Ivan Shapirov

CSC4780/6780

Homework 1

1. The inductive bias for this prediction task would be that the iris plants which are part of the same species would be closer to each other (their Euclidean distance would be smaller) compared to iris plants of different species.

   One restriction bias which I will train for this task would be to use a multi-dimensional graph with each descriptive feature being on one of the axes. Since all of the data has a ratio data scale, a graph will be an appropriate model as this will make it very easy to calculate the Euclidean distance between any data points.

   One preference bias which I will train would be that in the case that the 3 nearest neighbors of a data instance we are trying to predict have different species, then we will prefer the species who has the majority (if 2 setosa and 1 virginica we predict setosa) and if there is no majority, (1 of each species) then we predict the species with the shortest Euclidean distance of the 3.

2. Sepal length: data type – numeric, data scale - ratio
   Sepal width: data type – numeric, data scale - ratio
   Petal length: data type – numeric, data scale - ratio
   Petal width: data type – numeric, data scale - ratio
   Species: data type – categorical, data scale - nominal

3. Machine Learning is an ill-posed problem because the information given to create Machine Learning models is not enough to guarantee that the models will successfully predict the target feature for new data points 100% of the time. This is because the training data is only a small subset of all possible combinations, so there is no way to ensure that the rules the model created from the training data will accurately label all data. Another problem is that there are likely multiple models which all successfully give the values for the training data, but these models contradict each other when faced with new data instances outside of the training data. This means that simply getting a model which fits the training data is not enough for the model to accurately predict future data inputs.

4. 1. First, the factors which can cause customer churn should be evaluated (Business Understanding of CRISP-DM). One possible factor is the popularity of movies available on this company's streaming service compared to their competitor. It could be that the competitor has a catalog of more enjoyable and popular movies compared to this movie streaming service's catalog.

   2. A possible data analysis solution would be to create a model which predicts whether or not a movie will be enjoyed by customers of the online streaming service. This will allow the movie streaming service to only add movies to its catalog which customers will enjoy, both

increasing customer satisfaction of the catalog available and not wasting money to buy the rights to stream movies which will not be enjoyed.

The target feature can be collected by asking customers who watched a certain movie if they enjoyed the movie. If more than a certain threshold, for example 50%, enjoyed the movie, then that movie can be labeled as “enjoyable” and anything less will be “not enjoyable”. For descriptive features, the model can use box office sales of the movie (numeric, ratio), critic ratings (numeric ordinal), date of the movie release (numeric, interval), and genre (categorical, nominal) of the movie. All of this data can easily be collected online, and there should not be too many instances of missing data as this data is published for nearly every movie. The one data type which may pose some problems is critic reviews, as some movies which never gain attention do not get critics to review, but since these movies will rarely be considered by the streaming service (due to low popularity) those instances can be ignored. (Data Understanding, Preparation, and Modeling of CRISP-DM).

3. The predictive model will then train itself to best predict whether a movie is enjoyable or not based on the descriptive features. The business will use this model by inputting movies into the model, and the model will return whether or not each movie will be enjoyable. The streaming service will then select only those movies which are likely to be enjoyed by the consumers. Before deploying, the model should be tested on actual data to see its accuracy in selecting movies, and this accuracy should be compared to what the streaming service’s accuracy was in selecting movies. (Evaluation in CRISP-DM)

4. If this model increases the prediction accuracy of movie selection, the business can use this model to solve their problem of customer churn, because one of the factors identified as potentially causing the customer churn is the customers not enjoying the catalog of movies with which they are presented. Since the features can all be found out before a movie is added to the streaming service’s catalog, the business can know ahead of time whether or not movies will be enjoyed by their customers, resulting in more customers enjoying the movies on the streaming service and not switching to a competitor. (Deployment of CRISP-DM)