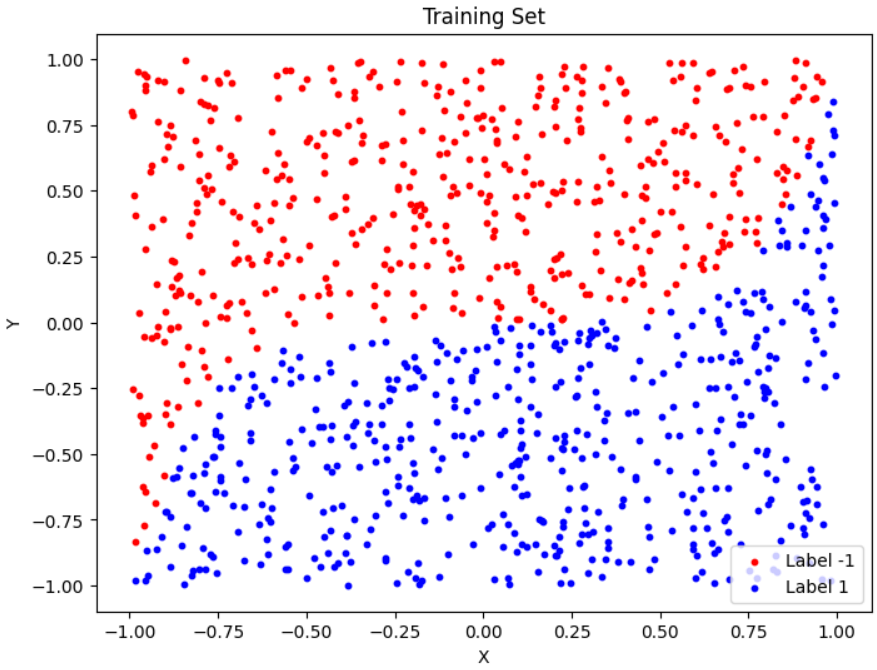
Introduction to Machine Learning

Exercise 3

Question 2:

a.



As we can see in this graph, the boundary between the 1 label (blue) and the -1 label (red) is not linear. This means these samples are not separatable by a linear soft SVM predictor, and there will be errors greater than 0, no matter what lambda we use. The kernel soft SVM increases the dimension of the training input, which helps it capture the complexity of the boundary, and it can create a separatable feature space, and consequently decrease the error.

b. The average validation error values:

- $\lambda = 1,\ k = 2$:

  error = 0.07300000000000001
- $\lambda = 1,\ k = 5$:

  error = 0.006999999999999999
- $\lambda = 1,\ k = 8$:

  error = 0.005
- $\lambda = 10,\ k = 2$:

  error = 0.07100000000000001

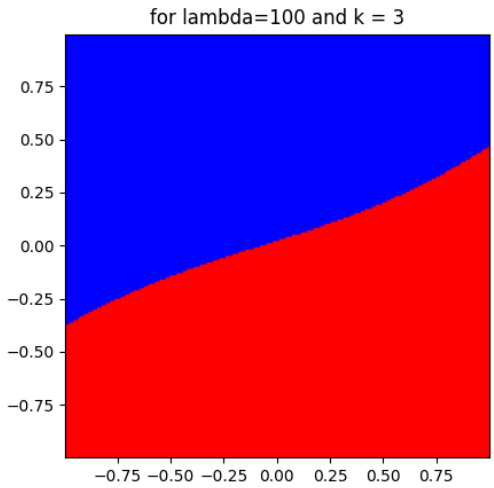
- $\lambda = 10,\ k = 5$, error = 0.015000000000000003
- $\lambda = 10,\ k = 8$, error = 0.009
- $\lambda = 100,\ k = 2$, error = 0.061
- $\lambda = 100,\ k = 5$, error = 0.039999999999999994
- $\lambda = 100,\ k = 8$, error = 0.013000000000000001

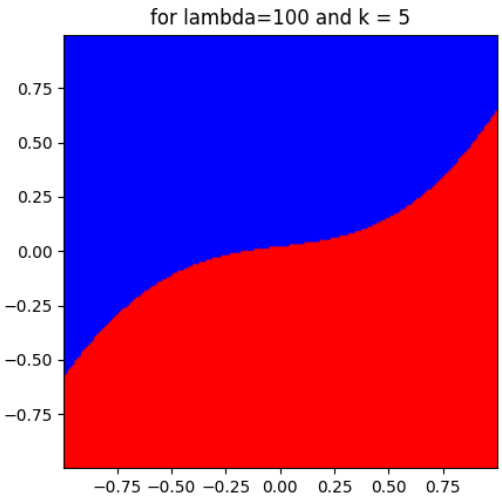
The best result was with $\lambda = 1,\ k = 8$ with an error in 0.005

