## Birla Institute of Technology & Science - Pilani, Hyderabad Campus
## Second Semester 2015-2016
## CS F415 / IS F415: Data Mining
## Comprehensive Examination – Part A

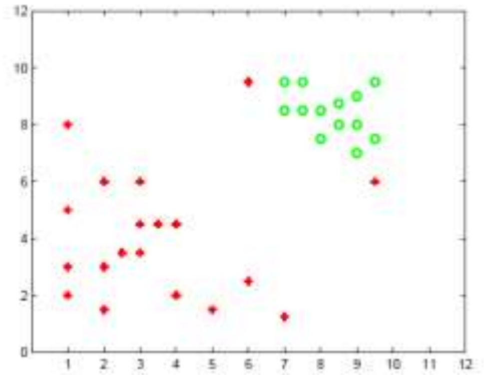
**Type: Closed** **Time: 180 mins** **Max Marks: 22** **Date: 12.05.2016**

**All parts of the same question should be answered together.**

**ID No.** **Name:**

1. The goal of this problem is to correctly classify test data points, given a training data set. You have been warned, however, that the training data comes from sensors which can be error-prone, so you should avoid trusting any specific point too much.

For this problem, assume that we are training an SVM with a quadratic kernel– that is, our kernel function is a polynomial kernel of degree 2. You are given the data set presented in the below mentioned Figure. The slack penalty C will determine the location of the separating hyperplane. Please answer the following questions qualitatively. Give a one sentence answer/justification for each and draw your solution in the appropriate part of the Figure at the end of the problem. [10 Marks]



Dataset for SVM slack penalty selection task

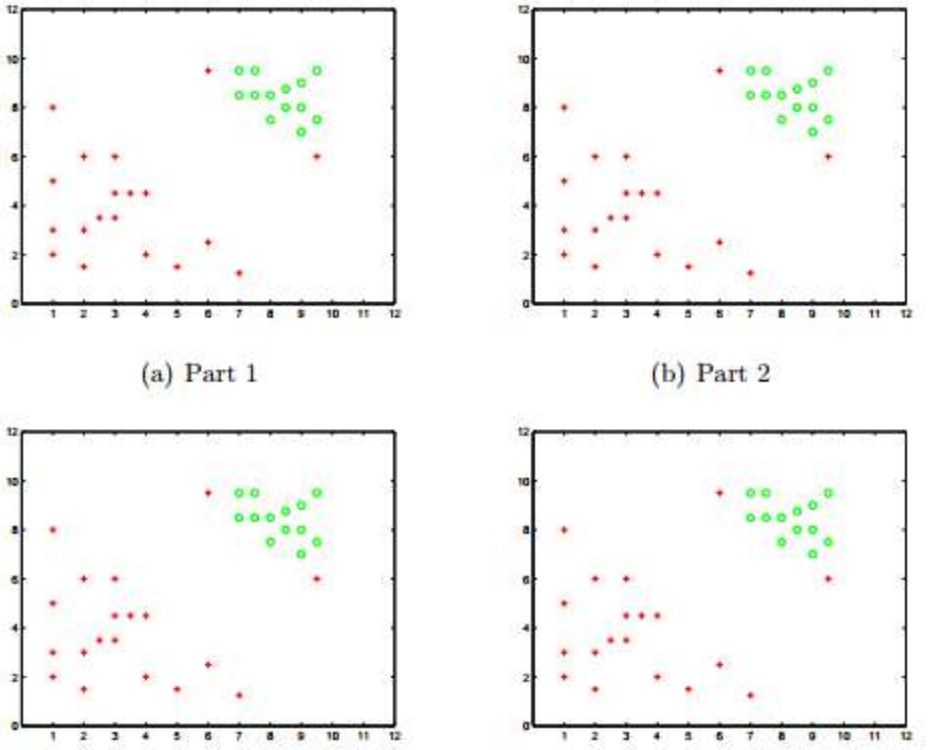
a. Where would the decision boundary be for very large values of C (i.e., C → ∞ (INFINITE) )? (remember that we are using an SVM with a quadratic kernel.) Draw on the figure below. Justify your answer.

b. For C ≈ 0, indicate in the figure below, where you would expect the decision boundary to be? Justify your answer.

c. Which of the two cases above would you expect to work better in the classification task? Why?

d. Draw a data point which will not change the decision boundary learned for very large values of C. Justify your answer.

e. Draw a data point which will significantly change the decision boundary learned for very large values of C. Justify your answer



(a) Part 1

(b) Part 2

(c) Part 4

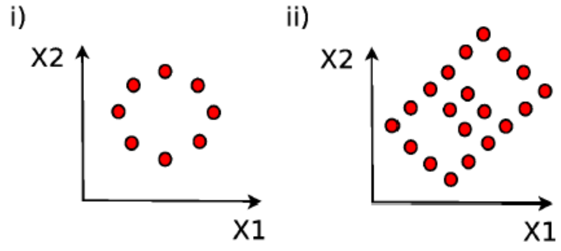
(d) Part 5

Draw your solutions for Problem here.

.

[PTO]

2. Which of the following unit vectors expressed in coordinates (X1, X2) correspond to theoretically correct directions of the 1st (p) and 2nd (q) principal components (via linear PCA) respectively for the data shown in the above mentioned Figure? Choose all correct options if there are multiple ones. [6 Marks]
Note: For each negative answer, -1 mark will be awarded.



Data in two-dimensions spanned by $X1$ and $X2$.

a) i) p(1, 0) q(0, 1) ii) p$(\frac{1}{\sqrt{(2)}}, \frac{1}{\sqrt{(2)}})$ q$(\frac{1}{\sqrt{(2)}}, \frac{-1}{\sqrt{(2)}})$

b) i) p(1, 0) q(0, 1) ii) p$(\frac{1}{\sqrt{(2)}}, \frac{-1}{\sqrt{(2)}})$ q$(\frac{1}{\sqrt{(2)}}, \frac{1}{\sqrt{(2)}})$

c) i) p(0, 1) q(1, 0) ii) p$(\frac{1}{\sqrt{(2)}}, \frac{1}{\sqrt{(2)}})$ q$(\frac{-1}{\sqrt{(2)}}, \frac{1}{\sqrt{(2)}})$

d) i) p(0, 1) q(1, 0) ii) p$(\frac{1}{\sqrt{(2)}}, \frac{1}{\sqrt{(2)}})$ q$(\frac{-1}{\sqrt{(2)}}, \frac{-1}{\sqrt{(2)}})$

e) All of the above are correct.

3. We would like to cluster the points in the below mentioned Figures (in fact both figured are the same) using k-means and GMM, respectively. In both cases we set k = 2. We perform several random restarts for each algorithm and chose the best one as discussed in class. For each method show the resulting cluster centers in the appropriate figure. [6 Marks]

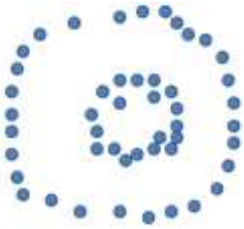


Figure : k-means

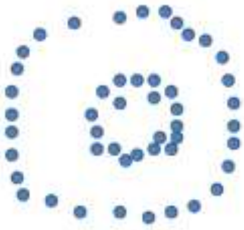


Figure : GMM