Gary Chen
804361027

Homework 4

EE M146
May 3, 2019

Problem 1

Given:

$$J(\tilde{W}) = \frac{1}{2}Tr\left[(\tilde{X}\tilde{W} - T)^T(\tilde{X}\tilde{W} - T)\right]$$

a) **Find the closed form solution of $\tilde{W}$ that minimizes the objective function $J(\tilde{W})$.**

Using definitions of the derivative of the trace: $\frac{\partial}{\partial Z}Tr(AZ) = A^T$ and $\frac{\partial}{\partial Z}Tr(Z^TAZ) = (A^T + A)Z$ onto an expanded $J(\tilde{W})$:

$$\begin{aligned} J(\tilde{W}) &= \frac{1}{2}Tr\left[\tilde{W}^T\tilde{X}^T\tilde{X}\tilde{W} - T^T\tilde{X}\tilde{W} - \tilde{W}^T\tilde{X}^TT + T^TT\right] \\ &= \frac{1}{2}\left[Tr\left[\tilde{W}^T\tilde{X}^T\tilde{X}\tilde{W}\right] - Tr\left[T^T\tilde{X}\tilde{W}\right] - Tr\left[\tilde{W}^T\tilde{X}^TT\right] + Tr\left[T^TT\right]\right] \end{aligned}$$

$$\begin{aligned} \frac{\partial J(\tilde{W})}{\partial \tilde{W}} = 0 &= \frac{1}{2}\left[\left[(\tilde{X}^T\tilde{X})^T + (\tilde{X}^T\tilde{X})\right]\tilde{W} - \tilde{X}^TT - T^T\tilde{X}\right] \\ \tilde{W} &= \left(\tilde{X}^TT\right)\left(\tilde{X}^T\tilde{X}\right)^{-1} \end{aligned}$$

b) **Show that $J(\tilde{W})$ has a unique minimum**

$$\begin{aligned} \frac{\partial^2 J(\tilde{W})}{\partial \tilde{W}^2} &= \tilde{X}^T\tilde{X} \\ &= \sum_k^K x_k^2 \\ &\geq 0 \end{aligned}$$

Problem 2

Show that the kernel function $K(x, x')$ satisfies the following generalization of the Cauchy-Schwartz inequality:

$$K(x_1, x_2)^2 \leq K(x_1, x_1)K(x_2, x_2)$$

The kernel function is the inner product of the feature maps of x and x':

$$K(x, x') = \langle \phi(x)\phi(x') \rangle = \phi(x)^T \phi(x') \tag{1}$$

Using the definition of the generalization of the Cauchy-Schwartz inequality:

$$\begin{aligned} |\phi(x)^T \phi(x')|^2 &\leq (\phi(x)^T \phi(x))(\phi(x')^T \phi(x')) \\ &\leq \|\phi(x)\|^2 \|\phi(x')\|^2 \end{aligned}$$

We see that the statement is true given the Cauchy-Schwartz inequality for vectors:

$$|u^T v|^2 \leq \|u\|^2 \|v\|^2$$

Problem 3

Given valid kernels $K_1(x, x')$ and $K_2(x, x')$, show that the following kernels are also valid:

a) $K(x, x') = K_1(x, x') + K_2(x, x')$

Let $K_i(x, x') = \phi_i^T \phi_i$, and $\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix}$ be the feature mapping for $K(x, x')$.

$$\begin{aligned} K(x, x') &= \Phi^T \Phi \\ &= \begin{bmatrix} \phi_1^T & \phi_2^T \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} \\ &= \phi_1^T \phi_1 + \phi_2^T \phi_2 \\ &= K_1(x, x') + K_2(x, x') \end{aligned}$$

We can show that it is a kernel To be a valid kernel, K must be symmetric and satisfy the Mercer condition $z^T K z \geq 0$ where z is an arbitrary vector.

$$z^T K(x, x') z = z^T K_1(x, x') z + z^T K_2(x, x') z$$

Where $z^T K_{1,2} z \geq 0$ since they are valid kernels.

$$\therefore z^T K(x, x') z \geq 0$$

Thus, $K(x, x')$ is a kernel with feature mapping Φ.

