Homework 3

Submission Instructions

- Homework is due on: **Tuesday 21/08/18 23:55**.
- Homework should be done **only in pairs**. Each pair is to do their own work, separate from the other pairs.
- We prefer you type your submission, however, you may submit scanned handwritten material as long as it is **clear and readable**.
- Submit **only one** PDF file. Please **write your ID** on the top of the file.
- Submission is done via **Moodle** website.

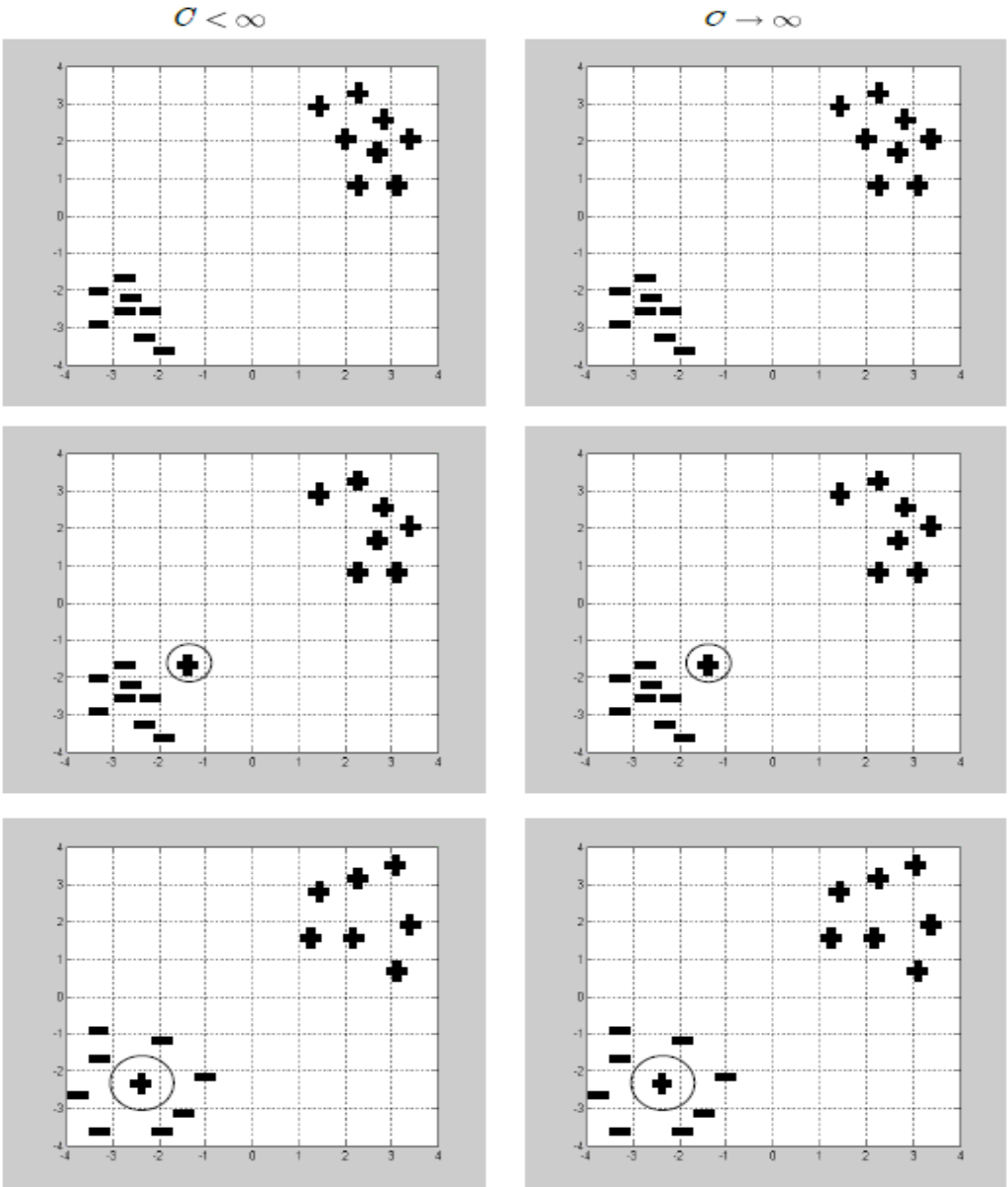
Question 1

(a) Sketch the separating hyperplane for the three datasets below and for two 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 exist, 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.



Question 2

Consider a training set $\{x_i\}_{i=1}^n$, $(x_i) \in \mathbb{R}^n$ with labels $y_i \in \{0, 1\}$ (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 $n + 1$.

Question 3

Consider two kernel functions $k_1, k_2 : X \times X \to \mathbb{R}$. t is known that the classification problem is linearly separable for k_1 but not for k_2. We define a new kernel function as

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

(a) Is $k_3(x, x')$ is a valid kernel function? If yes, then explicitly show that it satisfies the conditions required from a kernel function.

(b) Is the classification problem is linearly separable for $k_3(x, x')$?

