

I also tried other Machine Learning classification methods to discover more about the best classifier to use in this scenario. (the code is in bonus.py and the results are in bonus.txt)

LogisticRegression:

A disadvantage of it is that we can't solve non-linear problems with logistic regression since its decision surface is linear.

KNN:

We need to tune the hyperparameter (k) that best match the dataset and doesn't cause the model to overfit (large k) or underfit (small k) the model. I tried this model using `k=[5, 10, 15, 20, 25, 30]`

Nearest centroid:

It assigns to observations the label of the class of training samples whose mean(centroid) is closest to the observation.

Quadratic Discriminant Analysis:

It is a more general version of the linear classifier, because it separates measurements of two or more classes of objects or events by a quadric surface.

Results:

The results are in bonus.txt for different hyperparameters and classification models. Logistic Regression seems to be the best one among these classifiers.