Applied Machine Learning Homework 1

*by*

*Himanshu Goyal*

1. What is the difference between overfitting and under fitting?

Answer: overfitting happens when the algorithm fits the data too well, on the other hand under fitting occurs when the algorithm does not fit well the data. Under fitting leads to low variance whereas overfitting leads to high variance.

## 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?

Answer: we will accept this data set as this is rare disease and availability of data may be difficult. But this data set will have following concerns.

1. This will be an overfitting for the training data.
2. We need to have false negative as minimum as we can. If this model predicts that person has no disease despite of actually person having disease will have an adverse effect.
3. It will be difficult to create the training sets from the given data set.

## Now assume that all the measurements in the above problem are continuous measurements. How will you construct features that are meaningful for this task?

Answer: Continuous measurements can be converted to categorical variables.

1. Age Group (10-20, 20-40,40-60, 60+)
2. BMI (Underweight, Normal, Overweight)
3. Blood pressure (High, Medium, Low)
4. Define the following terms: training set, tuning set and test set.

Answer:

1. **Training Set**

Training set is the portion of labeled dataset used for learning i.e. to fit the parameters of classifiers.

1. **Tuning Set**

A set of examples used to tune the parameters of a classifier.

1. **Test Set**

A set of examples used only to assess the performance of a fully trained classifier.

## What is the problem with using accuracy as a performance metric? What are ROC Curves? How can you trade of between false positives and false negatives in a medical domain?

Answer: It is always crucial to check the accuracy of model. It is calculated as (TP + TN)/ (TP+ TN + FP + FN). Higher accuracy is not good when there is bias for a particular class.

If we consider the disease problem, we tend towards high accuracy model. In that case it is important to predict who has disease. In such scenarios, especially in industries like paramedical high accuracy is not good performance metric.

**ROC Curves:**

ROC curve is a plot between sensitivity (Y axis) and specificity (x axis). It stands for Receiver operating characteristics. These are mainly used in medicine and biology as a standard as they judge algorithm more than accuracy.

**Tradeoff between False Positive and False Negative in medical field**

We should build model with reduced specificity. This can be done by minimizing the false negative. If we minimize this, it will reduce the false positive.

## 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?

Answer:

|  |  |  |
| --- | --- | --- |
|  | **TRUE Positive** | **FALSE** |
| **TRUE** | 40 | 10 |
| **FALSE** | 50 | 100 |

False Positive Rate (FPR) = FP/ (FP + TN) = 50/ (50 + 100) = 0.33

False Negative Rate (FNR) = FN/ (FN + TP) = 10/ (10 + 40) = 0.20

1. 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.

Answer:

Following vector set can be derived with given features.

**Professor popularity:**

Prof- tenure, sub-rating, number of student advised, number of papers coauthored, sub-difficulty, number of courses he teaches

**Student- Success:**

Student IQ, Subject difficulty, subject rating, student year, number of papers coauthored

**Subject rating:**

Subject difficulty, student success, number of students enrolled for that course