Exercise 2 : SVM, Gradient Method , Perceptron and Regression

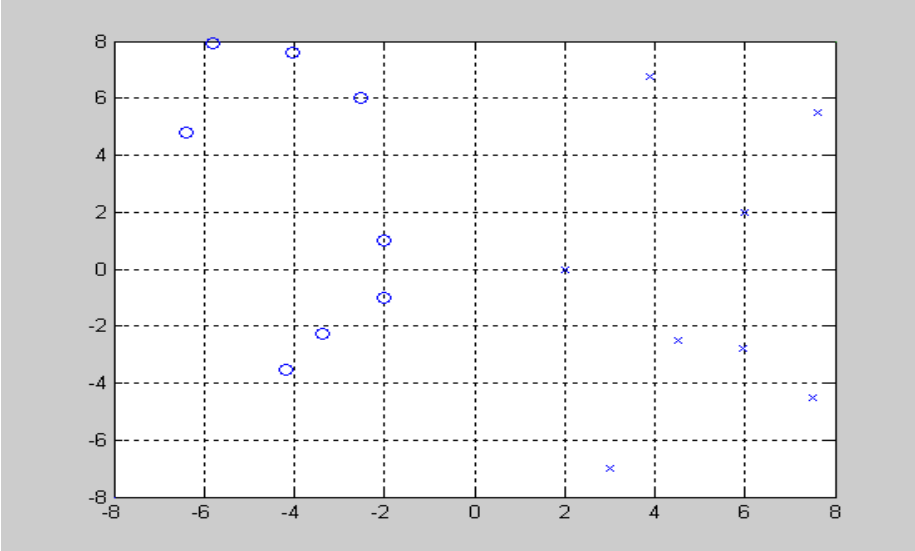
Homework Submission instructions:

- Homework is due on: Wednesday 16/8/2017
- Homework should be done in pairs. Each pair is to do their own work, separate from the other pairs (you can work in groups as long as the solution reflects your own work).
- We prefer you type your submission; however, you may submit scanned handwritten material as long as it is clear and readable.
- Please submit only one pdf file. **Make sure the file name contains your name and student ID** (don't submit a file name exercise2.pdf)
- Submission is done via Moodle website: https://moodle.technion.ac.il/course/view.php?id=4218

SVM

Problem 1

1) Consider the following linearly separable problem:



a. Which of the above data points are the support vectors?
b. What is the direction of the vector w ? Explain. (There is no need to explicitly calculate w)

