Stone Barrett

CSE 590: Introduction to Machine Learning

Exam 2

1)

a) Advantages: Can have better generalization performance, can improve accuracy of unstable models, can decrease the degree of overfitting, and can work well without heavily tuning the parameters

Disadvantages: Can perform poorly on sparse or high-dimensional data, requires more computational resources, can be more difficult to interpret

b)

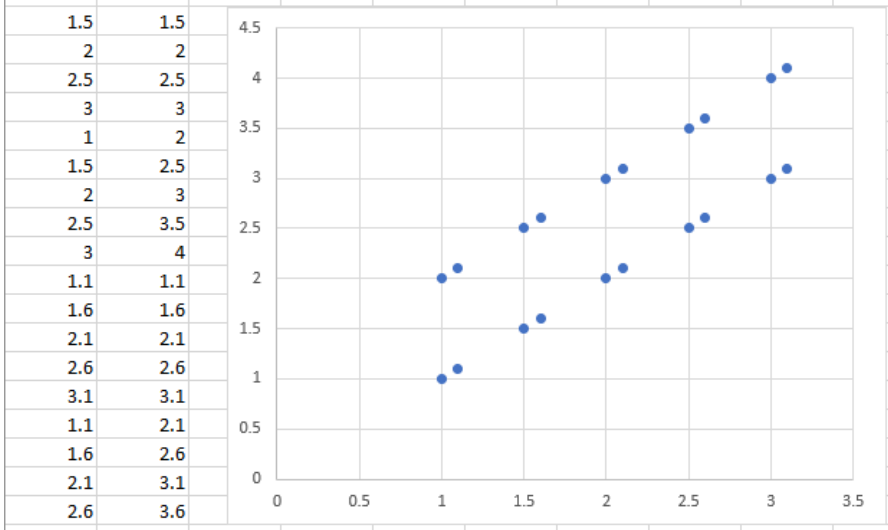
i) If an individual model is overfitting to the data, ensemble should be attempted

ii) If multiple models are being testing together in an ensemble, and the majority of their individual scores are significantly lower than the best individual score, the ensemble will perform worse than that individual

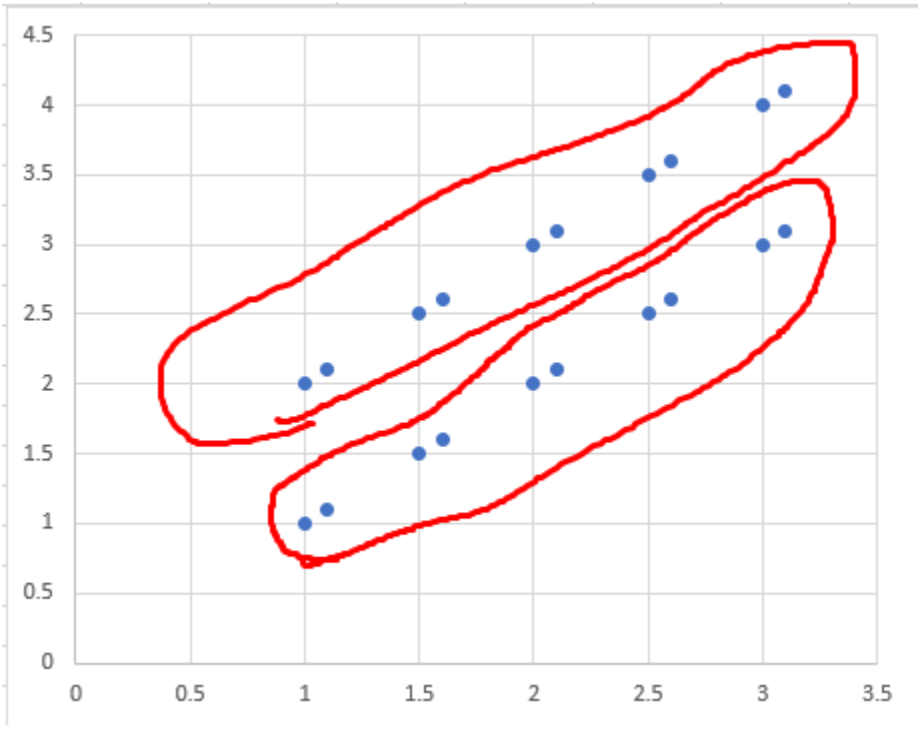
iii) If a model is significantly biased or the model has achieved very high testing accuracy already or is stable, ensemble will likely not improve performance

2)

