ISTE-470 Assignment 3

##### Fall 2221

Answer the following questions and submit them to the Assignment 3 dropbox by the specified due date.

Scoring: Questions 1-11 are worth 10 points each; Question 12 is worth 35 points.

1. How are both training sets and test sets used in classification?

The training sets are used to create the model

The test set is used to show the quality performance of the model.

2. “Only one decision tree may be built from a given data set”. Do you agree with this statement? Explain your reasoning.

I don’t think that only one decision tree must be built for a data set, I believe that

3. On Slide 16 of the Intro to Classification lecture, which attribute (Own Car, Car Type, or Student ID) provides the best split? Why?

Car type is the best attribute that gives the best split. This is because car type has a high degree of purity, while own car and student id are a lower degree of purity.

4. What is the accuracy of the Humidity model on Slide 23 of the Intro to Classification lecture? How did you compute it?

70 %

= 0.7

5. Given the two models on Slide 28 of the Intro to Classification lecture and your knowledge of accuracy and cost, which model should you choose? Why?

I would choose the model with 90% accuracy, I believe that the cost is worth the increase in 10% accuracy.

6. Why is overfitting a problem that should not be ignored?

Because when you overfit, you model ends up learning from the outliers and therefore decreasing the accuracy of your model.

1. Why is it advantageous to choose an odd k-value when using a Nearest Neighbor classifier for a two class classification problem?

Because having odd k-values prevent having the chance of getting ties on our data.

1. Compute the Euclidean distance between points p(4.6, 3.4, 1.4, 0.3) and q(6.3, 3.3, 4.7, 1.6). Show all of your work.

=3.93

1. Data-driven classifiers do not create a model in order to classify new data instances. What do they do instead? Use either of the two data-driven classifiers we learned about in class as an example to illustrate the overall classification methodology for a data-driven classifier.

Data driven classifiers don’t create a model, instead what they do is that they take a known set of data clusters, and then grabbing and unknown and classifying to the closes cluster.

For the k-nearest neighbor method, it takes clusters of known classified data, and then looks for the k-nearest neighbor, and whoever category gets the most amount it gets placed there.

1. What is the difference between the *I* (impulse) function and the *sign*function in an Artificial Neural Network?
2. Support Vector Machines can produce any number of hyperplanes for a given training set. How does the SVM decide which hyperplane is the **best** one to use as its model?

12a. (5 points) Open diabetes.arff in a text editor (Notepad++) and read about its attributes. Once you understand what the attributes represent, open the data set in Weka. Run the following classifiers using the default algorithm values and 10-folds cross validation: J48, Nearest Neighbor, Naïve Bayes, and Artificial Neural Network (ANN). Note the accuracies of each algorithm in the table below.

|  |  |
| --- | --- |
| **Classifier** | **Accuracy** |
| J48 | 77% |
| Nearest Neighbor | 75% |
| Naïve Bayes | 77% |
| ANN | 85% |

12b. (15 points) Discretize the diastolic blood pressure (pres), BMI (mass), and age attributes using the values shown in the tables below. Create a new ARFF file with this data in it and name it diabetes\_disc.arff. Include a screenshot of each attribute’s distribution in Weka after you have performed discretization on those attributes. Be sure to properly label each screenshot.

|  |
| --- |
| **Diastolic Blood Pressure** |
| low: [0, 90) |
| ideal: [90, 120] |
| prehigh: (120, 140] |
| high: > 140 |

Chart

Description automatically generated

|  |
| --- |
| **Body Mass Index (BMI)** |
| underweight: [0, 18.5) |
| normal: [18.5, 25] |
| overweight: > 25 |

Chart, bar chart

Description automatically generated

|  |
| --- |
| **Age** |
| young: [0, 40) |
| middle: [40, 60] |
| elderly: > 60 |

Chart, bar chart

Description automatically generated

12c. (10 points) Using the discretized data set, rerun J48, Nearest Neighbor, Naïve Bayes, and ANN and note their accuracies in the table below. How did the accuracies of each classifier change from the previous data set to now? Did discretization improve classifier performance or not for these classifiers?

|  |  |
| --- | --- |
| **Classifier** | **Accuracy** |
| J48 | 75.1302 % |
| Nearest Neighbor | 66.6667 % |
| Naïve Bayes | 76.0417 % |
| ANN | 73.9583 % |

Accuracies are no longer real numbers, they are instead decimals because they are more discrete. Discretization didn’t improve the performance for the classifiers. This is because discretizing the values makes it more vague and inaccurate.

12d. (5 points) Using the Nearest Neighbor classifier on the **continuous** data set (diabetes.arff), change the *k*-value to 3, 5, 7, and 9 and note the resultant accuracies in the table below. What happens to the classifier’s accuracy as *k* increases? Why might this happen?

|  |  |
| --- | --- |
| ***k*-value** | **Accuracy** |
| 3 | 72.6563 % |
| 5 | 73.1771 % |
| 7 | 74.7396 % |
| 9 | 72.1354 % |

Classifier accuracy gets higher the higher the k value is until it reaches 9 where it decreases. This happens because at 9 k becomes too large, and the neighborhood is including points from other classes making it less accurate.