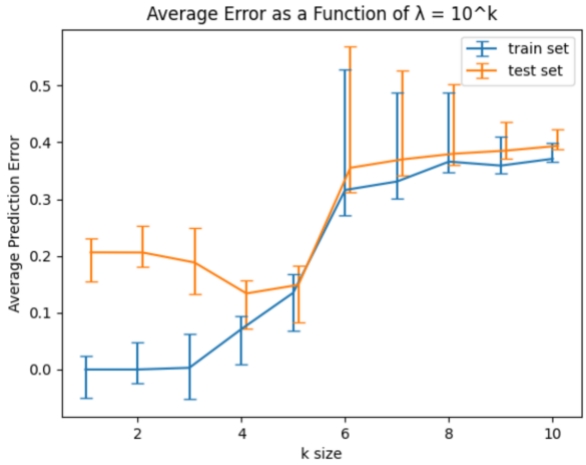
Exercise 2

Introduction to learning and analysis of big data

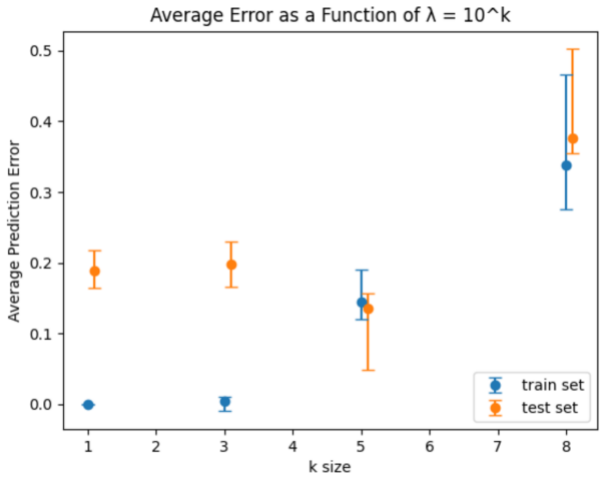
Nadav Shaked Kim Teldan

Question 2

a.



b.

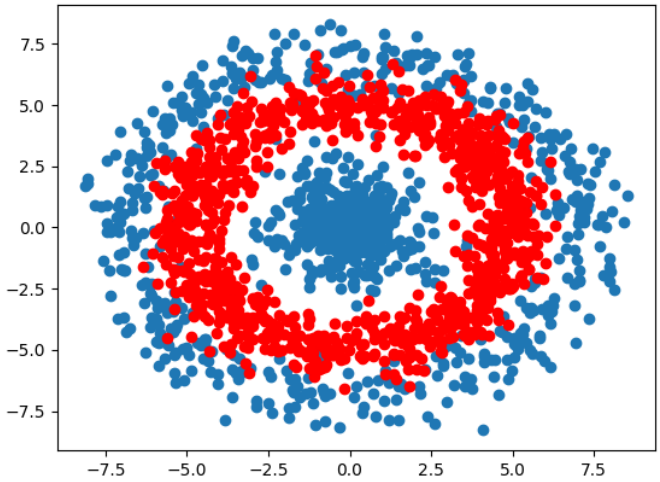


c. The results are match our expectation, as we expected smaller sample size will get a smaller training error and larger test error, As the sample size gets larger it may get harder to find a good seperator. As the λ increases the training error will increase too, because as λ increases - $\|w\|^2$ will have more effect and because of that we will aim to a smaller norm of w which brings to a bigger margin, so the hinge loss will have less effect, it causes over-fitting.

For the test set the results are non-monotonic (not increasing but also not decreasing), there is a trade-off between the punishment of norm of w, and the hinge-loss function. we can see the over-fitting of the training set if we compare it with the test set.

Question 4

a.



It is better to use kernel-svm, because as can be seen in the points scattering there is no linear predictor that can separate the points into two clearly-distinct areas.
If we will use basic soft-svm, there will be no linear predictor and possibly high loss that will cause high error. When we use kernel-svm we use a feature that increases the dimension of the problem - so we can separate the points by linear predictor in a better way. As seen in class, we will see next that gaussian-kernel will lead to much better results.

b.

$\lambda \setminus \sigma$	0.01	0.5	1
1	0.07999999999999999	**0.06549999999999999**	0.0875
10	0.07999999999999999	0.06549999999999999	0.0875
100	0.07999999999999999	0.06549999999999999	0.0875

For the best parameters $\lambda = 1$ and $\sigma = 0.5$, the test average error is: 0.045
Also, there are more parameters that has the same error.

