

Based on accuracy which model is the best one?

Looking at the accuracy values for each of the 12 models we used, I would choose the Linear Discriminant Analysis (LDA) model. This is because it had the greatest accuracy value out of the 11 different models (.973). At times (depending on each test) the Neural Network would tie with the LDA.

For each of the 11 other models, explain why you think it does not perform as well as the best one.

Starting with the Linear Regression, we can flat out discount that simply because we are fitting our data to a linear line, which won't match any of the complexities of our data, that is why we get such a low accuracy for that. Next, we used a Polynomial of degree 2 regression. This for sure helped the accuracy, increasing it all the way to .893. Since we are seeing data like a parabola, we can see an overall better fit. Though, when we increase to a Polynomial of degree 3 regression, the accuracy slightly goes down. This is because we overfit our data. Quadratic Discriminant Analysis most likely had a slightly less accurate result because of the overfitting characteristic of higher degree polynomials (kind of like degree 3 regression). kNN most likely didn't have as good of an accuracy because of the way points were distributed, leading to awkward distance calculations within kNN. Lets take a look also at the confusion matrices of some of the more accurate models within our test. Most of them are extremely close to what we see inside of LDA. While many are close, the predictions from K Neighbors Classifier are a little bit off, making it slightly less accurate. We can see similar from the new models we added in this assignemnt (Extra Trees, Random Forest, and Decision Tree). We can see the random forest have slightly less than a decision tree possibly from the way it chose observations in our dataset. Like I said above, the Neural Network would fluctuate in accuracy, with it either tying or being slightly under the accuracy in LDA. This difference in accuracy simply comes down to how the computer chose observations from within the dataset.

dbn.py

Does the program use k-fold cross-validation?

No the program doesn't use it

What percentage of the data set was used to train the DBN model?

80% of the data was used to train the mode, rest for training

How many samples are in the test set?

72 samples

How many samples are in the training set?

288 Samples

How many features are in test set?

8 features

How many classes?

10 classes

List the classes

it would just be the following indexes: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9