[AAA] Advanced Analytics and Applications

Summer Semester 2021

Problem Set 2

1 Revisiting concepts from the lecture

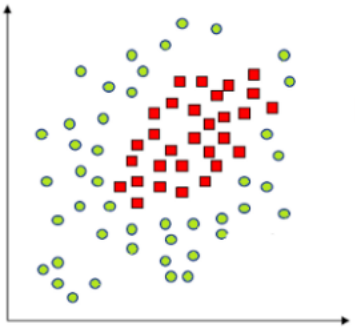


Figure 1 Random Data

1.1 Our objective is to find a classifier for the data above (red, and green dots) Why is a linear classifier not useful here?

1.2 Could we use a decision tree?

1.3 Watch the following YouTube video:
https://www.youtube.com/watch?v=Z2_yh2sice8&t=143s

Explain briefly how a linear classifier using an RBF kernel would approach the above problem?

2 SVM in a nutshell

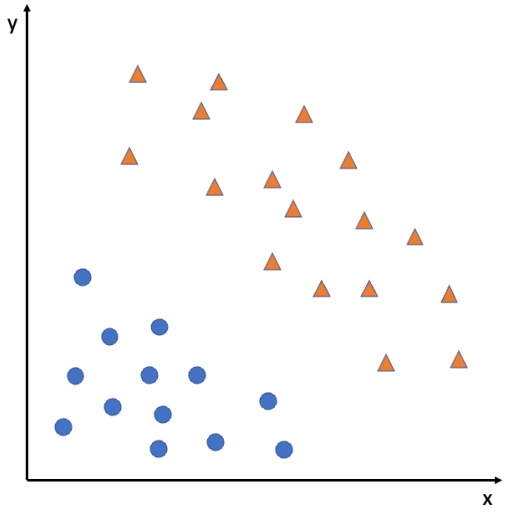


Figure 2 SVM

a) Before using any algorithm, how would you separate the data by drawing a line?

b) Draw several linear classifiers that separate the data correctly.

c) Roughly sketch the output of an SVM assuming a linear kernel. What is the difference between this output and the other linear classifiers?

d) Briefly explain what happens when support vectors are moved?

e) What happens when other than support vectors are moved?

3 Deep Understanding of Linear Decision Boundary

a) Read the following summary of the “perceptron” algorithm:

https://www.mathematik.uni-muenchen.de/~deckert/teaching/SS18/sec-steps.html

b) How does the algorithm work? Briefly explain the steps including the learning/update process.

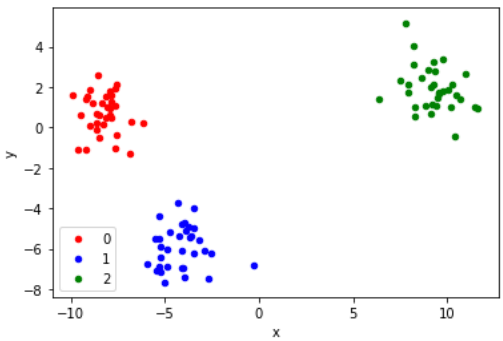
c) What is the key difference between the “Perceptron” classifier and a SVM classifier?

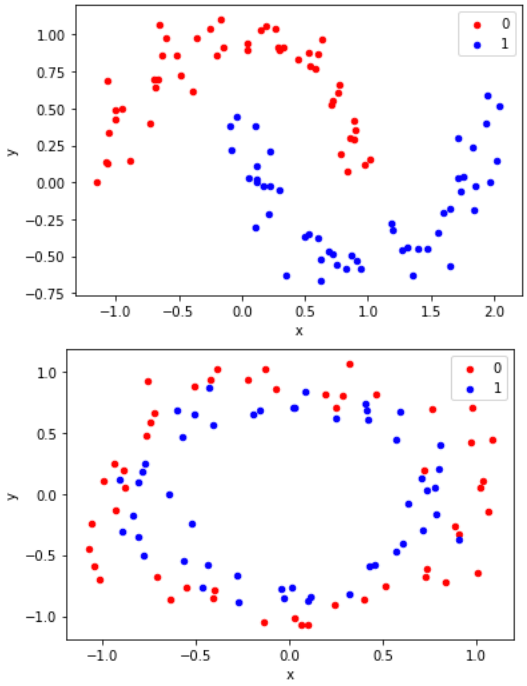
4 Programming

a) **Generating Toy Data**: Rather than using real world data, we want to get familiar with classifiers with toy data. Thus, we want to learn how to efficiently and quickly generate specific toy data following a predefined pattern (i.e., blob, circular, moon). The task is to read the documentation on

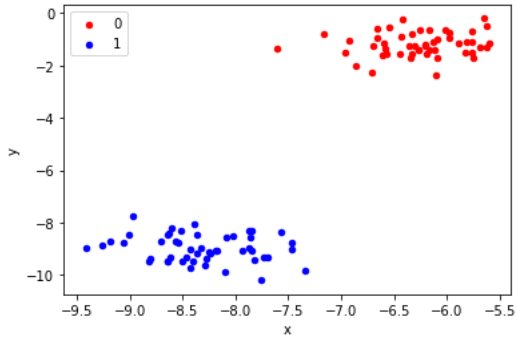
https://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets

and generate the following three datasets (blob, moon, circle)**.**





b) Again, generate a blob dataset with two clusters. Then train a support vector machine using a linear kernel on that dataset. Evaluate its performance (precision, recall).



c) Finally, generate circular data with two cluster centers. Then train a support vector machine using a linear kernel, and then repeat the training process but this time using a RBF kernel. Evaluate and compare both models.