1. Given the data set X with three input features and one output feature representing the classification of samples:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X: |  |  |  | O |
|  | 2.5 | 1.6 | 5.9 | 0 |
|  | 7.2 | 4.3 | 2.1 | 1 |
|  | 3.4 | 5.8 | 1.6 | 1 |
|  | 5.6 | 3.6 | 6.8 | 0 |
|  | 4.8 | 7.2 | 3.1 | 1 |
|  | 8.1 | 4.9 | 8.3 | 0 |
|  | 6.3 | 4.8 | 2.4 | 1 |

(a) Rank the features using a comparison of means and variances.

To get a treshold value I got the average of all data values in , , and = 4.77619

= {2.5, 5.6, 8.1}, 0 = {1.6, 3.6, 4.9}, 0 = {5.9, 6.8, 8.3},

1 = {7.2, 3.4, 4.8, 6.3}, 1 = {4.3, 5.8, 7.2, 4.8}, 1 = {2.1, 1.6, 3.1, 2.4}

SE(0 - 1) = = = 1.82317

SE(0 - 1) = = = 1.15328

SE(0 - 1) = = = 0.76702

= = 0.0137123801 < 4.77619

= = 1.871473825 < 4.77619

= = 6.127610753 > 4.77619

The analysis shows that and are candidates for reduction because the mean vallues are clsoe and the final test Is below the thresholld value. on the other hand has potential to be a distinguishing feature between the two classes 1 and 0.

1. Given four-dimensional samples where the first two dimensions are numeric and last two are categorical:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 2.7 | 3.4 | 1 | A |
| 3.1 | 6.2 | 2 | A |
| 4.5 | 2.8 | 1 | B |
| 5.3 | 5.8 | 2 | B |
| 6.6 | 3.1 | 1 | A |
| 5.0 | 4.1 | 2 | B |

(b) Apply relief algorithm under the assumption that is output (classification) feature.

() = () +

= 0.1/6 = 0.0167

= -12.94/6 = -2.1567

I used all samples in this question so m = n which is the number of rows. Based on the results of relief algorithm the feature is much more relevent then . I don't know what the threshold would be for this data but that would have to be determined by model using a folrmula with the model accuracy.

1. Apply the ChiMerge technique to recue the number of values for numeric attributes in problem #3

(a) Reduce the number of numeric values for feature and find the final, reduced number of intervals.

|  |  |
| --- | --- |
|  | O |
| 2.5 | 0 |
| 3.4 | 1 |
| 4.8 | 1 |
| 5.6 | 0 |
| 6.3 | 1 |
| 7.2 | 1 |
| 8.1 | 0 |

Sorted values for

Splitting interval points are 0, 2.95, 4.1, 5.2, 5.95, 6.75, 7.65

For = 10% threshold is = 2.706

Smallest splitting intervals are [5.2, 5.95] and [5.95, 6.75]

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=0 | K=1 | sum |
| [5.2, 5.95] | = | = | = |
| [5.95, 6.75] | = | = | = |
| sum | = | = | N = |

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=0 | K=1 | sum |
| [5.2, 5.95] | = 1 | = 0 | = 1 |
| [5.95, 6.75] | = 0 | = 1 | = 1 |
| sum | = 1 | = 1 | N = 2 |

= = 0.5

= = 0.5

= = 0.5

= = 0.5

= + + + = 2 < 2.706 we can assume that we can merge these since it is less than threshold

Intervals [2.95, 4.1], [4.1, 5.2]

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=0 | K=1 | sum |
| [2.95, 4.1] | = 0 | = 1 | =1 |
| [4.1, 5.2] | = 0 | = 1 | =1 |
| sum | =0 | =2 | N = 2 |

= 0.1

= = 1

= 0.1

= = 1

= + + + = 0.2 < 2.706 we can assume that we can merge these since it is less than threshold

Intervals [0, 2.95], [2.95, 5.2]

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=0 | K=1 | sum |
| [0, 2.95] | = 1 | = 0 | =1 |
| [2.95, 5.2] | = 0 | = 2 | =2 |
| sum | = 1 | = 2 | N = 3 |

= = 0.3333

= = 0.6667

= = 0.6667

= = 1.3333

= + + + = 3.00037 > 2.706 we don’t merge this since it is greater than threshold

Intervals [2.95, 5.2], [5.2, 6.75]

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=0 | K=1 | sum |
| [2.95, 5.2] | = 0 | = 2 | =2 |
| [5.2, 6.75] | = 1 | = 1 | =2 |
| sum | =1 | =3 | N = 4 |

= = 0.5

= = 1.5

= = 0.5

= = 1.5

= + + + = 1.3333 < 2.706 we can assume that we can merge these since it is less than threshold

