***Student: Victor Pugliese***

***R#: 11492336***

***Course: Independent Study on Python Machine Learning for Petroleum Engineering Application (PETR 5000)***

***Self-Homework #1***

1. ***How would you define Machine Learning?*** Machine learning is an approach for programming in which the program is capable to improve his performance using data from previous experiences.
2. ***Can you name four types of problems where it shines?*** Machine Learning shines for the following problems: the problem is solved using multiple/changing rules, when the problem is too complex for traditional approaches or there is no known algorithm, when new data is constantly updated and when the user wants to have a better insight of the problem.
3. ***What is a labeled training set?*** A labeled training set is the training data you feed to the system, and it includes the desired solution.
4. ***What are the two most common supervised tasks?*** Classification and Regression.
5. ***Can you name four common unsupervised tasks?*** Clustering, Association rule learning, Visualization and Dimensionality reduction.
6. ***What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?*** Reinforcement Learning.
7. ***What type of algorithm would you use to segment your customers into multiple groups?*** Clustering.
8. ***Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?*** Supervised learning problem.
9. ***What is an online learning system?*** It is a system which is trained incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches. Once it has learned about new data instances, it does not need them anymore, so you can discard them.
10. ***What is out-of-core learning?*** Out-of-core learning is an online learning approach used to handle huge datasets that cannot fit in one machine’s main memory. The algorithm loads part of the data, runs a training step on that data, and repeats the process until it has run on all the data.
11. ***What type of learning algorithm relies on a similarity measure to make predictions?*** Instance-based learning.
12. ***What is the difference between a model parameter and a learning algorithm’s hyperparameter?*** The hyperparameter is not affected by the learning algorithm itself; it must be set prior to training and remains constant during training. On the other hand, a model parameter is changing continuously as you are obtaining a better solution.
13. ***What do model-based learning algorithms search for? What is the most common strategy they use to succeed?*** How do they make predictions? Model-based learning algorithms search for predictions. The most common strategy they use to succeed is: Study the data, select a model, Train the model on the training data, and Make predictions on new cases. They make predictions fitting the regression parameters by minimizing a cost function. Once we get the best possible fitting parameter using the training data, the model is used to predict.
14. ***Can you name four of the main challenges in Machine Learning?*** Insufficient quantity of training data, Nonrepresentative training data, Poor quality data, Overfitting/Underfitting the training data.
15. ***If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?*** In this case, we overfit the training data. Three possible solutions are: simplifying the model by selecting one with fewer parameters, gathering more training data, and reducing the noise in the training data.
16. ***What is a test set and why would you want to use it?*** A test set is a set of data that is used to find how well your system will generalize to new cases. You would want to use it because you could have an estimation of the generalization error.
17. ***What is the purpose of a validation set?*** The purpose of a validation set is to tune the hyperparameter.
18. ***What can go wrong if you tune hyperparameters using the test set?*** The model parameters and the hyperparameter are adapted for “that” set. The system could have a large generalization error.
19. ***What is cross-validation and why would you prefer it to a validation set?*** Cross-validation is a technique in which the training set is split into complementary subsets, and each model is trained against a different combination of these subsets and validated against the remaining parts. We would prefer it to a validation set to avoid wasting too much training data in validation sets.