Data Mining & Predictive Analytics

2023-02-28

### Question 1

Based on the dataset given, the outlier is the score of 3 in the eighth exam.

To verify this numerically, we can calculate the mean and standard deviation of the dataset. The mean of the dataset is:

(mean) = (84 + 81 + 77 + 90 + 92 + 64 + 75 + 3) / 8 = 70.75

The standard deviation of the dataset is:

(std dev) = sqrt(((84-67.625)^2 + (81-67.625)^2 + … + (3-67.625)^2) / (8-1)) = 28.774

The score of 3 is more than 2 standard deviations away from the mean, which makes it an outlier.

Regarding the question of which measure, the mean or median, is more affected by the presence of an outlier, it depends on the number of observations in the dataset and how far the outlier is from the rest of the data. In this case, since we only have 8 observations, the mean is more affected by the presence of the outlier.

To show this, let’s compare the mean and median of the dataset with and without the outlier:

With outlier: (mean) = 70.75 (median) = 79

Without outlier: (mean) = (84 + 81 + 77 + 90 + 92 + 64 + 75) / 7 = 80.43 (median) = 81

We can see that the mean is significantly affected by the presence of the outlier, while the median is not affected as much. This is because the mean is sensitive to extreme values, while the median is resistant to them. Therefore, in datasets with outliers, it is often more appropriate to use the median as a measure of central tendency instead of the mean.

### Question 2

To classify record 6 with the k-Nearest Neighbor algorithm, we first need to calculate the distances between record 6 and the existing records. We will use the Euclidean distance for this calculation:

where is the value of feature k in record i, and n is the number of features.

Using this formula, we can calculate the distances between record 6 and the existing records:

Next, we need to select the k nearest neighbors to record 6. Since k=3, we will select the three nearest neighbors:

record 2: distance 5 record 3: distance 5 record 5: distance 5 Now, we need to determine the majority class of these three neighbors. In this case, all three neighbors belong to class “Large”. Therefore, we can classify record 6 as “Large”.

The reasoning behind this classification is that we are assuming that the k nearest neighbors are representative of the local area around record 6. Since all three of the nearest neighbors belong to class “Large”, it is likely that record 6 also belongs to class “Large”.

It is important to note that the k-Nearest Neighbor algorithm can be sensitive to the choice of distance metric and the value of k. In some cases, it may be beneficial to use a weighted voting scheme, where the closer neighbors have more influence on the classification. Additionally, choosing an appropriate value for k can be important in balancing between bias and variance in the model.

### Question 3

1. The multiple linear regression equation for predicting the donation amount based on the three variables, education, income, and number of children, can be written as:

Donation = 445.0583 + 31.79566 \* Education + 0.003698 \* Income + 45.69131 \* Kids

This equation shows how each of the three variables contributes to the predicted donation amount. The intercept term, 445.0583, represents the expected donation when all three variables are zero (which is not a realistic scenario in this case). The coefficients of the three predictor variables represent the change in the predicted donation amount for each unit increase in the respective variable, holding the other variables constant. For example, holding income and number of children constant, each additional year of education is associated with an increase of $31.79566 in the predicted donation amount.

1. The F-statistic of 7.974 and p-value of less than 0.001 indicate that the overall regression model is statistically significant, meaning that the model explains a significant amount of the variation in the donation amount. The R-squared value of 0.825119 indicates that 82.5% of the variation in the donation amount can be explained by the three predictor variables. The adjusted R-squared value of 0.813714 takes into account the number of predictor variables and penalizes the R-squared value if there are too many predictors in the model. This adjusted R-squared value is slightly lower than the R-squared value, indicating that adding the third predictor variable may not have improved the model as much as expected. The standard error of the estimate, Se, is a measure of the variability of the errors in the model, and is used to estimate the standard deviation of the errors.

The p-values for each of the four coefficients indicate that all three predictor variables are statistically significant in predicting the donation amount, with p-values less than 0.001. This means that it is highly unlikely that the coefficients are zero in the population.

1. To predict the expected donation by a person with 16 years of education, $90,000 annual income, and 2 children, we can substitute the values of the predictor variables into the regression equation:

Donation = 445.0583 + 31.79566 \* 16 + 0.003698 \* 90 + 45.69131 \* 2

Donation = 445.0583 + 508.73056 + 0.332682 + 91.38262

Donation = $1045.5043

Therefore, based on the regression model, we would expect a person with 16 years of education, $90,000 annual income, and 2 children to donate $1045.17444 to the charity. It is important to note that this is only an estimate, and the actual donation amount may differ due to other factors not accounted for in the model.

### Question 4

To derive the rules from the decision tree, we need to examine the splits and the nodes in the tree. The rules can be written as “if condition, then outcome,” where the condition is a combination of variables and values, and the outcome is a prediction of survival or not. We can also calculate the support and confidence for each rule to assess their reliability.\

To calculate the support and confidence for each rule, we need to count the number of instances in the dataset that match the conditions of the rule and also have the outcome variable equal to "yes". For example, for rule 1, we count the number of instances where gender = male, age = child, pclass is either 1st or 2nd, and survived = yes. This gives us a support of 16. To calculate the confidence, we divide the support by the total number of instances that match the conditions of the rule, regardless of the outcome variable. In this case, there are 16 instances where gender = male, age = child, and pclass is either 1st or 2nd, so the confidence is 100%. We repeat this process for each rule to obtain the support and confidence values.

Rule 1: If gender = male and age = child and pclass = 1st or 2nd, then survived = yes

Support: 16/2201 = 0.73% Confidence: 16/29 = 55.17%

Rule 2: If gender = male and age = child and pclass = 3rd, then survived = no

Support: 35/2201 = 1.59% Confidence: 35/64 = 54.69%

Rule 3: If gender = male and age = child and pclass = crew, then survived = no

Support: 0 Confidence: N/A (no instances in the node)

Rule 4: If gender = male and age = adult, then survived = no

Support: 1364/2201 = 62.00% Confidence: 1364/1731 = 78.78% Rule 5: If gender = female and pclass = 1st or 2nd, then survived = yes

Support: 344/2201 = 15.63% Confidence: 344/379 = 90.76% Rule 6: If gender = female and pclass = 3rd, then survived = no

Support: 126/2201 = 5.72% Confidence: 126/182 = 69.23%

From these rules, we can see that gender and age are important predictors of survival on the Titanic. Specifically, male children in 1st or 2nd class were more likely to survive, while male children in 3rd class and crew were not. Adult males were unlikely to survive. For females, those in 1st or 2nd class were very likely to survive, while those in 3rd class were not.

These rules can be useful for understanding the factors that influenced survival on the Titanic, and for making predictions about new passengers based on their gender, age, and class. However, it is important to note that these rules were derived from a single decision tree model, and may not be generalizable to other datasets or contexts. Additionally, the rules only consider a limited set of variables, and other factors such as individual characteristics, behavior, and location on the ship may also have played a role in survival.