Assignment 2 Solution – Applications of Machine Learning in Industries

Question 1. Discuss in detail the different ML techniques that are used in legend based automatic drug discovery process.

Solution1. Machine Learning algorithms have made a significant impact in drug discovery. Pharmaceutical companies and scientists have benefitted a lot from the rise of various ML algorithms, which can be used in drug discovery process. There are a lot of ML models which have been developed for predicting chemical, biological, and physical properties of a compound which would be of great assistance in this industry. These ML algorithms can be included in all the steps of the drug discovery process. For instance, they can be used to find new uses of old drugs, predict drug-protein reaction, and optimizing the bioactivity of molecules. Some of the algorithms which are highly useful are Random Forest, Naïve Bayes, and Support Vector Machines.

The properties of ML which are used even before the model are also important and require a brief discussion. So, basically there are 2 types of ML algorithms to consider here, supervised, and unsupervised. I would not go into the details of this, but these two methods can be combined as semi-supervised and reinforcement learning, where all the properties of the two methods can be utilized in various situations. For more accuracy and to decrease the false-positive rate, which is the most crucial part in the field of medical science, huge amounts of data is required for development and evolution of models. With the increasing data, processes like dimensionality reduction are important and significant. Interpretation of multi-dimensional data is also important and brings processes like principal component analysis into the picture as well.

Now I would like to discuss about the various ML algorithms that I earlier wrote about:

1. Random Forest: This algorithm is used for dealing with large datasets having multiple features. It is useful for negotiating with outliers, missing data and finding important features that influence the classification. The process of classification consists of several decision trees working together as an ensemble and each tree responsible for one prediction. The class with the most votes is the classified class. With the use of multiple

decision trees, the errors are minimized as there are a lot of trees, and not a single one, and all consider multiple factors. This concept of multiple trees also gives the name random forest. Coming to the drug discovery domain, this algorithm is mostly used for feature selection. The important factors utilized here are:

   a. It uses less parameters.
   b. It speeds up the training process.
   c. Represents missing data.

2. Naïve Bayes: Classification of biometric data is important and significant in drug discovery process, and more often than not this data is filled with non-related information, where Naïve Bayes has proved to be excellent. This algorithm is great for predictive modeling classification. The quality of Naïve Bayes that researchers have utilized in drug discovery process is that it is quite tolerant to random noise in the data. It also plays an important role in predicting ligand-target interactions.
3. Support Vector Machine: This is used in drug discovery process where the class of compounds are needed to be separated based on a feature. SVM quality is that it can differentiate between active and non-active compounds. It can also rank compounds form the drug database. It separates the classes consisting of compounds based on selected features and projects them into a chemical free space. A classification is either positive or negative, depending on the position it attains from the perspective of the hyperplane. This process ranks the compounds from most selective to least selective or vice versa. SVMs area also used to predict drugs that might have multiple bioactivities.

Question 2. Compare Statistical and Neural Machine Translation system in detail.

Answer 2. Let us talk about statistical machine translation first. This approach uses statistical methods which are based on analysis of bilingual text corpora. If we talk about the most basic approach, it is about probability. Let us consider a sentence B in the target language. Now, to train the model what we do is, we choose the best sentence A from the source language, which is most suitable to produce B when translated. Hence, in the process we maximize the probability P(A|B).

Therefore, we come to the conclusion that a standard statistical translation model requires the following:

a. Language model: To find the correct word in a given context.
b. Translation model: To find the best translation of a given word.
c. Method to find the correct order of the words in a phrase or a sentence.

Programs that have been using statistical machine translations:

a. Google translate (till 2016)
b. Microsoft translator (till 2016)

Advantages:

a. Reduced manpower from linguistic experts.
b. Once a model is trained, it can be used for multiple language pairs.
c. Improved translation quality if the model is suitable.

Disadvantages:

a. Requires huge amount of data.
b. Missing or incorrect data is difficult to fix.
c. Unsuitable with languages having differences in word order.

