ASSIGNMENT 20.1

Student: K. Anandaranga (Mar 2018 batch)

1. What are the three stages to build the hypotheses or model in machine learning?

1. **Data preparation and exploratory data analysis**
   a. Involves cleaning the data and transforming the data
   b. Exploring the data to understanding it better (descriptive statistics and a few inferential statistics)
   c. Separating the data into training set, validation set and testing set
2. **Algorithm training, evaluation and selection**
   a. Define the problem statement and identify the possible machine learning models that can help answer it
   b. Run specific machine learning algorithms on training dataset and compare predictive performance results; pick the best model(s)
   c. Test them on testing dataset to narrow down to best-fit model
3. **Model fine-tuning, deployment and monitoring**
   - Modify model hyper-parameters for the best-fit model
   - Check in validation dataset for further fine tuning
   - Deploy the final model; continue to monitor the model performance

2. What is the standard approach to supervised learning?

- Label the training dataset (for the known actual output)
- Learning algorithm uses the data inputs to predict its output; this is then compared to the known actual outputs, to calculate the model errors
- Learning algorithm is then modified to reduce the prediction error

3. What is Training set and Test set?

- Training set is the portion of the dataset that is used to train the model and help in fitting the best model
- Test set is a separate dataset that is used to assess the generalized error of the final chosen model

4. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?

- Ensemble method uses a combination of models to arrive at prediction
- Bagging is a method to reduce the variance; uses voting for classification and averaging for regression; generates additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data
- Boosting is used to reduce bias; it adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation on the next iteration

5. How can you avoid overfitting?

- Overfitting occurs when the model models the training data too well that it learns the 'noise' also as valid data signals
- A few ways to avoid overfitting includes –
    - Using more data, but make sure that it is still relevant data
    - Using ensemble methods, since they 'average' out the model results
    - Using simpler methods (over complex ones)
    - Validate that the model does not degrade a lot between training and test set
    - Adding a regularization term
- This is domain-dependent and dependent on the nature of the problem being solved, so practicality and common-sense should also prevail