, summarized representation of data itself, such that it can predict or otherwise determine an appropriate response even when it receives input that it hasn't seen before. The more accurately the model can come up with correct responses, the better the model has learned from the data inputs provided. An algorithm fits the model to the data, and this fitting process is training.

What are limitation of machine learning?

The benefits of machine learning translate to innovative applications that can improve the way processes and tasks are accomplished. However, despite its numerous advantages, there are still risks and challenges. Take note of the following cons or limitations of machine learning:

Error diagnosis and correction: One notable limitation of machine learning is its susceptibility to errors. Brynjolfsson and McAfee said that the actual problem with this inevitable fact is that when they do make errors, diagnosing and correcting them can be difficult because it will require going through the underlying complexities of the algorithms and associated processes.

Time constraints in learning: It is impossible to make immediate accurate predictions with a machine learning system. Remember that it learns through historical data. The bigger the data and the longer it is exposed to these data, the better it will perform

Problems with verification: Another limitation of machine learning is the lack of variability. Brynjolfsson and McAfee said that machine learning deals with statistical truths rather than literal truths. In situations that are not included in the historical data, it will be difficult to prove with complete certainty that the predictions made by a machine learning system is suitable in all scenarios.

Limitations of predictions: Brynjolfsson and McAfee reminded that unlike humans, computers are not good storytellers. Machine learning systems cannot always provide rational reasons for a particular prediction or decision. They are also limited to answering questions rather than posing them. In addition, these systems does not understand context. Depending on the provided data used for training, machine learning is also prone to hidden and unintentional biases. Human input is still important to better evaluate the outputs of these systems.

There are certain **challenges** an ML practitioner might face while developing an application from **zero** to bringing them to **production.**

1. Data collection
2. Less amount of training data.
3. Non-representative Training Data
4. Poor quality of data\Unwanted features
5. Overfitting of training data
6. Under fitting of training data

Q. Nr. 2 In this course we have studied following classification algorithms: Logistic regression, Neural Networks, and SVM. • Step by step explain how each algorithm works • under what circumstance one algorithm is preferred over other (strength and weaknesses) • How would you compare performance of these algorithm which metrics you will use, how you will interpret these metrics?

a) **Logistic regression**

Logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

How Logistic regression work

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a l*ogit*, from *logistic unit*, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probity model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

b) **Neural network**

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems.

How Neural Network Work

A neural network works similarly to the human brain's neural network. A "neuron" in a neural network is a mathematical function that collects and classifies information according to a specific architecture.

The network bears a strong resemblance to statistical methods such as curve fitting and regression analysis.

A neural network contains layers of interconnected nodes. Each node is a perceptron and is similar to a multiple linear regression. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear.

In a multi-layered perceptron (MLP), perceptrons are arranged in interconnected layers. The input layer collects input patterns. The output layer has classifications or output signals to which input patterns may map. For instance, the patterns may comprise a list of quantities for technical indicator about a security; potential outputs could be “buy,” “hold” or “sell.”

Hidden layers fine-tune the input weightings until the neural network’s margin of error is minimal. It is hypothesized that hidden layers extrapolate salient features in the input data that have predictive power regarding the outputs. This describes feature extraction, which accomplishes a utility similar to statistical techniques such as principal component analysis.

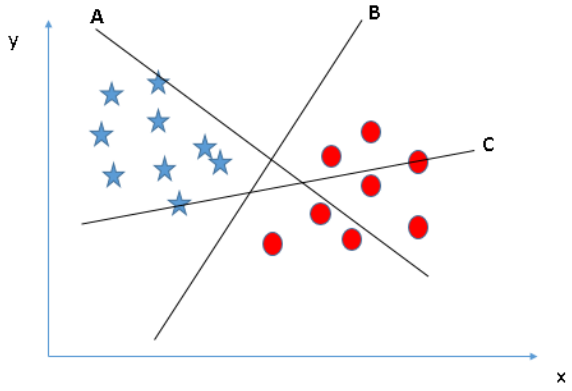
c) **SVM**

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

How Support Vector machine works?