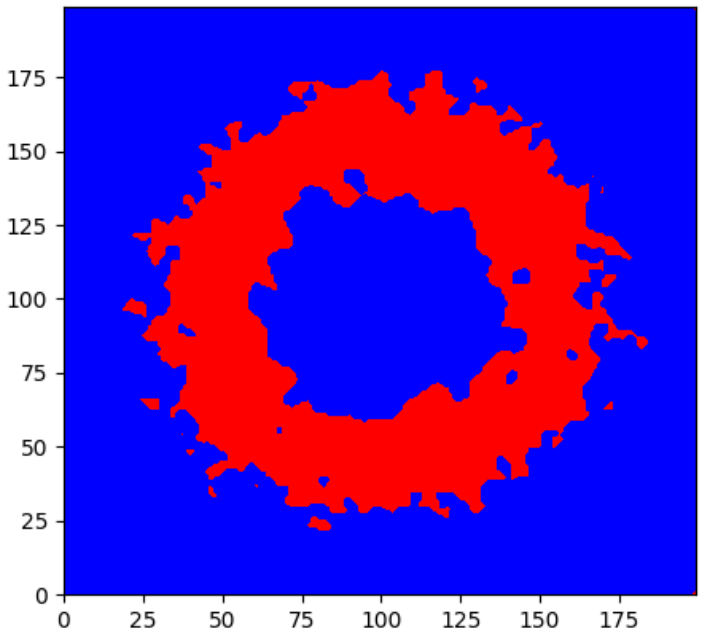
λ	Average Error
1	**0.49976750000000003**
10	0.49976750000000003
100	0.49976750000000003

For the best parameter $\lambda = 1$, the test average error is: 0.49976750000000003
Also, there are more parameters that has the same error.

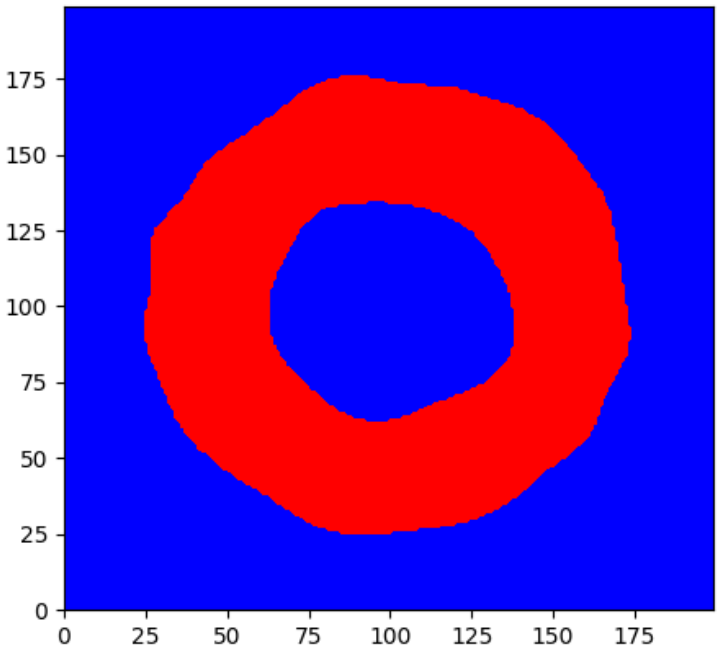
c. RBF gave us a much better validation error, as expected. RBF might give us a better validation error because it is able to find separators in the higher dimensions, in many cases where there is no separator in the original dimension. Therefore, its approximation error is much lower than regular soft-SVM's.

On the other hand, RBF might give us a worse validation error, because it chooses a higher-dimension separator out of many options (much bigger hypothesis class), a separator that might not work specifically for the validation sample (higher estimation error).

