COMP- 8220 Machine Learning Section A

Suhash Reddy Immareddy

45693242

Question 1. **Describe under which conditions overfitting can happen and provide three possible solutions that can reduce the problem of overfitting.**

Answer: Overfitting (low bias and high variance) Occurs depending on the model parameters that are used to fit the data and the data quality itself such that the model would perform well on the training data but does not generalize well when new data is given. The following are 3 among several solutions:

1) Train the model with less no of features which are important.
2) Using Ensemble Learning meaning: a model on top of another model, for example a bunch of decision trees running in parallel on top is a Random Forest, these provide good results because they take the average of predictions.
3) Cross Validation with hyperparameter tuning.

Question 2. **Suppose the features in your training set have very different scales. Explain what algorithms might suffer from this and what you can do about it.**

Answer: SVM, K nearest Neighbours and K-means are such models that would affect by the features if they are not on the same scale. So we need to scale the features such that they all comes in same scale range, the two algorithms that are used is Min Max Scaling (Normalization lie in range 0 to 1) and Standardisation (lie in range 0 to 1 with mean 0).

Question 3. **Explain when logistic regression should be used and why logistic regression is not called "logistic classification".**

Answer: The Logistic regression is used when the interest is to find the probability with which an Instance belongs to a class. Even though Logistic regression is classifying in this case we cannot call it as classification because of its underlying Statistics rule that are used like threshold techniques, above a certain threshold an instance is said to belongs to class 1 and vice versa.

Question 4. **The generalisation error of a model in machine learning can be expressed as the sum of three very different errors. Name these errors and briefly describe each of them.**

Answer: Three Generalization errors are: where models' complexity is directly proportional to variance and indirectly proportional to Bias. A good model will have low bias and low variance.

1. Bias: the assumptions made by the model on the data are wrong is measured in terms of bias.
2. Variance: models' sensitivity to the smallest variations present in data is measured by variance.
3. Irreducible error: This error is due to the noise present in the data itself.

Question 5. **Describe the fundamental idea behind Support Vector Machines (SVMs) and explain why it is important to scale the inputs when using SVMs.**

Answer: The Basic Idea behind SVMs is to draw a Hyperplane (which is a margin) such that the data will be divided into their respective classes (a line is a hyperplane in 2D, a plane in a 3D). Here SVM tries to maximise the margin between the data points and the plane so, if we have two features with different scales the margin would be affected by the feature with larger values.

Question 6. The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits. Illustrate how Gini impurity is calculated by providing an example.

Answer: Gini Impurity is used to measure the impurity of the node where 0 represents node is pure and vice versa.

For example: let us consider two child nodes where the Instances in two child nodes are [9, 2] and [1, 18] respectively and samples in Child nodes are 11 and 19, respectively. where Gini score is computed by subtracting the ratio of instance from child nodes by samples from 1.

Gini score for child node 1: 1 – (9/11) ^2 – (2/11) ^2 = 1 – 0.669 – 0.033 = 0.2971.

Gini score for child node 2: 1 – (1/19) ^2 – (18/19) ^2 = 1 – 0.0027 – 0.8968 = 0.1005.

Question 7. Name the main motivations for reducing the dimensionality of a dataset and explain what the main drawbacks are.

Answer: the following are the reasons for reducing the dimensionality:

1. When we have large no of features to train the model it would take a lot of time.
2. The larger the features the more the risk that model would overfit the data.
3. In some cases, not all features are relevant, most of the variation can be explained with fewer features.
4. Visualization would be easier if there are less features.

Question 8. Explain how the TF-IDF measure is calculated and name advantages and disadvantages of this measure.

Answer: Term Frequency and Inverse Document frequency is used to find the importance of a word to a document to which it belongs among several documents. It is calculated by multiplying the word frequency in a document to the log of ratio of total documents over no of documents that contain the word. [TF-IDF = TF * IDF]

Advantages: highly efficient

Disadvantages: Since TF-IDF uses Bag of words we cannot represent the semantics.

Question 9. Text processing is one of the most common tasks in many ML applications. Name the most important steps that are involved in data pre-processing and briefly describe each of these tasks.

Answer: The following are the important tasks which can all be performed by using built in NLTK modules:

1. Text normalisation: convert all the upper-case letters to lower-case.
2. Tokenisation: Here the given Input (sentence) will get broken down into small pieces called tokens (either words, sentences).

3. Part-of-Speech Tagging: here the tokenised word's will get assigned to their respective grammatical category for example (verb, noun, adjective….)
4. Stop words and punctuations: these are the words that do not add any value to the sentence by simply keeping them means more processing time requires. So, it is better to remove those words and as well as we need to remove the punctuation marks.
5. Stemming: Here we normalize the Words to their root form (like Texting to Text).
6. Lemmatisation: Lexical Resources like Wordnet is required to make sure that the word belongs to the dictionary.

Question 10. **In order for humans to trust machine learning methods, we need explain ability – models that are able to summarize the reasons for the behaviour of a machine, gain the trust of users, and produce insights about their decisions. Describe which machine learning algorithm that you came across comes closest to these requirements and explain why.**

Answer: To me SVM algorithm seem to satisfy all the above-mentioned requirements the reasons are as follows:

1. We can clearly see what the model is trying to do which is nothing but separating the data points using the hyperplane where the calculations involved in transformations like 2D to 3D a kernel trick is used to make the computation faster.
2. SVM has several types of Kernels to suit both linearly separable and non-linear data.
3. SVM Margin size can be controlled by C.
4. We can reproduce the similar results every time using the Random State.
5. The SVM model performs well on complex data and medium data.