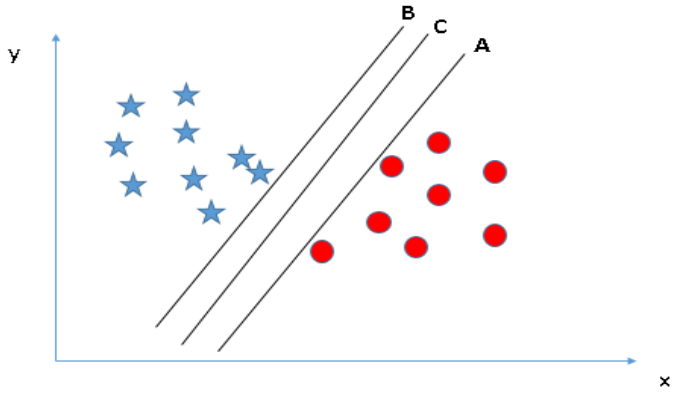
Scenario 1:

Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.

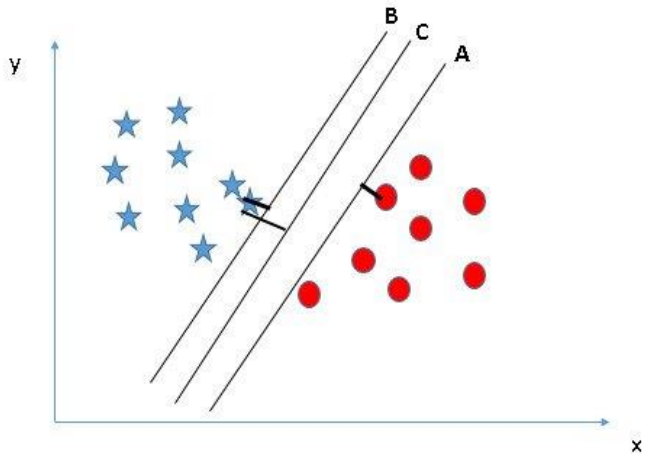


You need to remember a thumb rule to identify the right hyper-plane: “Select the hyper-plane which segregates the two classes better”. In this scenario, hyper-plane “B” has excellently performed this job.

Scenario 2: Here, we have three hyper-planes (A, B and C) and all are segregating the classes well. Now, How can we identify the right hyper-plane?

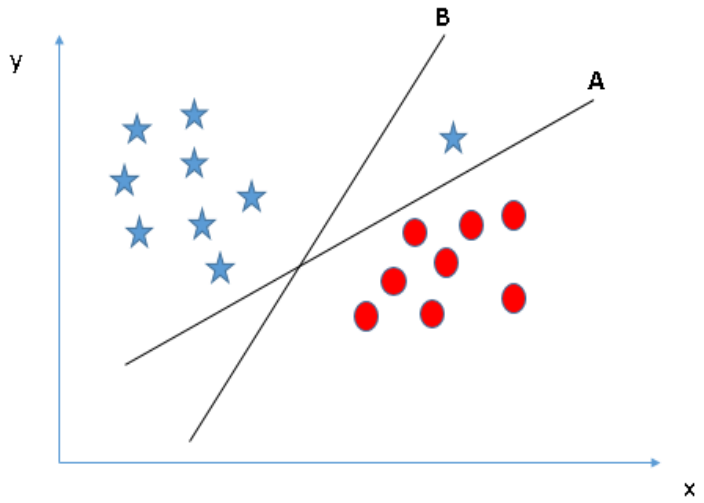


Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as Margin. Let's look at the below snapshot:



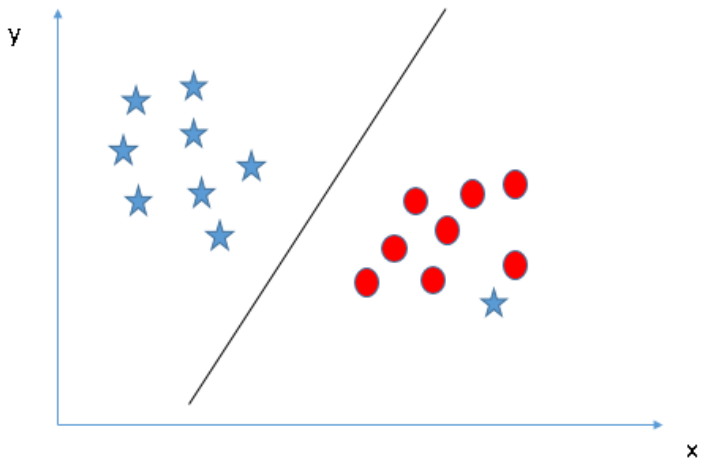
Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

Scenario 3:**Some of you may have selected the hyper-plane** B **as it has higher margin compared to** A. **But, here is the catch, SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is** A.