b) $K(x, x') = K_1(x, x')K_2(x, x')$

We start by writing out the inner products as explicit summations.

$$\begin{aligned} K(x, x') &= K_1(x, x')K_2(x, x') \\ &= \sum_i \sum_j \phi_i^1(x_1)\phi_i^1(x_1')\phi_j^2(x_2)\phi_j^2(x_2') \\ &= \sum_i \sum_j \phi_i^1(x_1)\phi_j^2(x_2)\phi_i^1(x_1')\phi_j^2(x_2') \end{aligned}$$

Where we define $K(x, x') = \sum_i \sum_j \Phi_{ij}(x)\Phi_{ij}(x')$, where $\Phi_{ij}(x) = \phi_i^1(x_1)\phi_j^2(x_2)$. This defines a kernel with feature map Φ.

c) $K(x, x') = \exp(K_1(x, x'))$

Taking a Taylor expansion of $K(x, x')$:

$$\exp(K(x, x')) = \sum_{n=1}^{\infty} \frac{K_1(x, x')^n}{n!}$$

We find a summation of powers of valid kernels, where each component has been proven to be a kernel, and the sum of kernels has been proven to be a kernel.

Problem 4

Show that the parameter b can be determined using the following equation:

$$b = \frac{1}{N_M} \sum_{n \in M} \left(y^{(i)} - \sum_{m \in S} \alpha_m y^{(m)} \langle x^n, x^m \rangle \right)$$

We start with the w condition and Lagrangian where $i = 1, \ldots, m$:

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \epsilon_i \tag{2}$$

$$\mathcal{L} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \epsilon_i + \sum_{i=1} \alpha_i(1 - \epsilon_i - y^{(i)}(w^T x^{(i)} + b)) - \sum_{i=1}^{m} \gamma_i \epsilon_i \tag{3}$$

We consider the components for $i \in S$ that contribute to defining the boundary where $\alpha_n \neq 0$. Being on the boundary, we find that the left hand side of equation 2 equals 1, such that $\epsilon_i = 0$, points with no slack.

$$\therefore b = \frac{1}{y^{(i)}} - w^T x^{(i)}$$

We note that $y^{(i)} = \pm 1$, thus, $1/y^{(i)} = y^{(i)}$.

$$b = y^{(i)} - w^T x^{(i)} \tag{4}$$

From the derivative of the Lagrangian with respect to the vector w, we find:

$$\frac{\partial \mathcal{L}}{\partial w} = 0 = w - \sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)}$$
$$w = \sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)}$$
$$w^T = \sum_{j=1}^{m} \alpha_j y^{(j)} (x^{(j)})^T$$

Given the condition we used to simplify equation 2, we only sum over the components of α_j, $y^{(j)}$, and $x^{(j)}$ for $j \in S$ and make a change of indices to avoid confusion ($i \to n, j \to m$).

$$b = y^{(n)} - \sum_{j \in S} \alpha_m y^{(m)} \langle x^{(n)}, x^{(m)} \rangle$$

This holds true for all points within the boundaries $n \in M$ where M is the set of indices where $0 < \alpha_n < C$. We can then average over all values in M and not change the result.

$$b = \frac{1}{N_M} \sum_{n \in M} \left(y^{(n)} - \sum_{m \in S} \alpha_m y^{(m)} \langle x^{(n)}, x^{(m)} \rangle \right)$$

Problem 5

Given 6 data points: 3 with negative labels: $x_1 = -1, x_2 = 0, x_3 = 1$, and 3 with positive labels $x_4 = -3, x_5 = -2, x_6 = 3$.