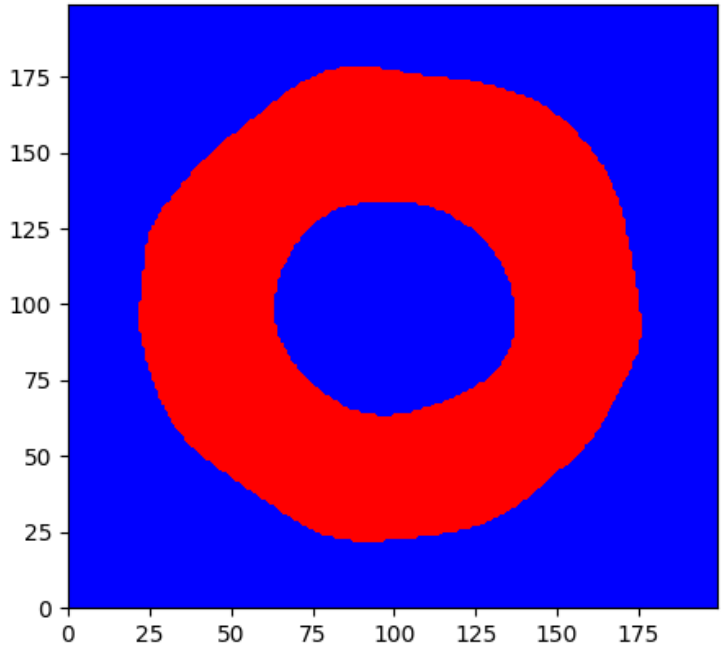
d. Grid for $\sigma = 0.01$



Grid for $\sigma = 0.5$



Grid for $\sigma = 1$



e. The difference we observe is that the lower the value ,the more fine-tuned the heatmap is, and the more the grid points match the closest sample points to them (larger estimation error). The lower the sigma, the lower effect far examples will have on our separator. Hence, our separator will be more effected by closer examples, and to our specific training set (over-fitting).

Question 5

For training sample $S = \big((x_1, y_1), \dots, (x_m, y_m)\big)$
We want to sole the soft-SVM optimization problem:

$$Minimize_{w \in \mathbb{R}^d}\ \lambda \|w\|^2 + \sum_{i=1}^{m} l^h\big(w, (x_i, y_i)\big)^2$$
$$s.t\ l^h\big(w, (x_i, y_i)\big) = \max\{0, 1 - y_i \langle w, x_i \rangle\}$$

a. A minimalization problem that equivalent to the problem above is:

$$Minimize_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^d}\ \lambda \|w\|^2 + \sum_{i=1}^{m} \xi_i^2$$
$$s.t\ \forall i, y_i \langle \alpha, G[i] \rangle \geq 1 - \xi_i, \xi_i \geq 0$$

Unlike the original soft-SVM we learned in class in this problem each product in the sigma always positive so we can forgive on the constraints of $\xi_i \geq 0$.

$l^h\big(w, (x_i, y_i)\big) = \max\{0, 1 - y_i \langle w, x_i \rangle\}$
so
$1 - y_i \langle w, x_i \rangle \leq \xi_i$
$y_i \langle w, x_i \rangle \geq 1 - \xi_i$

From the constraint we know that $\forall i,\ \xi_i \geq \max\{0, 1 - y_i\langle w, x_i\rangle\} = l^h\big(w, (x_i, y_i)\big)$
And because $l^h\big(w, (x_i, y_i)\big)$ is non-negonents, then:
$\xi_i^2 \geq (\max\{0, 1 - y_i\langle w, x_i\rangle\})^2 = l^h\big(w, (x_i, y_i)\big)^2$
So the optimal value we want to return is $\xi_i^2 = l^h\big(w, (x_i, y_i)\big)^2$

b. To solve the soft-SVM optimization problem

$$Minimize_{w\in\mathbb{R}^d, \xi\in\mathbb{R}^d}\ \lambda\|w\|^2 + \sum_{i=1}^{m}\xi_i^2$$

by quadratic problem solver

$$Minimize_{z\in\mathbb{R}^n}\frac{1}{2}z^T * H * z + \langle u, z\rangle$$
$$s.t. Az \geq v$$

Note: we no longer have linear components in our objective - ξ_i, are quadratic.
We have to set:

u: $(0, \dots, 0) \in \mathbb{R}^{m+d}$

v: $\left(\underbrace{0, \dots, 0}_{d\ times}, \underbrace{1, \dots, 1}_{m\ times}\right) \in \mathbb{R}^{m+d}$

H: $\begin{bmatrix} 2\lambda * I_d & 0 \\ 0 & 2 * I_m \end{bmatrix} \in \mathbb{R}^{(m+d)x(m+d)}$

A: $\begin{bmatrix} 0 & I_m \\ xy^* & I_m \end{bmatrix} \in \mathbb{R}^{(2m)x(m+d)}$

$s.t. \forall i, row\ i\ of\ xy^*\ is\ y_iG_{i,1}, y_iG_{i,2}, \dots, y_iG_{i,m}$

z: $(w_1, \dots, w_d, \xi_1, \dots, \xi_m) \in \mathbb{R}^{m+d}$

Question 6

To solve this optimization problem by the representer theorem

$$Minimize_w\ f(\langle w, \psi(x_i)\rangle, \dots, \langle w, \psi(x_m)\rangle) + R(\|w\|)$$

