Homework 3

Yuan Li, N19728558, NetID: yl6606

Part II: Programming and Questions

1. Answer:

When k = 1, the 11, 12 and 17 examples were not correctly classified.

2. Answer:

When k = 1, the accuracy is 92.10%

3. Answer:

When k = 3, the 11 example was not correctly classified.

4. Answer:

When k = 3, the accuracy is 97.36%

5. Answer:

When k = 5, the 11 and 38 examples were not correctly classified.

6. Answer:

When k = 5, the accuracy is 94.73%

7. Answer:

When using zero-R, the accuracy is 23.68%

8. Answer:

I tried manhattan distance and cosine distance with the same train data and test data.
When k = 1, the accuracy is $acc_{man} = 92.10\%, acc_{cos} = 92.10\%$.
When k = 3, the accuracy is $acc_{man} = 97.36\%, acc_{cos} = 97.36\%$.
When k = 5, the accuracy is $acc_{man} = 94.73\%, acc_{cos} = 94.73\%$.
So the accuracy is the same as the euclidean distance.

Manhattan distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates, and euclidean distance is just the connection distance between two points of their Cartesian coordinates. Cosine distance(also as Cosine similarity) finds the normalized dot product of the two attributes.

So I think when the data is distributed in multiple small pieces, all three distances can reach similar results.

9. Answer:

The cross-validation accuracy for k=3 is 92.84%.
The cross-validation accuracy for k=7 is 94.66%.
The cross-validation accuracy for k=99 is 18.81%.
When using euclidean distance, the best k should be 7, with the accuracy of 94.66%.