<u>ENGG 680 Assignment 5</u>

Name: Deep Vyas

UCID#: 30139014

Date: Mar 24, 2022

Q1: What are the two main types of learning systems? Describe each type with its main areas of applications and the most popular learning tasks used for those tasks.

➔ There are mainly two types of learning systems,

1) **Supervised Learning:**
With supervised learning, you can predict the outcome of unlabeled data by using labeled training data. A team of highly skilled data scientists needs to devote time and technical, scaling, expertise to building and deploying an accurate, high-performing supervised machine learning model. Furthermore, data scientists must rebuild models to make sure insights remain valid until the data changes. It is possible to collect data or produce data based on previous experiences through supervised learning.
Applications:
- House Pricing Prediction
- Image Recognition
- Weather Prediction
- Feedback Classification

2) **Unsupervised Learning:**
The technique of unsupervised learning involves not supervising the machine learning model. Instead, you allow it to discover information on its own. When compared to supervised learning algorithms, unsupervised learning can handle more complex processing tasks. In comparison to deep learning and reinforcement learning, unsupervised learning can yield more unpredictable results.
Applications:
- Finding customer Segments
- Reducing the complexity of a problem
- Feature selection

Q2: For each of the following scenarios, determine the type of machine learning task required:

a. Segmenting your customers into multiple groups - **Unsupervised Learning**

b. Making a robot to walk in various unknown terrains - **Reinforcement Learning**

c. Detecting spam emails - **Supervised Learning**

Q3: What is model-based learning and what are the main steps taken by a model-based learning system?

➔ Each new application is designed to be tailored to a model-based approach. Instead of having to transform a problem to fit some standard algorithm, model-based machine learning offers a custom algorithm specific to the problem. In model-based machine learning, the assumptions about the problem domain are made explicit through models.

- ➔ Model-Based Learning Steps:
  - Study the Data
  - Select a model
  - Train model on training data
  - Test model on testing data
  - Apply the model to predict new unseen data

Q4: Describe four main problems prevalent in the realm of machine learning and describe the main methods currently used to combat these issues.

- ➔ There are some below challenges in machine learning

1) **Insufficient Quantity of Training Data:** Using data repositories that contain massive amounts of data is often recommended when building Machine Learning models. However, a simple Machine Learning model can be built from only a few hundred records, although generally true. Rather complex systems, however, do require large datasets. Image Recognition models, for example, may require millions of records.
   The below technique is used to overcome this issue.
   a. Model Complexity
   b. Transfer Learning
   c. Data Augmentation
   d. Synthetic Data

2) **Nonrepresentative Training Data:** Having an accurate representation of the population in the training data is crucial for generalizing the model. Even if the sample size is large, you may still end up with sampling noise, which is unbiased data based on chance. However, even a large sample can be unrepresentative of the sampling method is flawed. This is known as Sampling Bias.
   To overcome this issue, one should select a training set that represents the cases you wish to train. The most common method process used in practice by ML practitioners is known as Bias-Variance Trade-off which finds a sweet spot for low Variance and Bias.

3) **Poor-Quality Data:** If the data that is being fed into the model is already dirty, then the model will not be able to perform well. There is typically some amount of error, outlier, and noise in dirty data. Missing observations can also cause data to be dirty. To solve this problem predictive models are trained by cleaning data before it is used for predictions by Data scientists.

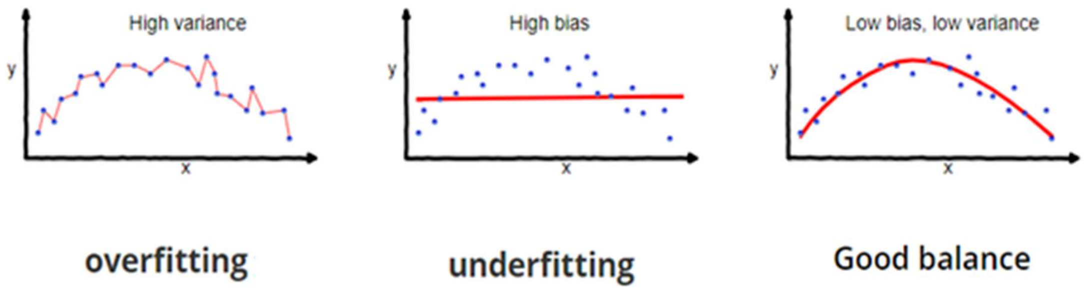
4) **Irrelevant Features:** If you feed bad data to a model with lots of features, then the model will ultimately return garbage. In practice, Feature Engineering is used to overcome this issue. Feature Engineering involves the below process.
   a. Feature Selection
   b. Feature Extraction

Q5: What is cross-validation and why does it become important during model training?

- ➔ Validating the stability of a machine learning model is vital whenever you develop one. It is important that the model you use has captured most of the patterns from the data and does not pick up too much on the noise, that is, it has low bias and variance. The model should be stable if it predicts with high accuracy even on unseen data and if it is consistent for a wide range of input data.

- ➔ **Overfitting**: Because your model was trained with the given data, it understands the data quite well, captures even minute variations(noise), and is generalizing quite well over the given data. In the case of completely new, previously unseen data, the model might not be able to generalize based on that data and might not predict as accurately as it has in the past. This problem is called Overfitting.

➔ **Underfitting**: It happens occasionally that the model is unable to find patterns in the training set, and thus it fails to train. This would not work well on the test set in this case. This problem is called Underfitting.



Using a technique called Cross-Validation, we can know whether our model is over-fitting.

Q6: Explain the concept and importance of regularization in machine learning systems.

➔ In this method, extra information is added to the model to prevent it from becoming overfit. By using this technique, we can avoid the problem of overfitting by learning a simpler or more flexible model. By regularizing, independent variables are reduced in magnitude while keeping the same number of variables.

➔ **Why Regularization?**
Machine learning models sometimes perform well when trained on data but not when tested. Overfitted models are unable to predict the output for unseen data since there is noise in the output, and thus overfitting results in the model not being able to predict the output. Noisy data points are those which do not necessarily reflect your data's true nature but are instead due to chance. Therefore, regularization techniques are used to prevent overfitting.