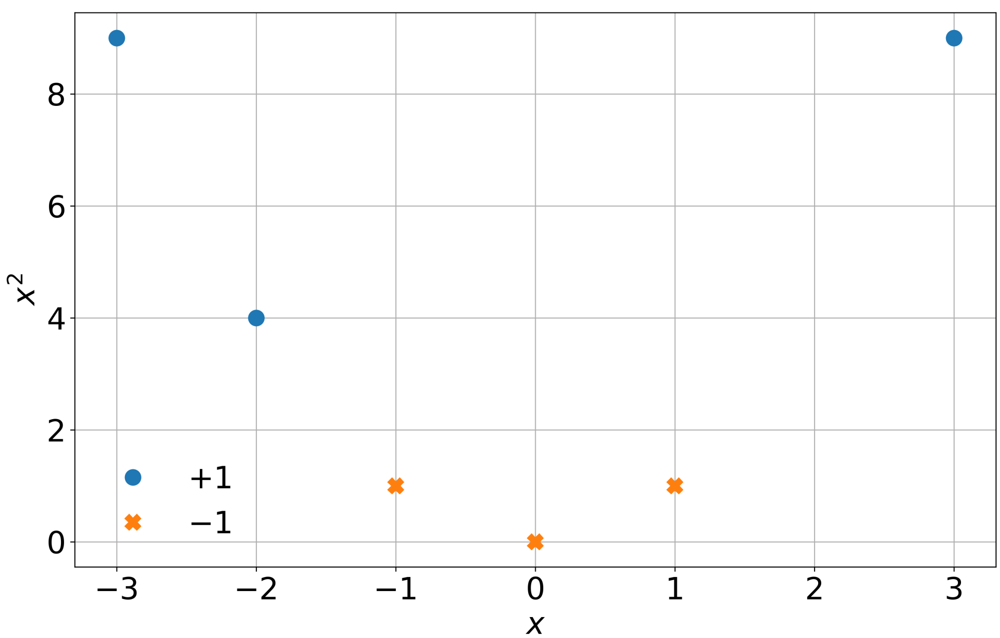
a) **Consider a linear classifier of form $f(x) = \textbf{sign}(w_1 x + w_0)$. Write down the optimal value of w and its classification accuracy with the 6 data points.**

Setting $w_1 = -1$ and $w_0 = -1.5$, we get an accuracy of 5/6.

b) **Given two samples x and z in $\mathbb{R}$, define the kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as $K(x, z) = xz(1 + xz)$, find the corresponding feature map $\phi(x)$.**

$$\begin{aligned}\phi(x)\phi(z) &= xz(1 + xz) \\ &= xz + x^2z^2 \\ &= \begin{bmatrix} x & x^2 \end{bmatrix} \begin{bmatrix} z \\ z^2 \end{bmatrix} \\ \therefore \phi(x) &= \begin{bmatrix} x \\ x^2 \end{bmatrix}\end{aligned}$$

c) **Apply $\phi(x)$ to the data and plot the points in the induced feature space $\mathbb{R}^2$. Are these points linearly separable now?**



After using the feature space, the points are now linearly separable.

d) **Draw a maximum margin hyperplane that can be parameterized by $w_1\phi_1(x) + w_2\phi_2(x) + w_0 = 0$ and circle the support vectors.**

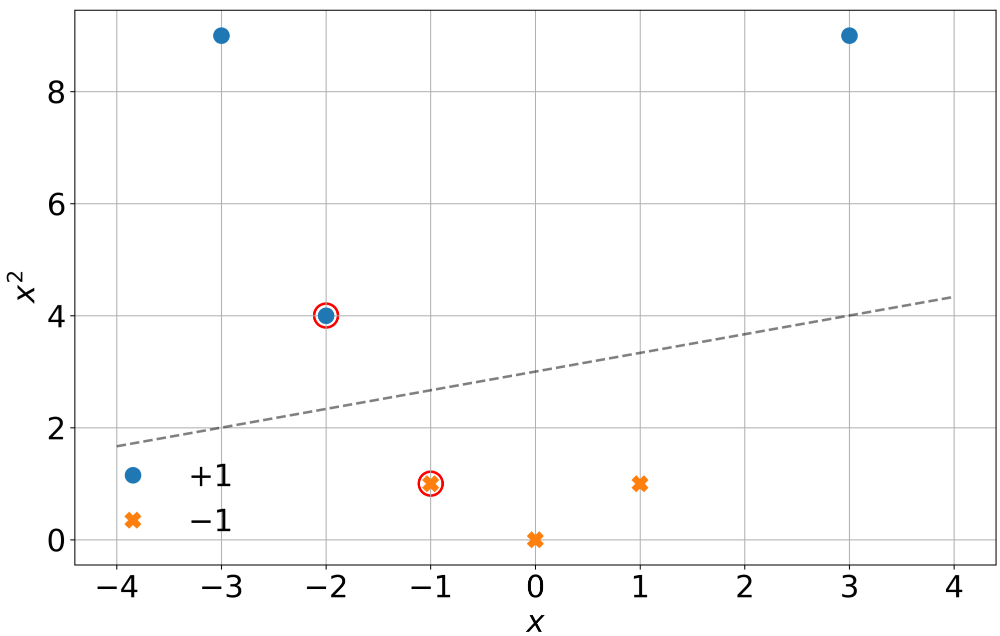
We start by picking the points that are closest together as support vectors:

$$\phi(x_1) = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \phi(x_2) = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$$

We bisect the vector between the two points to find the line perpendicular to both, being:

$$-0.2\phi_1 + 0.6\phi_2 - 1.8 = 0$$

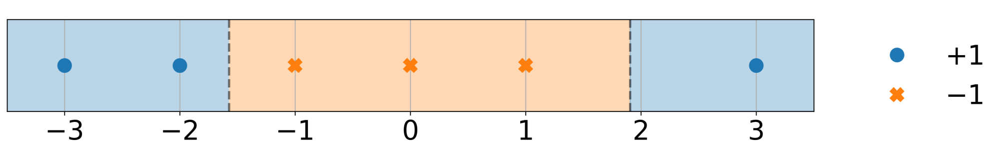
Where $w = \begin{bmatrix} -0.2 \\ 0.6 \end{bmatrix}$ and $b = -1.8$.



e) **Draw the decision boundary of the separating hyperplane you found in (d) in the original $\mathbb{R}$ feature space.**

To map the decision boundary onto the original feature space, we solve for the intersection of the boundary on the two spaces.

$$\begin{aligned} 0 &= w^T\phi(x) + b \\ 0 &= w_0 x + w_1 x^2 + b \\ \therefore x &= \frac{-w_0 \pm \sqrt{w_0^2 - 4w_1 b}}{2w_1} \\ &= 1.91, -1.57 \end{aligned}$$



f) **Find α_i, w, and b in**

$$h(x) = \text{sign}\left(\sum_{n\in S} \alpha_n y_n K(x_n, x) + b\right) = \text{sign}(w^T\phi(x) + b)$$

Do this by solving the dual form of the quadratic program. How are w and b related to your solution in part (d)?

With the dual problem:

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle \tag{5}$$

And values:

$$x_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad y_1 = -1$$
$$x_2 = \begin{bmatrix} -2 \\ 4 \end{bmatrix} \quad y_2 = 1$$

We find the condition:

$$\begin{aligned} \sum_{i=1}^{n} \alpha_i y_i &= 0 \\ \alpha_1(-1) + \alpha_2(1) &= 0 \\ \alpha_1 &= \alpha_2 \end{aligned}$$

Where we let $\alpha = \alpha_1 = \alpha_2$. Thus, plugging this back into equation 5:

$$W(\alpha) = \alpha + \alpha - \frac{1}{2}\left[(-1)\alpha^2\langle x_1, x_2\rangle + (-1)\alpha^2\langle x_2, x_1\rangle + (1)\alpha^2\langle x_1, x_1\rangle + (1)\alpha^2\langle x_2, x_2\rangle\right]$$

Where $\langle x_1, x_2\rangle = 6$, $\langle x_1, x_1\rangle = 2$, and $\langle x_2, x_2\rangle = 20$

$$\begin{aligned} &= 2\alpha - \frac{1}{2}\left[-12\alpha^2 + 2\alpha^2 + 20\alpha^2\right] \\ &= 2\alpha - 5\alpha^2 \end{aligned}$$

Maximizing:

$$\frac{\partial W(\alpha)}{\partial \alpha} = 0 = 2 - 10\alpha$$
$$\alpha = \frac{1}{5}$$