We have to satisfy:

f: $\mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$

R: $\mathbb{R}_+ \to \mathbb{R}\ is\ a\ monotonic\ non\ decreasing\ function$

In our case we want to solve

$$Minimize_{w\in\mathbb{R}^d} r * \|w\|_2^4 + \sum_{i=1}^{m}\exp^{|\langle w, x_i\rangle - y_i|}$$

To satisfy the objectives we can define:

f: $f(\langle w, \psi(x_i)\rangle, \dots, \langle w, \psi(x_m)\rangle) = \sum_{i=1}^{m}\exp^{|\langle w, x_i\rangle - y_i|}$
$f: \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$

R: $R(x) = r * x^4, for\ any\ x > 0$
$R: \mathbb{R}_+ \to \mathbb{R}$

In our case any input to R is non-negative because $\|w\|_2$ is non-negative for any w
We want a monotonic non-decreasing function, so we have to demand $R'(x) \geq 0$,

We will find which are satisfy the equation:

$$3r * x^3 \geq 0$$
$$x \geq 0, so\ r \geq 0$$

Therefore for every $r \geq 0$, the representer theorem is satisfied, for the objective.

Question 7

a. Let $X = \mathbb{R}^d$, and $\forall x, x' \in X, K(x, x') = -x_1 * x_1'$
Assume in contradiction that exist ψ such that $K(x, x') = \langle \psi(x), \psi(x') \rangle$
For $x_1 = x_1' \neq 0$
$\langle \psi(x), \psi(x') \rangle = \langle \psi(x), \psi(x) \rangle = \|\psi(x)\|^2 > 0$
But
$K(x, x') = K(x, x) = -x_1 * x_1 < 0$
Contradiction.

b. Let $X = \mathbb{R}^d$, and $\forall x, x' \in X, K(x, x') = (x_1 + x_2)(x_3' + x_4')$
Assume in contradiction that exist ψ such that $K(x, x') = \langle \psi(x), \psi(x') \rangle$
Let $x = (1, 2, 0, \dots, 0), x' = (0, 0, 3, 3, 0, \dots, 0)$
So
$18 = K(x, x') = \langle \psi(x), \psi(x') \rangle \doteq \langle \psi(x'), \psi(x) \rangle = K(x', x) = 0$
Contradiction.
* Inner product is symmetric.

c. Kernel Perceptron Algorithm:
Input $A\ Gram\ matrix\ G\ where\ G_{i,j} = K(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle$
Output $\alpha \in \mathbb{R}^m \quad s.t. \forall i \leq m, w = \sum_{i=1}^{m} \alpha_i \psi(x_i)$

1. $t \leftarrow 1, \alpha^1 = (0, \dots, 0) \in \mathbb{R}^m$
2. ***while*** $\exists i\ s.t.\ y_i\left(\sum_{i=1}^{m} \alpha_i^t\ G_{i,j}\right) \leq 0\ do$ (*)
3. $\alpha_i^{t+1} \leftarrow \alpha_i^t + y_i$
4. $t \leftarrow t + 1$
5. ***end while***
6. $return\ \alpha^t$

Explanation:
When we got an example i that satisfy the $y_i\left(\sum_{i=1}^{m} \alpha_i^t\ G_{i,j}\right) \leq 0$, we want to maintain α, so

$$y_i \sum_{j=1}^{m} \alpha_j^{t+1} * K(x_i, x_j) = y_i \left(\sum_{i=1}^{m} \alpha_j * K(x_i, x_j) + y_i * K(x_i, x_j) \right)$$
$$= y_i \sum_{i=1}^{m} \alpha_j * K(x_i, x_j) + y_i^2 * K(x_i, x_j) = y_i \sum_{i=1}^{m} \alpha_j * K(x_i, x_j) + K(x_i, x_j)$$

So each update will make it closer to positive, which will reduce the error on the sample set.

(*)

$$y_j\langle w, \psi(x_j)\rangle \leq 0$$

$$y_j \langle \sum_{i=1}^{m} \alpha_i * \psi(x_i)\,, \psi(x_j)\rangle \leq 0$$

$$y_j \left(\sum_{i=1}^{m} \langle \alpha_i * \psi(x_i), \psi(x_j)\rangle \right) \leq 0$$

$$y_j \left(\sum_{i=1}^{m} \alpha_i * \langle \psi(x_i), \psi(x_j)\rangle \right) \leq 0$$

$$y_j \left(\sum_{i=1}^{m} \alpha_i * K(x_i, x_j) \right) \leq 0$$

$$y_j \left(\sum_{i=1}^{m} \alpha_i * G_{i,j} \right) \leq 0$$