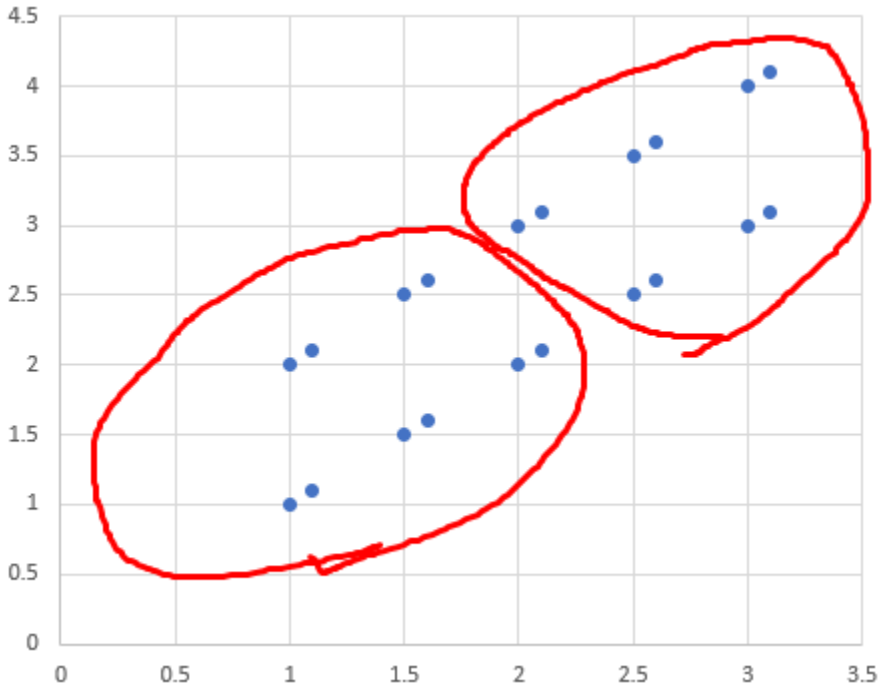
x	y
1	1
1.5	1.5
2	2
2.5	2.5
3	3
1	2
1.5	2.5
2	3
2.5	3.5
3	4
1.1	1.1
1.6	1.6
2.1	2.1
2.6	2.6
3.1	3.1
1.1	2.1
1.6	2.6
2.1	3.1
2.6	3.6
3.1	4.1



The dataset here should be clustered as such:

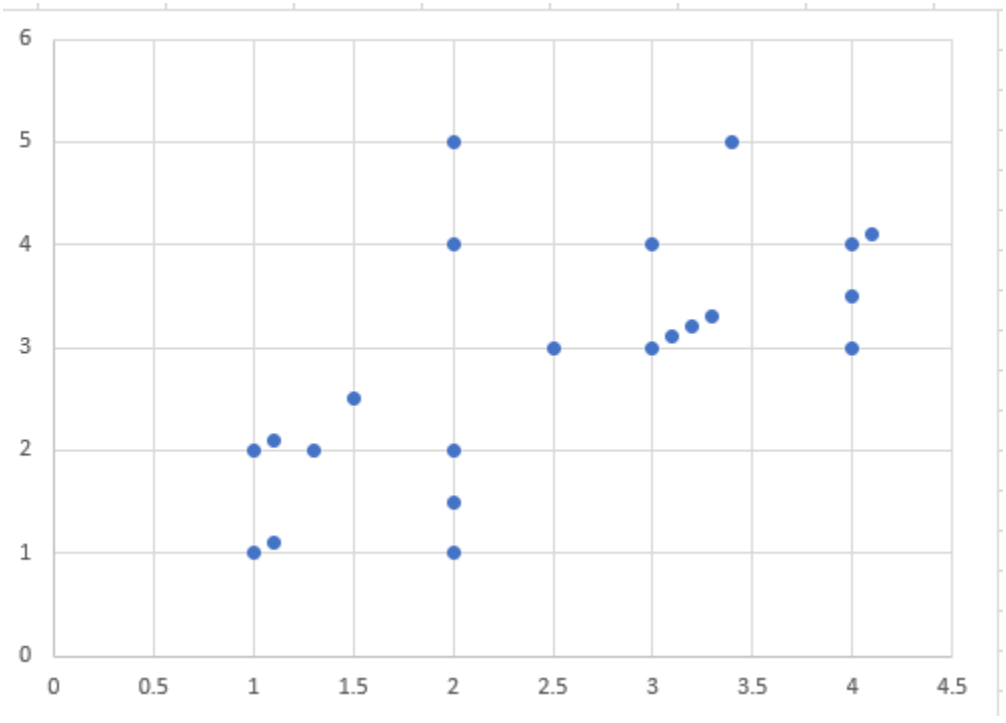


K-means would fail because it considers each direction to have equal weight. K-means might cluster it like this:

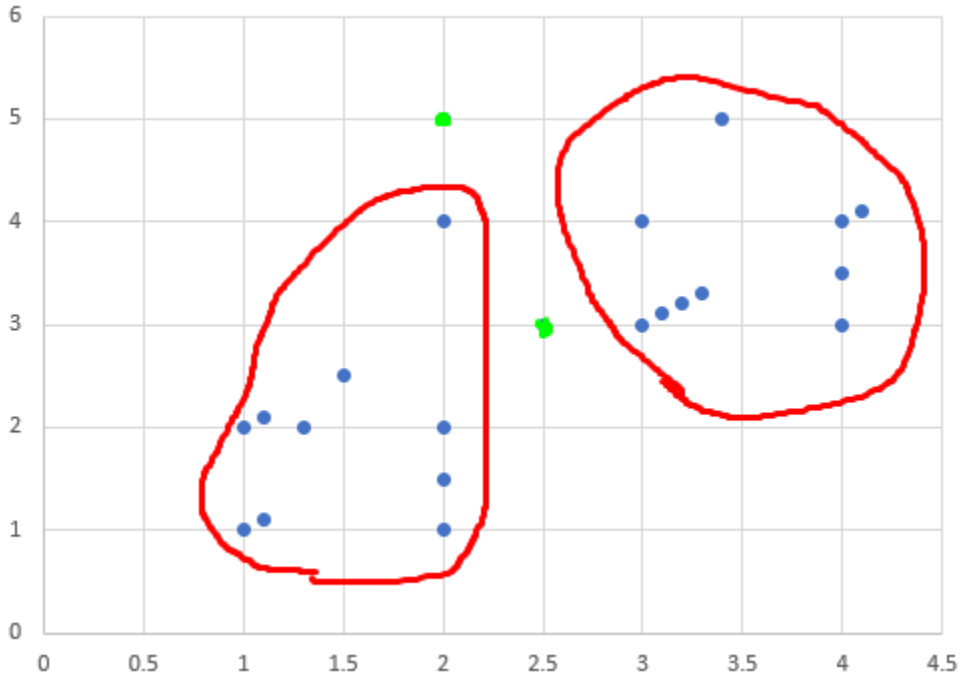


Agglomerative, however, continuously groups the nearest clusters together, meaning the corresponding groups in the same class would reach each other in that process before any of them reached a group from the other class. Average linkage should work to analyze this set.

3)



This dataset should be clustered as such (green is noise):



K-Means could not classify this correctly because of the (2,4) spot as well as the noise. Considering multiple directions as equal weights would again be its downfall. Agglomerative clustering also does not perform well with noise and in this case would fail due to all linkages being thrown off by the noise proximity to clusters. DBSCAN would work because of its ability to cluster even complex shapes and weed out noise points.

4) Data can be reduced to 2 or 3 dimensions in order to visualize it on a graph

Dimensionality can also be reduced to find patterns in relationships between specific features

Dimensionality can also be reduced to weed out features/variables that may throw off prediction such as an attribute that has many outliers in a dataset

5) If data is extremely imbalanced in a binary classification task, some models may score very high or very low in sheer accuracy, so an F1 score that considered precision and recall could return more reliable results. This would ensure that the model isn't just "remembering" the data and getting a high test score.

Another example with binary classification is a situation where one needs to not just analyze accuracy, but also the false positive and/or false negative rate in order to make a business decision and allocation funding or resources based on those values. In this case, the AUC-ROC curve would serve quite well to show how common false findings are compared to trues.

6) No. “svm” is not defined as an instance of the SVC class, so there is no model to be fitted to.

- *Load Iris data*
- *Normalize data to have zero mean and unit variance*
- *X_train, X_test, y_train, y_test = train_test_split (iris.data, iris.target)*
- *svm = SVC()*
- *svm.fit (X_train, y_train)*
- *score = svm.score (X_test, y_test)*

7)

a) No. The highest scores on the P2 Axis are all at the highest value tested, meaning that there are potentially higher values that could result in higher scores. P2 could have been .1, 1, 10, 100. P1, however, does not have this problem. The local maximum has been found at P1 = 1, so the range of tested values can be assumed correct.

b) P2: No, there is not a clear local maximum. A bell curve cannot be seen. Could have been .1, 1, 10, 100 or 1, 10, 100, 1000. Could be either because the scores seem to still be increasing fairly significantly between .1 and 1, so the optimal value could be further away.

P1: Scores decrease on either side of .1, making it the local maximum. The correct range was tested.

c) The highest score in the entire grid can be found towards the middle, meaning it is a local maximum for both parameters. Both ranges tested are correct.