To then find w, we use the form:

$$\begin{aligned} w &= \sum_{i=1} \alpha_i y_i x_i \\ &= \frac{1}{5}\left((-1)\begin{bmatrix}-1\\1\end{bmatrix} + (1)\begin{bmatrix}-2\\4\end{bmatrix}\right) \\ &= \frac{1}{5}\begin{bmatrix}-1\\3\end{bmatrix} \\ &= \begin{bmatrix}-0.2\\0.6\end{bmatrix} \end{aligned}$$

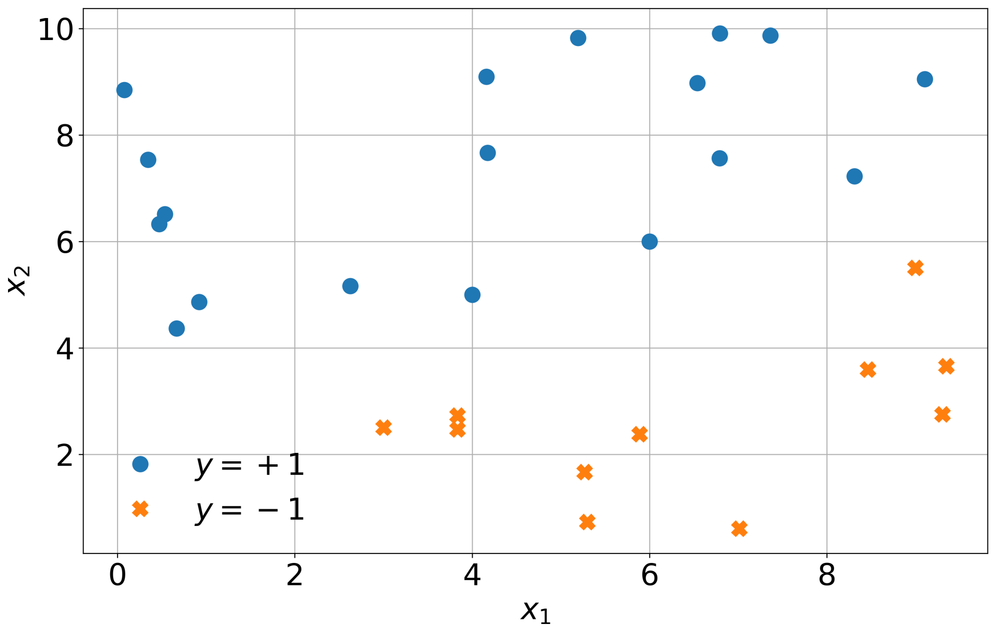
Considering the sign condition for $i = 1$:

$$\begin{aligned} 1 &= y_1(w^T x_1 + b) \\ 1 &= (-1)\left(\begin{bmatrix}-0.2 & 0.6\end{bmatrix}\begin{bmatrix}-1\\1\end{bmatrix} + b\right) \\ -1 &= 0.8 + b \\ b &= -1.8 \end{aligned}$$

These values are identical to the ones found in part (d).

Problem 6

a) **Visualization: Use different color to plot data with different labels in the 2-D feature space. Is the data linearly separable?**



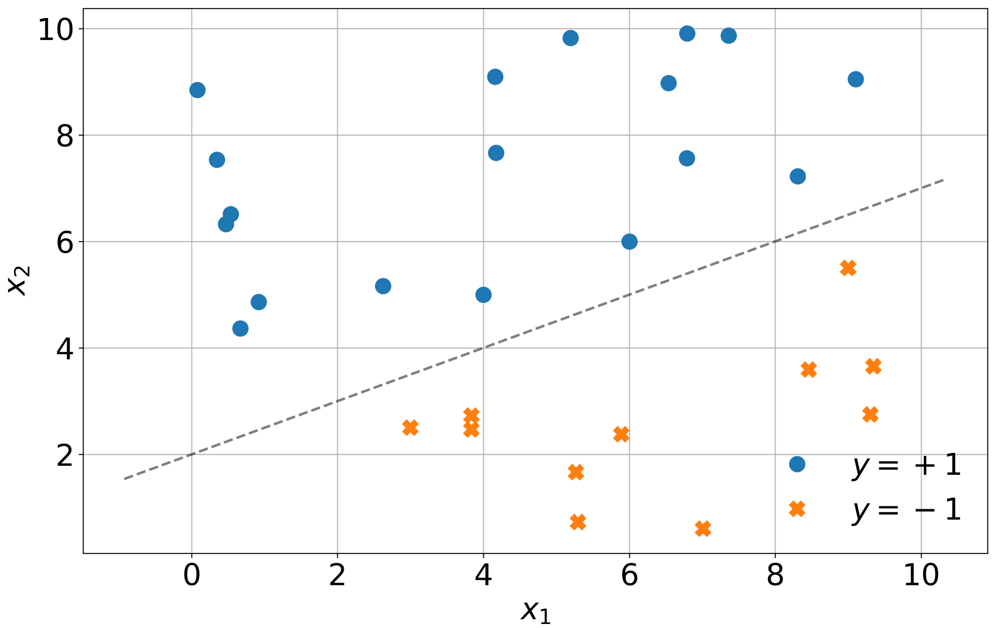
The data looks like it is linearly separable.