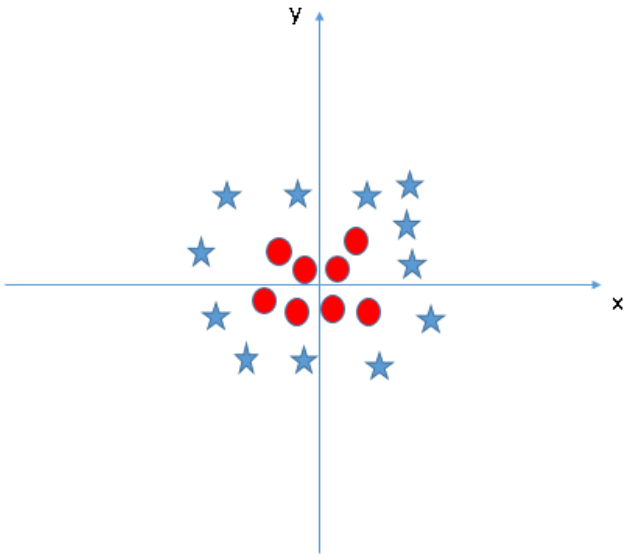


Scenario 4:**Below, I am unable to segregate the two classes using a straight line, as one of the stars lies in the territory of other(circle) class as an outlier.**

As I have already mentioned, one star at other end is like an outlier for star class. The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Hence, we can say, SVM classification is robust to outliers.

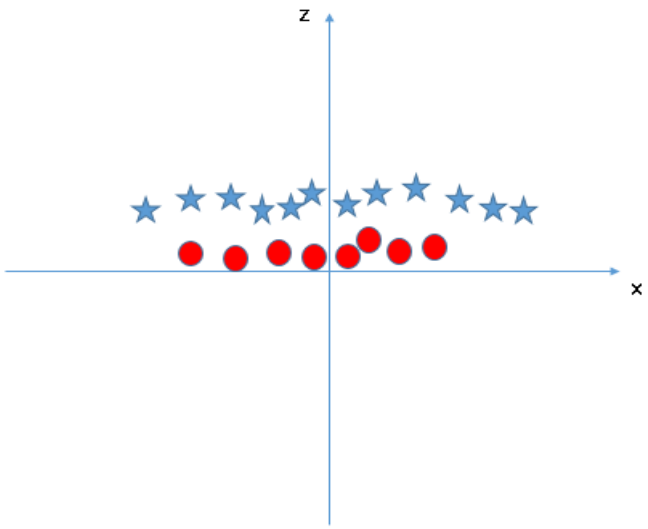


Scenario 5: In the scenario below, we can't have linear hyper-plane between the two



classes, so how does SVM classify these two classes? Till now, we have only looked at the linear hyper-plane.

SVM can solve this problem. Easily! It solves this problem by introducing additional feature. Here, we will add a new feature z=x^2+y^2. Now, let's plot the data points on axis x and z:



Under what circumstance one algorithm is preferred over other (strength and weaknesses)

SVMs are really good when you have a high dimensionality dataset and you don't have a lot of data. With or without a kernel they are not very likely to overfit and produce good results.

Logistic Regression is a very good all-purpose algorithm, if you need probabilities or you have a lot of data LR is usually good. Same if you have only a few features.
NNs are very flexible you usually need a lot of data and they are particularly useful for data such as sound, images, video and other multimedia data.
So in general:
1)reduced number of features => LR
2)a lot of features but not a lot of data => SVM
3)a lot of features and a lot of data => NN
Of course that is an over-simplification and we'll find plenty of counter-examples but I think it is sound as a rough guideline.

How would you compare performance of these algorithm which metrics you will use, how you

will interpret these metrics

SVMs are really good when you have a high dimensionality dataset and you don't have a lot of data. With or without a kernel they are not very likely to overfit and produce good results.

Logistic Regression is a very good all-purpose algorithm, if you need probabilities or you have a lot of data LR is usually good. Same if you have only a few features.
NNs are very flexible you usually need a lot of data and they are particularly useful for data such as sound, images, video and other multimedia data.

Q. Nr. 3

Solve attached assignment titled as K-means Clustering and Principal Component Analysis (Ex7)Download assignment package from following url: https://drive.google.com/file/d/1AB1Lx0QXTFOhUapdaHCtr8I7-_w3lKab/view?usp=sharing

Answer Attached

Q. Nr. 4

Solve attached assignment titled as Anomaly Detection and Recommender System (Ex-8). Download assignment package from following

url: https://drive.google.com/file/d/1F4itfynGibfPlcLQVHaqE8g2fEM2jev9/view?usp=sharing

Answer Attached