Data Mining

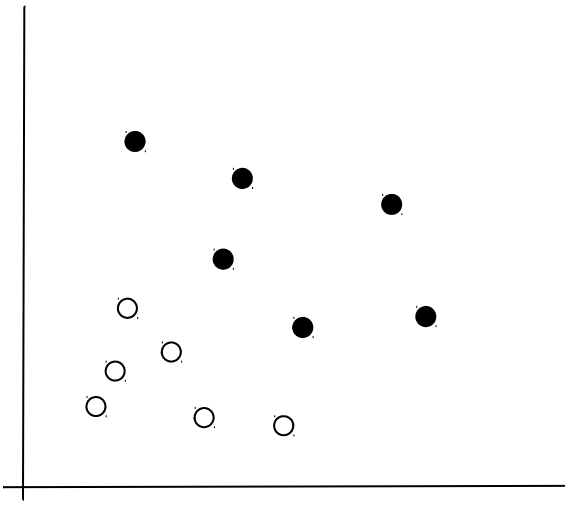
Homework 5

Important Notes:

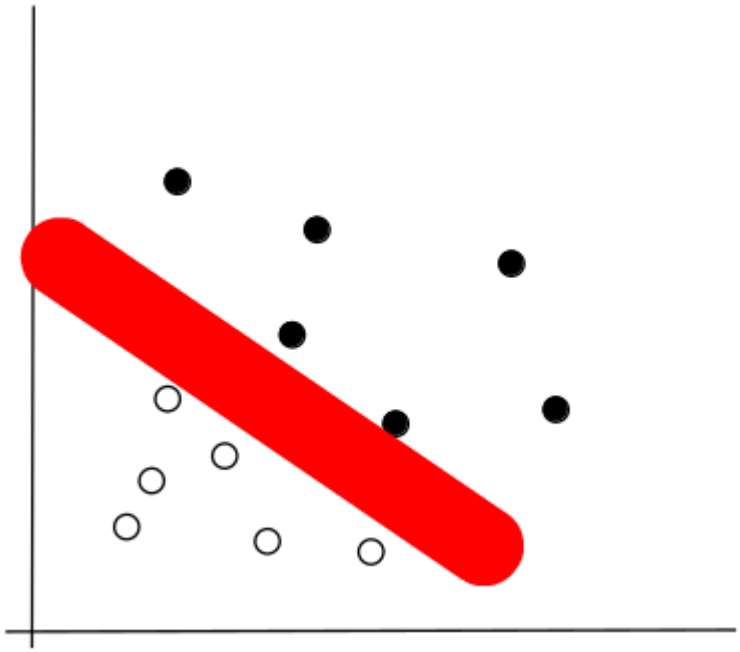
1. Submit in electronic form before 11:59pm on Wednesday, May 17, 2017
2. No late homework will be accepted.
3. The homework should be completed and submitted by each individual.
4. The homework should be submitted through Gradescope. Entry Code: **9BW66M**
5. The homework should be written in English.
6. The HW is worth it 10 points.
7. The [Research] questions require from you to do some research on the Web and get to understand things that were not covered during the lecture.
8. For questions, please use Piazza (English only!)

Exercise 1: Maximum Margin Hyperplanes [4 pts]

Suppose you are given the data shown below:



Assume that the white spheres are negative class examples and the black dots positive class examples. Illustrate in this figure the Maximum Marginal Hyperplane a Support Vector Machine will choose to separate these two classes and indicate which examples constitute the support vectors. Explain why is this the Maximum Marginal Hyperplane.



The maximum Marginal Hyperplane is the biggest line possible between the two classes, which results in the following image (40 pixels)

Exercise 2: Decision function [3 pts]

Assume that you have trained a support vector machine, and you have obtained the following parameters $\theta = [-2, 1/2, 3, -1, 5]$ for the hypothesis: $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$. Based on this hypothesis function decide the class of the following testing examples, $\mathbf{x}=[x_1, x_2, x_3, x_4]$: $\mathbf{x_{t1}} = [2, 2, 0, 4]$, and $\mathbf{x_{t2}} = [1, 1, 4, 0]$.

-2 + (½ * 2) + (3 * 2) + (-1 * 0) + (5 * 4) = 25. The class is positive (+)

-2 + (½ * 1) + (3 * 1) + (-1 * 4) + (5 * 0) = -2.5. The class is negative (-)

Exercise 3: Slack variables [3 pts]

Draw a 2-dimensional dataset which is not linearly separable (it should have at least 10 data points, 5 of a positive class and 5 of a negative class). Assume that you train two SVMs over these data points. For the first one set the parameter C to large value, while for the second one set the parameter C to a small value. Show over the dataset you have drawn two Maximum Margin Hyperplanes that are likely to be found by the two SVM for the two values of C, and explain why these two hyperplanes are likely to be found given these values of C.