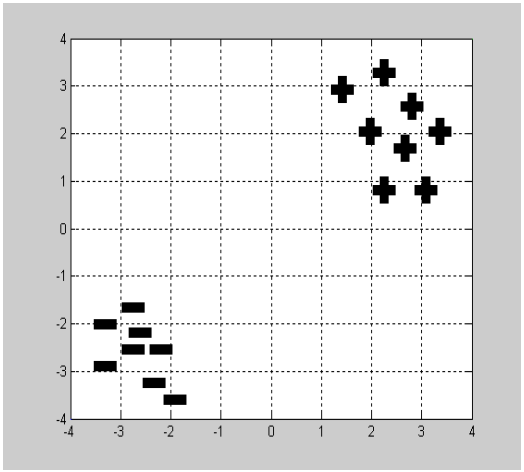
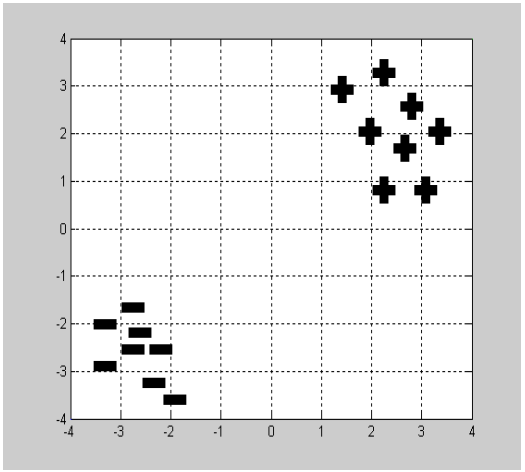
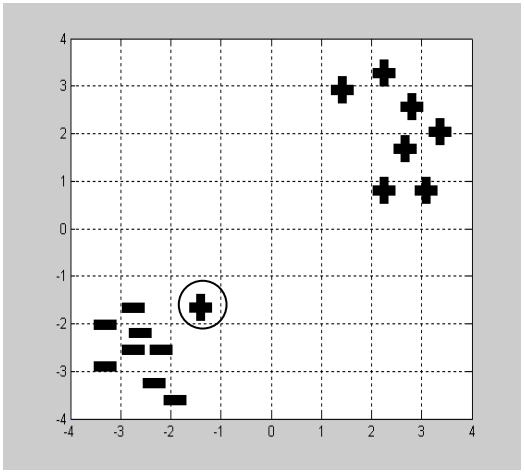
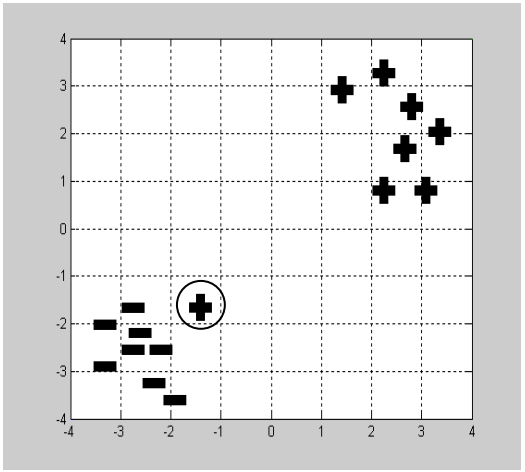
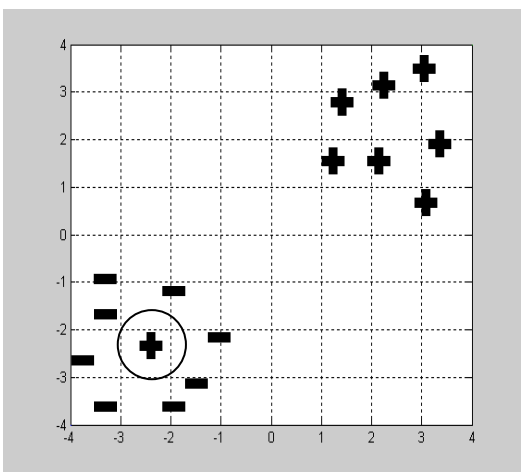
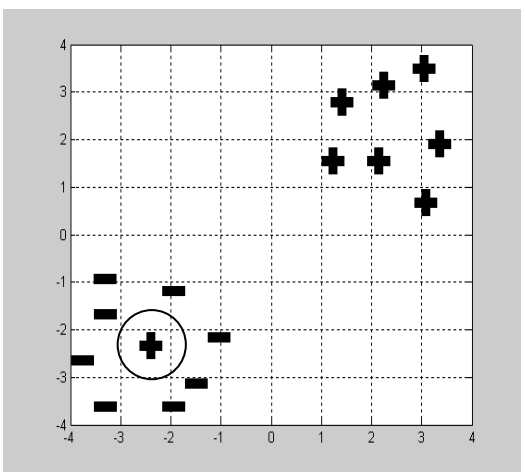
Problem 2

2)

a. Sketch the separating hyperplane for the following 3 dataset (see below) and for 2 values of C :
   - In the left column sketch the hyperplane for the case $C \to \infty$
   - In the right column sketch the hyperplane for the case $C < \infty$ .If the separation hyperplane does not exit, explain why.

b. In the last two problems (4 last figures) there is a circled data point, what is the suitable value of ξ (Equal to 0, between 0 to 1, greater than 1) for that point? Explain.

(You should attach this page to your homework)

$C < \infty$	$C \to \infty$
	
	
	

Problem 3

3) Assume the following training set: $\{x_1, x_2, ..., x_n\} \in \mathbb{R}^2$ with labels to two classes (binary problem). After training a SVM classifier with $C \to \infty$ the number of support vector received was $k = 2$ $(k < n)$. Later a new example x_{n+1} was added to the training set and a new classifier was learned. Determine which of the following options are possible, there could be more than one possible option (It is recommended to explain with a sketch)
   a. The number of support vector remained $k = 2$
   b. The number of support vector grew to $k+1$
   c. The number of support vector grew to $n+1$

Kernel Methods

Problem 4

4) Consider two kernel functions $k_1, k_2 : X \times X \to \mathbb{R}$. It is known that the classification problem is linearly separable for k_1 but not for k_2.
   a. Is the following function is a valid kernel function? (if yes, then explicitly show that it satisfies the conditions required from a kernel function)

$$k_3(x, x') = k_1(x, x') + k_2(x, x')$$

   b. Is the classification problem is linearly separable for k_3 ?

