Intelligent Systems

20-EECE-5136

Homework 1

Due Tuesday, October 3rd, 2017 A. D.

By Tomas Seymour

M05873855

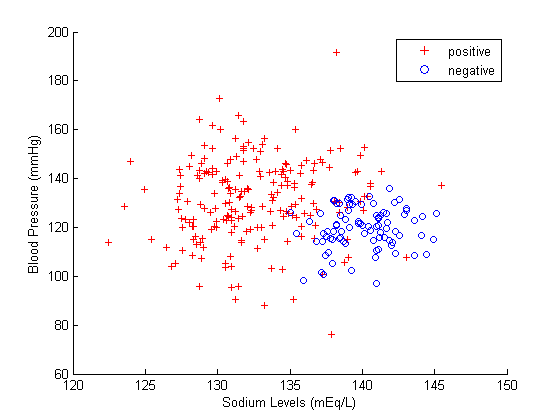
**Problem 1**

Problem Description:

Data was collected from 300 patients including their blood sodium level (L), their blood pressure (P), and whether they tested positive or negative for a disease.

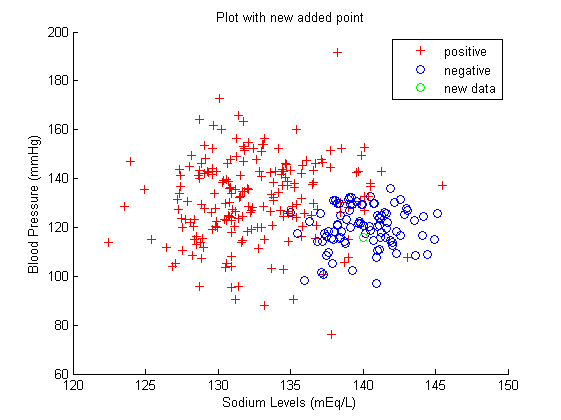
Use k-nearest neighbor algorithm and a neighborhood algorithm to classify new patient data entries added to the space for patient features L and P.   
The k-nearest neighbor algorithm is the standard algorithm where the k amount of the nearest data points respective to the new data point are accounted for, then whatever class is the most prevalent in those k neighbors is what the new data point is classified as.   
The neighborhood algorithm is a bit different where instead of choosing the k nearest data points to the new data point, instead a circle is drawn with the new data point as the centre. Then, whatever class is the most prevalent in the circle is the class which the new data point is associated as.

Results:

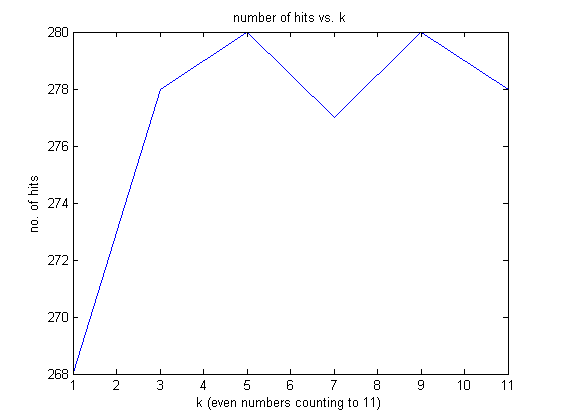


*Figure 1: P scatter plotted against feature L of the 300 patients. Color and shape of the dots represent if that patient was diagnosed positive or negative for disease.*

**Problem 1**

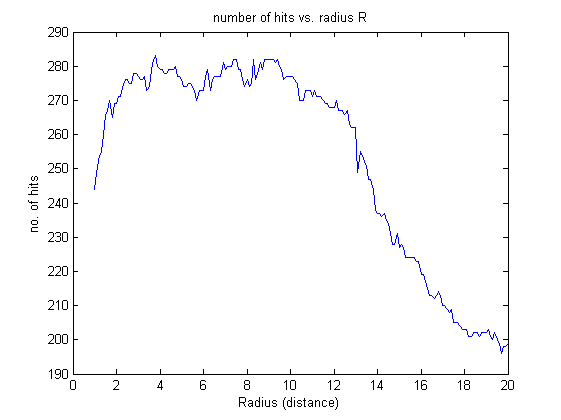


*Figure 2: using k-NN with k = 4 neighbors, a new patient with L of 140.0 and P of 116.0 was classified as testing negative for disease.*

**

*Figure 3: Using k-NN algorithm, with increasing k, the frequency of hits, i.e. the amount of times that each old data point in the test set was correctly assigned to the class that the patient was tested as, is plotted.*

**Problem 1**

**

*Figure 4: Using neighborhood algorithm, we remove every point from the test set, then include it as a new patient. Then, with increasing radius of the circle drawn around the point, plot how many hits there were for every point in the set, i.e. whether it was classified accurate to what it was classified in the given data.*

Discussion:

The k-NN algorithms gave the highest amount of hits at k = 5 and k = 9.   
The neighborhood algorithm gave the highest amount of hits at around radius = 4.

The max no. of hits seems to be pretty similar across both algorithms. However, when the radius of the circle about the “new” data point is 4, the neighborhood algorithm has a very slight advantage over the k-NN algorithm.

Conclusion:

Both algorithms proved to be somewhat affective in classification of the test data under certain conditions. As a result, new data would be accurately classified if the class was previously unknown.

I would recommend if one was using the k-NN algorithm to use a k between 5 and 9 for the maximum hit response. I would recommend if one was using the neighborhood algorithm to choose a radius between 4 and 10.

Both seem to have relatively the same maximum hit response: around 280. This suggests that there are around 20 data points that can be perceived as potential outliers, which did not receive hits. I don’t know for certain if these 20 data points are the same or not, but it could be easily determined from the model what these points are and attempt to scale these values for maximum efficiency.