Applied Machine Learning Homework 1

Jivitesh Poojary: ID 2000116443

Answers:

1. What is the key reason for over fitting? Explain with an example. (2 points)

i. Over fitting is a scenario which occurs when while expanding a decision tree by using an algorithm you create a large number of nodes covering all types of possible solutions. This is undesirable as we not actually training with the data we are just memorizing the values with the possible outcomes. As a result, we observe that the accuracy of the algorithm may be high on a training dataset however, a lower accuracy is observed on a test dataset. Thus it become too complex and has a poor predictive performance.

We can say the key reason for overfitting would be that we are not putting limitations on learning model while running the algorithm. We should set a threshold after which no new nodes are created and try to find outcomes, which can be accommodated by searching or modifying existing nodes. So stopping it late and nor regulating the process are some reasons for overfitting.

- ✓ Eg: Consider we are using a dataset containing genetic information of individuals for predicting the possibility of tuberculosis. For each individual with a different genetic makeup if we were to create a decision tree based on each unique feature. Over time after rigorous training, we may find the common genetic makeup for tuberculosis sufferers. However, when we test it with any new data the result may show a lower accuracy. We can thus conclude that the system was not learning but just memorizing.

---

2. Let us consider the following task: A physician (say Mike) approaches with you with a data set about a rare disease. His data set contains 10 different clinical measurements such as cholesterol (HDL and LDL), blood pressure (SBP and DBP), BMI, age etc. His data set has about 20 people with the disease and 480 others without it. Will you accept this data set?
What will be your concerns? (3 points)

i. I will accept the dataset. The reason behind it is that the disease being rare, we may find difficulty in finding any data on the topic. By working on the dataset there is a possibility that the diagnostic processes may improve in correctly classifying the disease. However, we have to keep in mind the concerns and risks should be mitigated wherever possible.

ii. I will be having some concerns regarding the data, some of them are listed below:

- Small size of dataset: The given dataset has information of only 500 people. As the disease is rare we would want to have the information of larger population of people. When you look at the bigger dataset of human population which is the actual affected group the above dataset seems miniscule. In the dataset there are 20 people suffering from the disease, the proportionate value may increase or decrease when calculated on a large set of people.
- Skewed dataset: If we were to use the dataset, we would be predicting the occurrence of the disease with an accuracy of 96% (480/500). This value is significantly larger as the people having the disease are less. There is a high possibility of getting a false positive outcome of no disease when applied to a larger dataset.
- Fewer clinical measurements: The task is about identifying the factors that aid in the occurrence of a rare disease. There are only 10 clinical measurements that are provided. As a result, there is a possibility of some critical measurements that serve as a differentiating factor have been somewhere missed out and we may not be able to identify the factors that increase the possibility of the disease.
- Accuracy of the available data: The physician who has approached us with dataset may have collected the data diligently. However, we have to take into consideration the possibility of human error while collecting the data.

---

3. Now assume that all the measurements in the above problem are continuous measurements. What are the construct features that are meaningful for this task? (2 points)

i. The data that is obtained can be in multiple forms, it can be discrete, nominal, boolean or it can be continuous. As the name suggests discrete data is in the form of possible set of discrete value eg: size of shirts – small, medium, large. On the other hand, continuous data is usually numerical and has a range of values associated with it eg: height of students in the classroom – 4ft to 6ft this will include even the fractional values of height.

ii. The following are the construct features that are meaningful for the task:

- Ignore hierarchical information: In this case, we only refer to the leaf nodes and ignore hierarchical information. As a result, we may have to travers a number of nodes instead of focusing on a selected few.
- Precise values: Using continuous data we can get accurate values for metrics that serve as a differentiating / identification characteristic. We get a range of values and we can find specific ranges of measurements where we need to work on. For instance, the range of 'Normal' blood pressure varies with the age group, gender, location and other habits. However the probability of a continuous datapoint occurring is miniscule as we can always go to a more finer value.
- Complex system: Working with continuous data when not required may make the system more complex unnecessarily even when it is not of importance. For instance, if we are measuring the RGB colour of the shade of the exterior car and its performance, we may get a continuous data but it is making the system complex and not affecting our outcome.

---

4. Define the following terms: training set, tuning set and test set (2 points)

i. Training Set – It is a subset of the examples present in the dataset of examples that is used to obtain a model by training the classifier. It occupies a larger proportion of the dataset. It is useful for predicting relationships in the feature variables of a class. The elements of a training set are randomly selected by performing a split on the dataset. We apply learning algorithm on the training set data and

ii. Tuning Set – It is a subset of the examples present in the training set that is used to select the best possible set of parameter settings that are obtained from repeated application on all of training set examples.

iii. Test Set – It is also a subset of the examples present in the dataset distinct from the training set, it is used for validating how good the model is at predicting and estimating the properties of the model. The test set after being split from the original data has to kept aside and only be used during the testing phase. If the test data is leaked or exposed to the model before time we will be jeopardizing the entire process.

---

5. What is the problem with using accuracy as a performance metric? What are ROC Curves? How can you do a trade-off between false positives and false negatives in a medical domain? (2 points)

i. Accuracy is defined as the percentage of examples that are correctly classified over the total set of examples.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

Here there is a possibility of a true value being incorrectly classified as false and a false value being classified as true. Since accuracy is calculated on the basis of false positive and false negative, there is a cost associated with values being incorrectly classified even when the overall accuracy may be high. Another reason being that the dataset itself may be biased or skewed towards a particular value. When we are predicting using this data we are bound to get most of the test results biased towards this value. Having said this, accuracy serves as a decent performance metric when the dataset is not biased.

