AML Assignment 2 **Due Date: 9th Sept , 2019** **Total Marks: 90**

Instructions: You are allowed to discuss but the final answer should be your own. Any instance of cheating will be considered as academic dishonesty and penalty will be applied.

Q1 [20 Marks]. Data points are: Negative: (-1, 0) (2, -2) Positive: (1, 0)
Recall that for SVMs, the negative class is represented by a desired output of -1 and the positive class by a desired output of 1.

(i) For each of the following separators (for the data shown), indicate whether they satisfy all the conditions required for a support vector machine, assuming a linear kernel. Justify your answers very briefly.

a) x1 + x2 = 0
b) x1 + 1.5x2 = 0
c) x1 + 2x2 = 0
d) 2x1 + 3x2 = 0

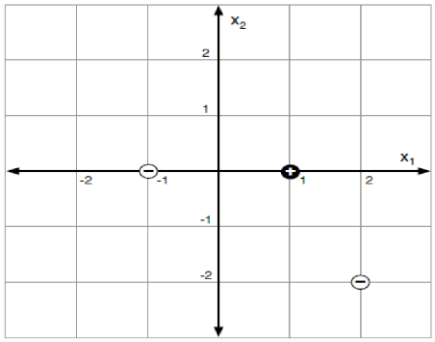


Figure Data points diagram

Write a clear and brief justification for each of the points listed in the question.

(ii) For each of the kernel choices below, find the decision boundary diagram that best matches. In these diagrams, the brightness of a point represents the magnitude of the SVM output; red means positive output and blue means negative. The black circles are the negative training points and the white circles are the positive training points.

a) Polynomial kernel, degree 2
b) Polynomial kernel, degree 3
c) Radial basis kernel, sigma = 0.5
d) (d) Radial basis kernel, sigma = 1.0

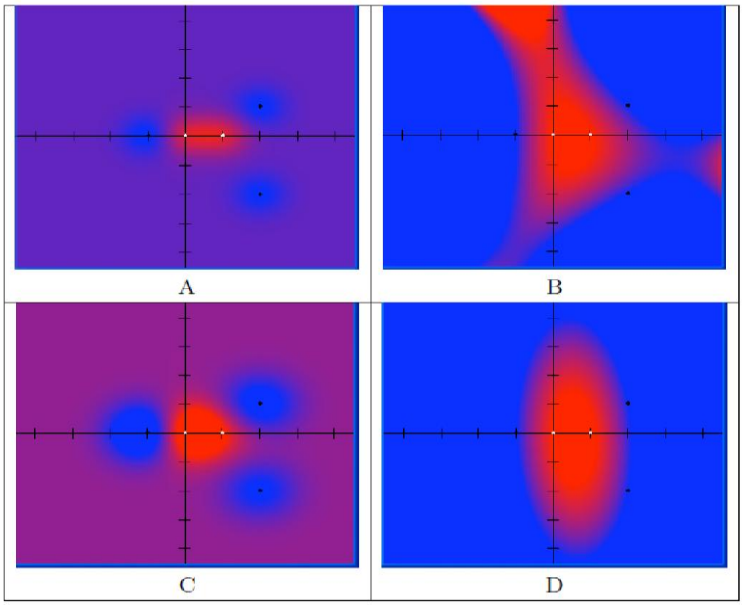


Figure Decision boundary diagram

Along with the chosen diagram, write the concise justification of your answer as well.

Q2 [40 Marks]. Use breast cancer classification dataset for multi-class classification. Train the SVM using following kernel functions and for each of the kernel learn the SVM using OVA and AVA techniques (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)):

- Linear

- Gaussian (radial basis) Kernel for each of the following values of the parameter $\gamma \in \{1, 10, 100, 1000, 10000\}$

a) Plot the decision boundary on top of a scatterplot of the data. Please make the boundary lines easy to see.
b) State-based on the plot—whether the model overfits, underfits, or performs well
c) [For OVA Linear Kernel SVM only] Perform 10-fold cross validation to pick the appropriate γ for this dataset. Show how the testing and training errors averaged across folds change with γ. What's the value of γ you would choose and what are its corresponding test/training errors?

You need to submit the code along with the 1-2 page analysis report regarding the experimentations you have performed. Decision boundary plots are also required as mentioned in (a). Do not forget to mention the accuracy computed in each step.

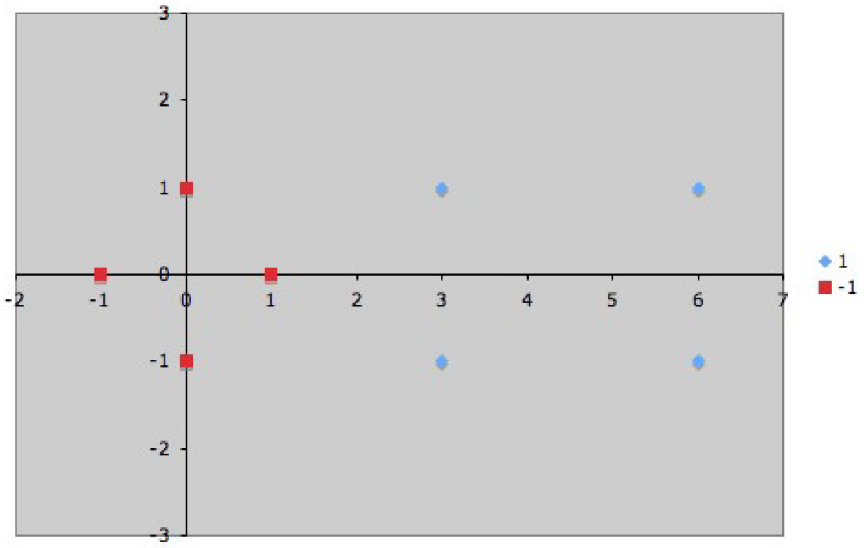
Q3 [20 Marks]. Perform online SVM learning on the above database. Perform the detailed analysis including the benefits of online SVM learning compared to traditional kernel learning.

You need to submit the code along with the 1-2 page analysis report regarding the experimentations. Also breif description of the algorithm used.

Hints (reading material of some online SVM learning):
http://digitalimaginggroup.ca/members/Shuo/Incremental%20Support%20Vector%20Learning%20for%20Ordinal%20Regression.pdf
http://web.mit.edu/seyda/www/Papers/GHC06_ACMSRC_abstract.pdf

https://www.semanticscholar.org/paper/Online-SVM-learning%3A-from-classification-to-data-Tax-Laskov/fc5664f83c49f88b079d3da355609db8b03aa706

Q4 [10 Marks]. Suppose we are given the following positively and negatively labeled data points in R^2. Find the SVM that can accurately discriminates the two classes (i.e., discriminating hyperplane)



Show the weight vector and bias and the steps used in computation.

Q5 [10 Marks]. What are the limitations of SVM when applied for large scale database? What is the computation complexity of SVM and how it limits its implementation for large scale database. Provide one detailed solutions/algorithm (along with pseudo code) which can be applied to solve the problem large scale data classification using SVM.