Problem 5

5) Which of the classifiers below have a zero training error on the following dataset:

x	Y
(-1,-1)	-1
(-1,+1)	1
(+1,-1)	1
(+1,+1)	-1

   a. Linear SVM
   b. SVM with a polynomial kernel function of degree 2
   c. SVM with a Gaussian kernel function $K_\lambda(x, z) = e^{\frac{-\|x-z\|}{\lambda}}$

Gradient Algorithm

Problem 6

6) Given is the following function:

$$f(x,y) = -20\left(\frac{x}{2} - x^2 - y^5\right)\exp\left(-x^2 - y^2\right)$$

a. Using MATLAB, plot this function in the range of $-3 \le x, y \le 3$. (You may use MATLAB functions `mesh` and `meshgrid`)
b. Implement in MATALB the gradient descent method for finding the minimum point. (Attach your code)
c. Initialize your algorithm with the following values:
   - As the initial guess, choose $[x_0, y_0] = [0.1, 1]$
   - As a step size, choose $\eta = 0.01$.

   Sketch the convergence graph of the algorithm (i.e. the value of the function at each step). To what point if any the algorithm converges?

d. Initialize your algorithm with the following values:
   - As the initial guess, choose $[x_0, y_0] = [1.5, -1]$
   - As a step size, choose $\eta = 0.05$.

   Sketch the convergence graph of the algorithm (i.e. the value of the function at each step). To what point if any the algorithm converges? Which phenomenon can be observed?

e. Initialize your algorithm with the following values:
   - As the initial guess, choose $[x_0, y_0] = [1.5, -1]$
   - As a step size, choose $\eta = 0.01$.

   Sketch the convergence graph of the algorithm (i.e. the value of the function at each step). To what point if any the algorithm converges? Compare your results with the results of section (d)

Perceptron

Problem 7

7) **Reminder: The perceptron learning algorithm**

input: set of labeled examples $\{(x_i, d_i)\}_{i=1}^{n}$ where $d_i \in \{-1,1\}$ and $x_i \in \mathbb{R}^d$

Initialization: the weights vector w_0 is initialized with zeros

For each step $t = 1,2,\ldots$:

- Choose one example x_t from the dataset
- Calculate the perceptron output for that sample using the current weight vector w_t :

$$y_t = sign\left(w_t^T x_t\right)$$

- Update the weights vector $w_{t+1} = w_t + \eta\left(d_t - y_t\right)x_t$

Assume:

- $\eta = \frac{1}{2}$
- The example that the algorithm receives it step t, x_t , is the i-th column of an orthogonal matrix sized $D \times n$ (i.e. the norm of each example is one $\|x_t\| = 1$
- The dataset is linearly separable
- The algorithm can receive each example only one time.

How many updates will the algorithm preform, at the least?

Regression

Problem 8

8) The purpose of this problem is to explore the properties of regression with regularization. Consider the following series of points $\{x_k, y_k\}_{k=1}^n$ where $y_k \in \mathbb{R}$, $x_k \in \mathbb{R}^l$ and $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^l \to \mathbb{R}^d$ is a given mapping.

a. For the following cost function :

$$E_{SSE,\lambda} = \sum_{k=1}^{n} y_k - \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_k)^2 + \lambda \sum_{i=1}^{d} w_i^2$$

i. Show that solution has the following form: $\mathbf{w}^* = (\Phi^T\Phi + R)^{-1}\Phi^T\mathbf{y}$ where Φ is a matrix defined by $\Phi = \boldsymbol{\varphi}(\mathbf{x}_1)\ \boldsymbol{\varphi}(\mathbf{x}_2)\cdots\boldsymbol{\varphi}(\mathbf{x}_n)^T$ and $\mathbf{y} = y_1\ y_2\cdots y_n^T$.

(Hint: express $E_{SSE,\lambda}$ in a vector form using Φ, y and w, then calculate the gradient with respect to w)

ii. What is the matrix R in this case?

iii. What happens for high values of λ ?

b. Show that the obtained matrix $(\Phi^T\Phi + R)$ is always invertible. (Hint : Show that the obtained matrix $(\Phi^T\Phi + R)$ is a positive defined matrix)

c.

i. What are the optimal weights for $\boldsymbol{\varphi}(\cdot) \equiv (0,\ldots,0)$?

ii. What are the optimal weights for $\boldsymbol{\varphi}(\cdot) \equiv (1,\ldots,1)$? (Hint : use the Sherman-Morrison formula (look in Wikipedia))

d. Given is the following cost function:

$$E_{SSE,\lambda} = \sum_{k=1}^{n} y_k - \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_k)^2 + \lambda \sum_{i=1}^{d} w_i$$

Is a high value of the parameter λ ensures that we will get low values (in absolute value) of the weights?