Intervals [6.75, 7.65], [7.65, 8.1]

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=0 | K=1 | sum |
| [6.75, 7.65] | = 0 | = 1 | = 1 |
| [7.65, 8.1] | = 1 | = 0 | = 1 |
| sum | = 1 | = 1 | N = 2 |

= = 0.5

= = 0.5

= = 0.5

= = 0.5

= + + + = 2 < 2.706 we can assume that we can merge these since it is less than threshold

Intervals [2.95, 6.75] [6.75, 8.1]

|  |  |  |  |
| --- | --- | --- | --- |
|  | K=0 | K=1 | sum |
| [2.95, 6.75] | = 1 | = 3 | = 4 |
| [6.75, 8.1] | = 1 | = 1 | = 2 |
| sum | = 2 | = 4 | N = 6 |

= = 1.3333

= = 2.6667

= = 0.6667

= = 1.3333

= + + + = 0.374920 < 2.706 we can assume that we can merge these since it is less than threshold

Final intervals are [0, 2.95] and [2.95, 8.1]

1. Explain the differences between averaged and voted combined solutions when random samples are used to reduced dimensionality of a large data set.

Averaged and voted combined solutions are different in how they select the final solution. In the case of the voted solution it looks for the most common solution that is in the set of solutions. In the case of the average solution it takes the average of all the solutions in the set of solutions and then it uses this average as the final solution. Example solutions = {1, 1, 2} voted would be 1 and averaged would be 1.333.

1. Given the data set: F = {4, 2, 1, 6, 4, 3, 1, 7, 2, 2}. Apply two iteration of bin method for value reduction with best cutoffs. Initial number of bins is 3. What are the final medians of bins, and what is the total minimized error?

Ordered F = {1, 1, 2, 2, 2, 3, 4, 4, 6, 7}

Bin1 = {1, 1, 2} Bin2 = {2, 2, 3} Bin3 = {4, 4, 6, 7}

Smoothing based on median value

Bin1 = {1, 1, 1} Bin2 = {2, 2, 2} Bin3 = {5, 5, 5, 5}

ER = 0+0+1+0+0+1+1+1+1+2 = 7

Move values to different bins to try and reduce the error

Bin1 = {1, 1} Bin2 = {2, 2, 2} Bin3 = {3, 4, 4, 6, 7}

ER of this bin set up is = 0+0+0+0+0+1+0+0+2+3=6

Smaller then the original

1. Discuss situations in which we would use the interpolated functions given in Figures 4.3b, c, and d as "the best" data-mining model.

For figure b in the example it is a linear approximation which could be useful for looking for values between given data points. Such as an approximating function that has a minimum distance from all given data points.

Figure c is and example of nonlinear approximation which could be seen in some learning methods such as multilayer artificial neural networks. That because the output of an approximating function depends nonlinearly on parameters.

Figure d is an example of a quadratic approximation could be using for estimating the values of parameters.

1. Explain the difference between interpolation of loss function for classification problems and for regression problems.

The interpolation of loss function for classification problem is designed to guess what a data point is classified as and determine a loss value for values that are guessed and how close the guess was to the actual classification of the data point. A regression loss function tries to predict the next value in a data set and will interpolate a loss function that will calculate a loss value based on how close the data point predicted was to the actual data point that was supposed to go next.

1. Analyze the differences between validation and verification of inductive-based models.

For inductive based models validation is to check that the model behaves with satisfactory accuracy that is expected by the users. This confirms that the model is the right mode for the system.

Verification is for checking that the model is transformed from the data as intended into new representations with sufficient accuracy. This deals with building the mode right, which means the model corresponds correctly to the data.

1. A nearest neighbor approach is best used:
   1. With large-sized data sets.
   2. When irrelevant attributes have been removed from the data.
   3. When a generalized model of data is desirable.
   4. When an explanation of what has been found is of primary importance.

Select only one choice and give additional explanations.

If there are no irrelevant attributes in the data any more that would mean that hopefully that the data points in the model are condensed which would help make condensed nearest neighbor problem which should speedup the model using this approach. There is also less data for the model to calculate distance from so that it is not wasting time on irrelevant data that could slow it down in finding the nearest neighbor.

1. Show that accuracy is a function of sensitivity and specificity.

Accuracy = (sensitivity)(prevalence) + (specificity)(1-prevalence)

They are both also used in the ROC curve which is used to get the optimal threshold value.

An example could be a diagnostics model that diagnoses a patient as sick or not sick. The rate at which tests are evaluating a patient to be sick when they are not actually sick would be the sensitivity in a problem, and the rate at which patients are not diagnosed as sick when they are sick would be the specificity of the problem. Both of these values are used in a confusion matrix when determining the accuracy of the model to determine how many diagnoses the model got correct.