

I have created two confusion matrices for this homework. One for the nearest mean algorithm and the other for the nearest neighbor. The results were as follows: (Columns represent the predictions and rows represent the actual classes)

Nearest Mean:

$$\begin{pmatrix} 25 & 0 & 0 \\ 0 & 22 & 2 \\ 0 & 3 & 23 \end{pmatrix}$$

Nearest Neighbor:

$$\begin{pmatrix} 25 & 0 & 0 \\ 0 & 23 & 1 \\ 0 & 2 & 24 \end{pmatrix}$$

The percentage of inaccurate predictions for the Nearest Mean algorithm was %6.67.

The percentage of inaccurate predictions for the Nearest Neighbor algorithm was %4.

In class, we were predicting that the nearest neighbor algorithm would perform better compared to the nearest mean algorithm. Our test result justifies this hypothesis. Such an outcome might be intuitive, but it is hard come up with good reasoning for it.

We can think of a set of data which has a distribution that is divided into two distant clusters. Their mean will probably be somewhere near the middle of the two clusters. If we take a test data from the middle, the nearest mean algorithm will select this class even though it is very unlikely that the data is actually in that class. The nearest neighbor algorithm will probably be more reliable in such situations.