- ✓ Eg: - For instance if 95 % of an example are immune to smallpox then classifying a person to having an immunity to small pox has a success rate of 95%. However, if the same person is actually suffering from smallpox there is grave possibility of spreading the disease.

ii. ROC Curve – Receiver Operating Characteristics (ROC) is a graph that compares the true positive and false positive rate of a classifying algorithm. We can look at it as a probabilistic classifier. It was developed during WWII for radar research. It was developed because measuring performance of algorithms on the basis of accuracy does not give us the best picture of performance.

In an ideal scenario there should be no false positives and only true positives in our results however algorithms and data are not perfect in the real world and hence we have the need of calculating the curve. We plot the curve by using the true positive rate (TP) on the Y – axis and false positive rate (FP) on the X - axis. Different algorithms can work better in different parts of ROC space. This depends on cost of false + vs false -. We can compare two algorithms on a ROC curve by integrating the area below the curve for the give range of required values. Larger the area better the algorithm.

iii. In the medical domain, values of false positives and false negatives are critical as they deal with matters of life and death. ROC curves can be used to measure the accuracy of any diagnostic test and the likelihood of a disease in the future. Practitioners can set a threshold for the success or failure of a test and adjust it according to find the performance of a test. Similarly, we can use ROC curves on the data obtained from clinical trial, the performance of a drug should match the standards set out the regulatory body of the government (could be Food and Drug Administration department)

---

6. Consider a classification problem - the goal is to predict the people at risk of heart attack. The algorithm classifies correctly 40 out of the 50 people who were at risk in the test set and classifies correctly 100 out of 150 people who were not at risk. Draw the confusion matrix. What is the false positive rate and false negative rate? (3 points)

i. Confusion matrix: It is a tool that gives us information on the cost of incorrect correct classification. It also gives the user an idea about the performance of a model in terms of false positives and false negatives. It gives us an idea of the performance of the model and the costs associated with it.

n =200	Model Result Risk	Model Result No-Risk	
50	True Positive (TP) = 40	False Negative (FN) = 10	Expected Result Risk
150	False Positive (FP) = 50	True Negative (TN) = 100	Expected Result No-Risk
	90	110	

i. False Positive Rate – It is defined as the probability of incorrectly classifying a negative value as positive. False positive rate is calculated as the ratio of the False Positive (FP) values divided by the sum of True Negative (TN) values and False Positive (FP) values. It serves as a false alarm in cases when do not expect an outcome to appear as true. Eg: Classifying a person suffering from ulcer as suffering from skin cancer.

$$\textit{False Positive Rate} = \frac{FP}{TN + FP}$$

ii. False Negative Rate – It is defined as the probability of incorrectly classifying a positive value as negative. False positive rate is calculated as the ratio of the False Negative (FN) values divided by the sum of True Positive (TP) values and False Negative (FN) values. It is a serious issue as in this case we do not consider a true value and wrongly classify it as false. Eg: Incorrectly classifying a sick patient to be healthy there is a possibility of spread of the disease.

$$\textit{False Negative Rate} = \frac{TP}{TP + FN}$$

---

7. Consider a university data set - There are professors who teach courses that are taken by students. In addition, professors advise students and there can be more than one professor who advises a student. Finally, professors and students co-author papers and of course, there can be more than 1 professor and 1 student in each paper. Assume that each professor has the following attributes - popularity and tenure level. Each student is described by his/her IQ level, years in the program and success in the program. Each course is described by its difficulty level and the average rating of the students.

Now, create flat feature vector data sets for the following prediction tasks: predicting the success of a student, predicting the popularity of the professor and predicting the rating of a course. For each of these clearly define the features and present the same. (6 points)

Feature Vector:
Professor – {Popularity, Tenure level}
Student – {IQ level, Years in the program, Success in the program}
Course – {Difficulty level, Average rating of the students}

Flat Feature Vector Representation:

	Predicting the success of a student	Predicting the popularity of the professor	Predicting the rating of a course
Professor - Popularity	0	1	1
Professor - Tenure level	0	1	1
Student - IQ level	1	0	0
Student - Years in the program	1	0	0
Student - Success in the program	1	1	1
Course - Difficulty level	1	1	1
Course - Average rating of the students	0	1	1

Prediction task:

i. Predicting the success of a student:
   - The feature vector dataset we shall be choosing here is {IQ level, Years in the program, Success in the program, Difficulty level}
   - I believe that the success of student depends on the students' abilities, experience and the difficult of the course. Professor's popularity, tenure level and the average course ratings given by student are insignificant in this context.
   - Some additional features that can be possible are career guidance, research opportunities, building a network, CGPA, interaction in the class etc.

ii. Predicting the popularity of the professor:
   - The feature vector dataset we shall be choosing here is {Popularity, Tenure level, Years in the program, Success in the program, Difficulty level, Average rating of the students}
   - Popularity of the professor depends on the tenure level, success of the student in the program, difficulty level of the course and the average rating of the students.
   - Some other features that can be possible are industry ties, research done, current research, available funding for future research, industry ready teaching methodology, etc.

iii. Predicting the rating of a course:
   - The feature vector dataset we shall be choosing here is {Popularity, Tenure level, Years in the program, Success in the program, Difficulty level, Average rating of the students}
   - Rating of the course depends on the features of the professor, some features of student and features of the course itself.
   - Some other features that can be possible are industry ready curriculum, project on real life scenarios, opportunity for collaboration, learning experience of students, involvement of students, etc.