Question 4

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

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

Question 5

Consider the following function:

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

(a) Plot this function in the range of $-3 \leq x, y \leq 3$ (You may use MATLAB functions *mesh* and *meshgrid*).

(b) Implement the gradient descent method for finding the minimum point. Attach your code to your submitted pdf file.

(c) Initialize your algorithm with the following values:

- $[x_0, y_0] = [0.1, 1]$.
- $\eta = 0.01$.

Plot 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:

- $[x_0, y_0] = [1.5, -1]$.
- $\eta = 0.05$.

Plot the convergence graph of the algorithm. To what point if any the algorithm converges? Which phenomenon can be observed?

(e) Initialize your algorithm with the following values:

- $[x_0, y_0] = [1.5, -1]$.
- $\eta = 0.01$.

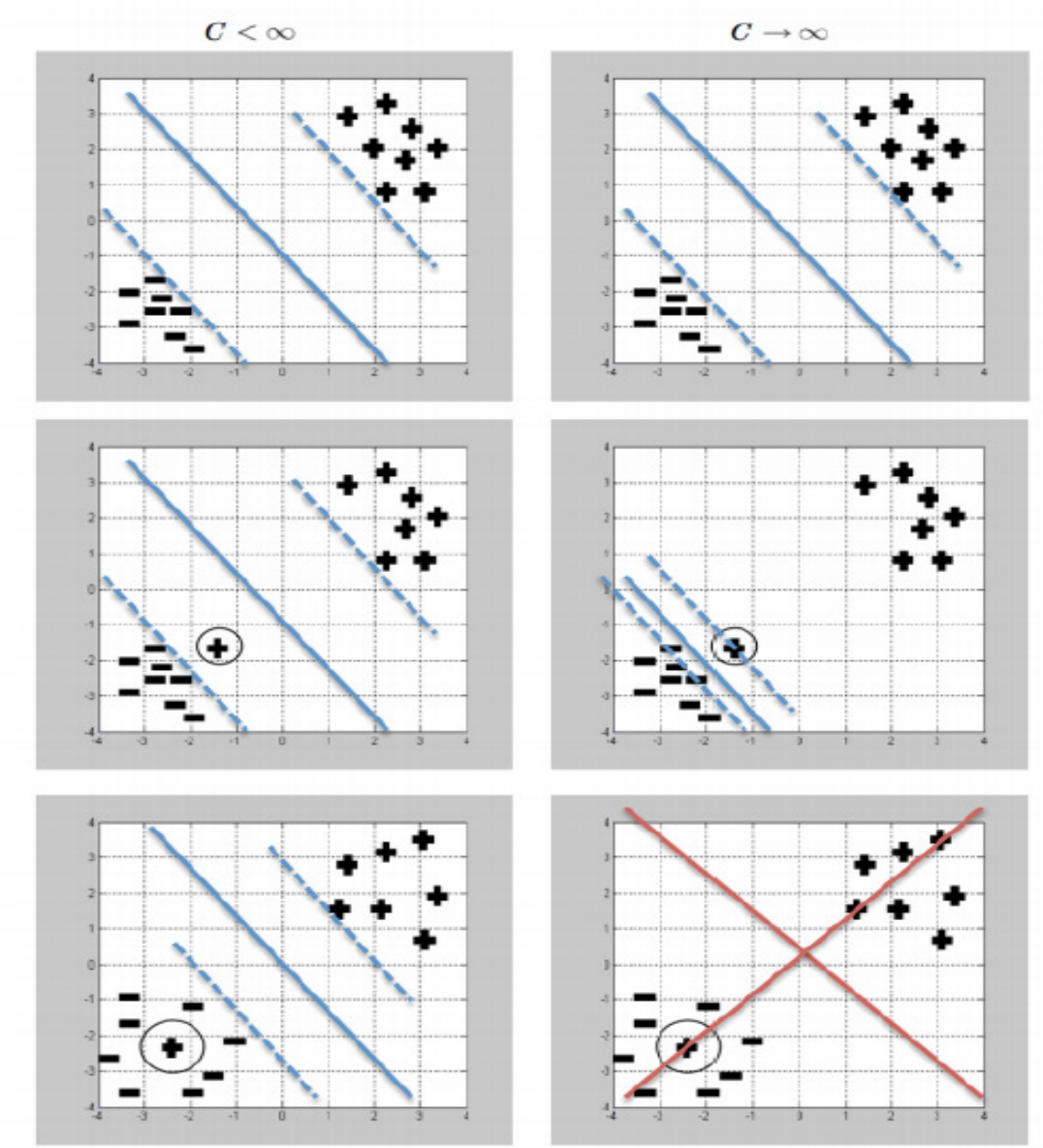
Plot the convergence graph of the algorithm. To what point if any the algorithm converges? Compare your results with the results of part (d).

