

Assignment – 01

- Q1. I am not actively working on any project now but I have been thinking of a model that can predict students dropout, a model that can predict their future career path. I am not sure about heuristic design choices.
- Q2. I am thinking of a use-case that can find the similarity between two documents written in different languages. For instance consider documents written in German and Hindi. One way is to translate all Hindi documents into German (or vice-versa) and then estimate the similarity between them. Other way is to translate all documents into English and find the similarity. However the complexity may be reduced if we can avoid the translation.
- Q3. Data is considered as an abstraction of a part of the real world. Machine learning models are trained to gain intelligence from the selected set of data, the set that the engineer considers relevant to the problem at hand and believes the desired results can be achieved. This process of selecting the appropriate set of data that properly represents the problem at hand is fairly difficult. If right data is not available, however good the algorithm is, a perfect model cannot be trained. So it is not just about more data but right data.
- Q4. End to end training approach can be used to solve problems in domains where large amounts of training data is available and high validation is not mandatory. Machine translation, speech recognition, speech to text translation are the various areas where I think end-to-end training approach can be used.

Machine learning (ML) is all about “learning from data”. Tom Mitchell defines Machine Learning as *the study of computer algorithms that improve automatically through experience*. Whatever the ML problem we work on, the core components will be the same; Data, Model, Loss Function and the Learning Algorithm. ML models take in data of one type and make predictions of a possibly different type. Although there are many models, we see a rapid progress in statistical models, applications and algorithms in recent times. Loss function is a formal measure of how well the model is. Squared error is the most common objective function which is also easy to optimize. Any ML model uses a learning agent, this learning agent uses a learning algorithm to learn from data by trying to minimize the loss function. Gradient descent is the most popular optimization algorithm.

Main aim is to make machines intelligent and how do we make it? They should be able to learn on their own. Broadly there are three types of learning (i) Supervised Learning (ii) Unsupervised Learning (iii) Reinforcement Learning.

Supervised Learning, as the name suggests, it is learning under supervision. Underlying concept behind this is labeled data. Label is something particular about the data that we want to predict. For e.g. if we want to differentiate images of cats and dogs, collect a bunch of images of both these animals, feed all of them to a model that tries to learn the difference. Now this model uses the labeled dataset to explicitly learn the difference between cats and dogs. If a new image is fed to the model and it makes a correct prediction then the loss function tells the model to reinforce its learning, otherwise the model tries to learn from the experience. Consider another e.g., in Netflix for each movie, the genre, duration, cast and crew, and other details are defined. For each movie we already know if the user liked that movie or not. Based on that, ML model will try and predict the new movies that the user is most likely to enjoy watching. Supervised learning can be used to solve different problems like classification, regression etc.

Unsupervised Learning is learning without supervision. There is no labeled data. The model tries to find the patterns in data. We use unlabeled data because it takes lot of resources to label data, the process of labeling can be noisy and error prone due to human error or the biases people labeling the data might have. Moreover labels make sense when we know exactly what we want to predict. Its not the case always, sometimes we may want to discover new patterns in our data. Unsupervised learning can be used to solve different problems like clustering. Cluster formation is based on similarities in data. Whenever we want to study data and find patterns in that, we use unsupervised learning.

Reinforcement Learning is reward based learning. It is like giving a toffee to the child whenever he gets something right and punish him when he does something wrong (although we should never do that). In reinforcement learning, the agent not just make predictions but also choose actions that impact the environment. Based on the impact of the action, agent receives a reward from the environment. Mostly used in game playing. In the game of chess, we associate the state of checkmate with highest reward and different states with lower rewards. So the model tries to select actions that maximize the rewards it get.

Huge amounts of data is made available from internet applications that serve hundreds of millions of users online, cheap and high-quality sensors, by significant progress in computation and mostly through GPUs. It is evident that RAM has not kept in pace with growth in data and increase in computational power has outpaced that of the data available. This means statistical models need to become more memory efficient, so they moved from linear models and kernel methods to deep networks. Few ideas that helped researchers achieve tremendous progress in this area over the last decade are dropout, attention mechanism, multistage designs, GANs, distributed machine learning systems, deep learning frameworks etc.