Now, let us talk about neural machine translation in detail. As the name suggests, we use neural networks in this approach to achieve translation. With the emergence of sequence models, the possibility of training a network with an input and output sequence became possible. The emergence of neural machine translation was quick, and in no time, they left behind statistical models. Even google and Microsoft switched to neural network-based systems post 2016.

Programs that use neural machine translation-based systems:

a. Google (post 2016)
b. Microsoft (post 2016)
c. Facebook

Advantage: End to end model is build, without any requirement of a pipeline.

Disadvantages: Now, there are quite a few difficulties associated to this.

a. Training data should not be unbalanced. Model will struggle to learn from rare of frequent data samples.
b. In some languages there are a lot of words which are used rarely, which might create a problem.

Some key differences which we can observe is that:

a. SMT models obtains better bleu scores than NMT models.
b. Process of decoding is faster in NMT than SMT based models.
c. NMT models are superior in terms of output fluency.

Question 3.

a. Application of ML based techniques in manufacturing industry for fault assessment.
   Solution. The use of ML based techniques is increasing in manufacturing industry due to advantages in data collections systems and multiple algorithms that are available in these times. If the diagnosis of faults in the manufacturing process is timely, it provides key advantage to the company to stay competitive in the industry and maintain the quick delivery and top quality of the product. Therefore, figuring out ways to detect this is important and the discussion points are:
   1. Bayesian models: BN is a directed acyclic graph whose nodes represent random variables, and their conditional dependencies are depicted by directed arcs linking the nodes. Data sources in Bayesian networks are important and include Quality management systems, manufacturing execution systems, recipe management systems etc. These models are used to model uncertainty and can be used to model and visualize hierarchal levels of causes and effects that have been deduced from predictions. Since these models are resistant to noise, they are very useful in this domain, where their might be noise as most of the data is collected from machines and it is next to impossible to supervise the data.
   2. Support Vector Machines: SVM uses kernel functions such as radial based function or polynomial kernel to find the best hyperplane that separates the data. Radial based

functions are real time functions where values depend on the distance between input and some fixed point. Applications of SVM are mostly in fault localization. Fault localization refers to deducing failures from the set of observed failure indications. As soon as classifications comes into the picture that too with small dataset, SVMs are one of the best models that can be used. Computational time is fast in SVMs and it is also good in modeling linear as well as non-linear relationships.

b. Application of ML based techniques in retail industry.

Solution: In the retail industry, machine learning concepts such as:

1. Predicting customer behavior: Companies and retailers can use this data to understand the need of the customers. Machine learning algorithms can be used to figure out various features which influence customer behavior, which include quality of the product, price of the product, after sale service, behavior and treatment meted out to the customers etc. All these variables can impact customer behavior and machine learning algorithms can prove to be of great help in predicting these and suggesting the improvements required.
2. Marketing behavior: Another aspect of retail is the marketing required for a product and its position in the retail store to the rest of the products. Taking an example from one of the giants in retail industry Walmart. Their beer bottles and baby diapers case study is famous and is a perfect example of use of machine learning techniques in the retail industry. It was observed that on Friday evenings office going men came to purchase diapers as told by there wives. And they used to pick beer bottles as well. So, Walmart made a fortune by putting the beer bottles closed to the diapers.

c. Evaluation of Machine Translation using BLEU.

Solution: Bleu score is a sort of a function or a judging parameter of a translation that takes place from one language to another. Its full form is Bilingual Evaluation Understudy Score. The score was developed for evaluating the predictions made by automatic machine translation systems. It is not perfect, but does some advantages:

1. It is quick and requires less computational power to calculate.
2. It is easy to understand.

3. It is language independent.

4. It correlates highly with human evaluation.

In the NMT project done in Application of Machine Learning in Industries Lab: we used two kinds of Bleu scores available in NLTK library.

I. Sentence Bleu – The sentence is provided as a list of tokens.

II. Corpus Bleu – It calculates bleu score for multiple sentences. A list of source languages sentences needs to be passed along with translated sentences in the targeted language.

Submitted By:

Bharat Goyal,

500068877.