Solution

Question 1

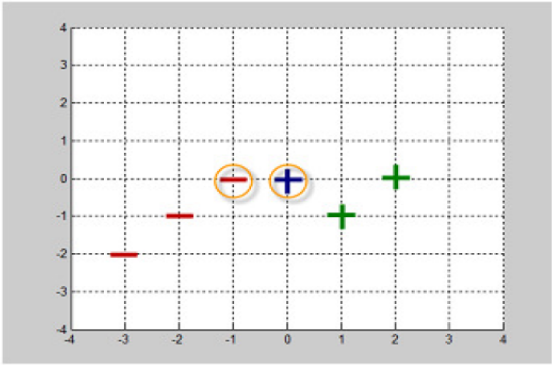
(a) The hyperplane are shown below. The last set points is not separable for $C \to \infty$ since the plus marked in a red circle is surrounded by minus examples.

(b) For the second row, when $C < \infty$ then $< 0\xi < 1$ since the circled plus reside on the negative side of the hyperplane. When $C \to \infty$ then $\xi = 0$. For the last row, when $C < \infty$ then $\xi > 1$. When $C \to \infty$ we would like the error to be $\xi = 0$ but there in no solution as we saw in (a)

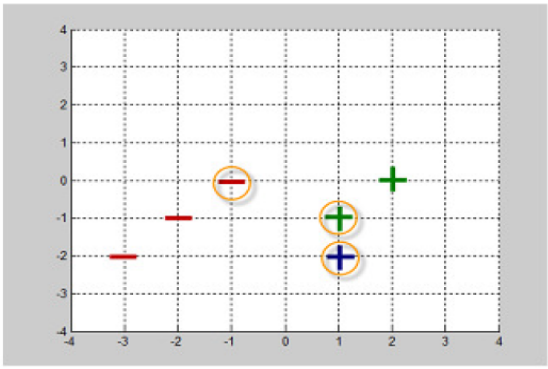


Question 2

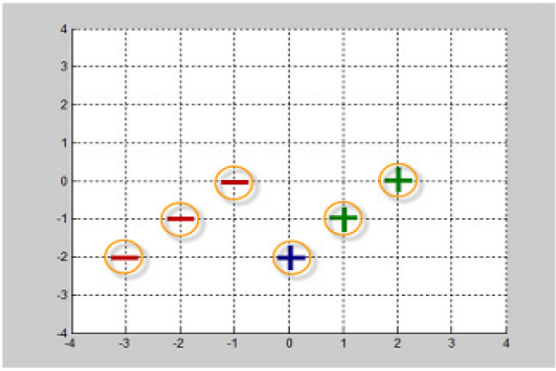
1. Assume that we have 5 original samples points as in Figure 1 without the blue plus. In this case we have two support vectors. Now, the blue point is added. Still we have two support vectors marked in yellow circle.



2. When the blue point is added like in Figure 2, the number of support vectors grows from $k = 2$ to $k + 1 = 3$.



3. When the blue point is added like in Figure 3, the number of support vectors grows to $n + 1 = 6$.



Question 3

(a) A kernel function should satisfy the following conditions:

- Symmetry

$$k_3(x,z) = k_1(x,z) + k_2(x,z) = k_1(z,x) + k_2(z,x) = k_3(z,x).$$

- PSD - since k_1 and k_2 are kernel functions then the matrices K_1 and K_2 defined by $(K_i)_j l = k_i(x_m, x_l)$ are PSD:

$$\begin{aligned} & c^T K_1 c \geq 0,\ c^T K_2 c \geq 0\ \forall c \in \mathbb{R}^n \\ \Rightarrow\ & c^T K_3 c = c^T (K_1 + K_2) c = c^T K_1 c + c^T K_2 c \geq 0\ \forall c \in \mathbb{R}^n. \end{aligned}$$

Therefore, $k_3(x,z)$ is a kernel function.

(b) The problem is linearly separable for k_3:
Denote $k_1(x,z) = \phi_1(x)^T \phi_1(z)$ and $k_2(x,z) = \phi_2(x)^T \phi_2(z)$. Hence, we can write

$$k_3(x,z) = \phi_1(x)^T \phi_1(z) + \phi_2(x)^T \phi_2(z) = [\phi_1(x),\ \phi_2(x)]^T [\phi_1(z),\ \phi_2(z)] \triangleq \phi_3(x)^T \phi_3(z)$$

This implies that the feature vector obtained by the mapping of k_3 contains both the features obtained by the mapping of k_1 and k_2. Since the problem is linearly separable for k_1, then, it will linearly separable also for k_3.

Question 4

(a) Linear SVM - the points are not linearly separable, hence, linear SVM cannot achieve a zero training error.

(b) SVM with a polynomial kernel function of degree 2: The features obtained by a this mapping are: $x_1^2, x_2^2, \sqrt{2}x_1x_2$. Since the mapping function contains a multiplication between the first and the second dimension of the input the classifier will achieve a zero training error.

(c) SVM with a Gaussian kernel function $K_\lambda(x,z) = e^{-\frac{||x-z||^2}{\lambda}}$: For a sufficiently large λ, the kernel matrix approaches the identity matrix. Thus, for each sample from the training set:

$$w^T \phi(x_i) = \sum_k \alpha_k y_k K(x_i, x_k) \approx a_i y_i K(x_i, x_i) = \alpha_i y_i.\ \Rightarrow sign(\alpha_i y_i) = y_i.$$

Therefore a zero training error is achieved.