Workshop Week 3

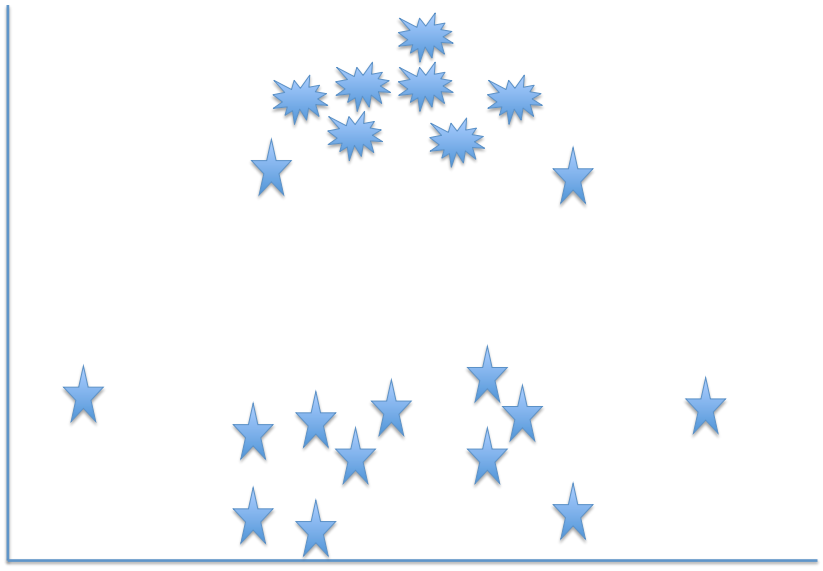
MBUSA Machine Learning, 2024

1 Theoretical

1) Consider an instance that is correctly classified and is far from the decision boundary. Why would an SVM's decision boundary not be affected by this instance, but the one learned by logistic regression would be affected?

2) Why does using kernels facilitate using SVMs with high dimensional feature spaces, without large overhead in the running time?

3) Suppose we are using an SVM with polynomial kernel of degree 2 on the dataset pictured below. This dataset may contain noisy instances in it. Answer the following questions



a) Draw the decision boundary if the value of the C parameter is very large.

b) Draw the decision boundary if the C parameter is near zero.

c) Which of these boundaries do you think will be more appropriate?

d) Insert an instance that would make a big change to the decision boundary when C is very large.

e) Insert an instance that would not make a big change to the decision boundary when C is very large.

2 Practical

The following libraries will be helpful for developing answers in Python

- ROC curves
- Cross validation
- Support vector machine model tutorial
- Support vector machine API

There is a partially filled Jupyter notebook you can use for creating answers to the questions - the file *week3-partial.ipynb*

For the *cats.csv* file use a support vector machine to predict the gender of a cat.

1) How does varying the C parameter affect the classification result of the SVM?

2) How does using different polynomial degrees affect the classification result of the SVM?

3) How does using RBF kernel affect the classification results of the SVM?

4) Which kernel do you think works best for this data and why?