b) **The Primal Problem: Use CVX to solve the primal problem of this form:**

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1,\ i = 1, \ldots, m \end{aligned}$$

Report w and b. Plot the hyperplane defined by w and b.

We find the hyperplane to be defined by $w = [-0.5, 1]$ and $b = -2$



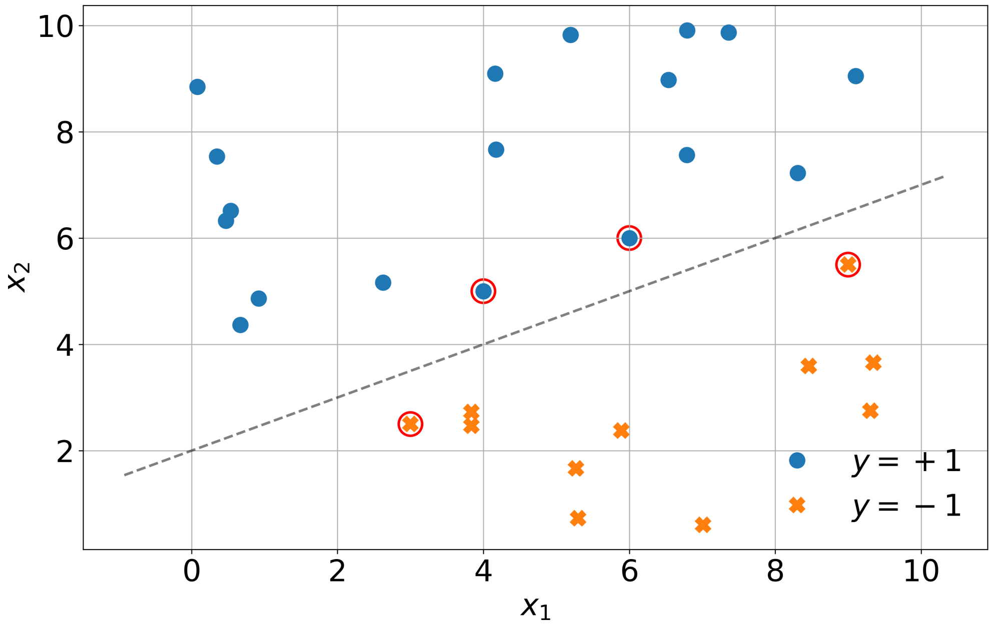
c) **The Dual Problem: Use CVX to solve the dual problem of this form:**

$$\begin{aligned} \min_{w,b} \quad & W(\alpha) = \textstyle\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & 0 \le \alpha_i, i = 1, \ldots, m \\ & \textstyle\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \end{aligned}$$

Use the resulting α to identify the support vectors on the plot. Report your nonzero α_i's is. How many support vectors do you have? Circle those support vectors.

We find that the second summation term in the dual problem can be written as $z^T P z$ where $z \equiv \sum_i \alpha_i y_i$ and $P = \langle x^{(i)}, x^{(j)} \rangle$. P is a kernel, which is convex and can thus be solved with the cvxpy quad form.

We find $a_i = 0.38, 0.24, 0.46, 0.16$, meaning that there are 4 support vectors


10
8
6
4
2
x_2
0
2
4
6
8
10
x_1
$y = +1$
$y = -1$