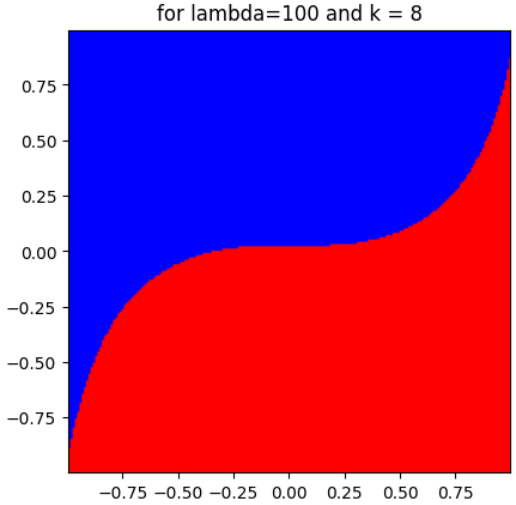
The error rate for the test data with $\lambda = 1,\ k = 8$ is: 0.01

c. Polynomial soft SVM might get a better validation error than linear soft SVM since the polynomial kernels can capture a complex boundary between labels, a nonlinear boundary as we have in this assignment will be separated better using nonlinear kernel function, that will give higher dimensions to the sample data. On the contrary, Linear soft SVM might have a better validation error than polynomial soft-SVM since polynomial soft SVM might be overfitting to the sample data, due to it capturing noise and not being able to generalize.

d.







Question 3:

a. Let us assume that $K(x,x') \coloneqq \big(2x(7)+x(3)\big)\cdot x'(2)$ is a kernel function, meaning there exists such mapping $\psi(x)$ such that $K(x,x') = \langle\psi(x),\psi(x')\rangle$, $\psi\colon \mathbb{R}^d \to \mathbb{R}^k$.
$x = (0,0,1,0,0,0,1)$ and $x' = (0,1,0,0,0,0,0)$. Now we place them in the formula:
$K(x,x') = \big(2x(7)+x(3)\big)\cdot x'(2) = (2\cdot 1+1)\cdot 1 = 3 \neq 0 = (2\cdot 0+0)\cdot 0 = K(x',x)$
A kernel function must be symmetrical, hence we reached a contradiction.

b. Let us assume that $K(x,x') \coloneqq 5 - \big(x(1) - x(2)\big)\cdot\big(x'(1) - x'(2)\big)$ is a kernel function, meaning there exists such mapping $\psi(x)$ such that $K(x,x') = \langle\psi(x),\psi(x')\rangle$, $\psi\colon \mathbb{R}^d \to \mathbb{R}^k$.
we will use $x = x'$ so $K(x,x) = \langle\psi(x),\psi(x)\rangle = ||\psi(x)||_2 \geq 0$. We place $x = (0,3)$ so

$$K(x,x) = 5 - \big(x(1) - x(2)\big)\cdot\big(x'(1) - x'(2)\big) =$$

$$5 - \big(x(1)-x(2)\big)^2 = 5 - (-3)^2 = -4 < 0$$

We reached a contradiction.

c. $f(x,x') =$

$$\big(x(1)x'(1)\big)^6 + e^{\left(x(3)+x(5)+x'(3)+x'(5)\right)} + \frac{1}{x(1)x'(1)} + \big(x(4)+x(6)\big)\cdot\big(x'(4)+x'(6)\big)$$

We will define a feature map $\psi(x) = \Big(x(1)^6, e^{x(3)+x(5)}, x(1)^{-1}, x(4)+x(6)\Big)$, now

$\langle\psi(x),\psi(x')\rangle = x(1)^6\cdot x'(1)^6 + e^{x(3)+x(5)}\cdot e^{x'(3)+x'(5)} + x(1)^{-1}\cdot x'(1)^{-1} +$
$\big(x(4)+x(6)\big)\cdot\big(x'(4)+x'(6)\big) =$

$\big(x(1)x'(1)\big)^6 + e^{\left(x(3)+x(5)+x'(3)+x'(5)\right)} + \frac{1}{x(1)x'(1)} + \big(x(4)+x(6)\big)\cdot\big(x'(4)+x'(6)\big) =$
$f(x,x')$
we saw $f(x,x') = \langle\psi(x),\psi(x')\rangle$ so $f(x,x')$ is a kernel function.

Question 4:

a. Based on convex function definition, $g\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ is convex if for every $u, v \in \mathbb{R}^d, a \in [0,1] : g(au+(1-a)v) \leq a * g(u) + (1-a) * g(v)$
Now, let show that g is not necessarily a convex function for any $a_1 \dots a_k \in \mathbb{R}$
Let $a_1 \dots a_k < 0$, $b \in [0,1]$ and $u, v \in \mathbb{R}^d$ then:
$g(bu+(1-b)v) = \sum_{i=1}^{k} a_i f_i(bu+(1-$
$b)v) \underset{f\ is\ convex\ and\ a\ is\ negative}{>} \sum_{i=1}^{k} a_i b f_i(u) + a_i(1-b)f_i(v))$

Explaining the last step – since f is convex we get that $f(au + (1-a)v) \leq a * f(u) + (1-a) * f(v)$ but since the coefficient $a_1 \dots a_k \in \mathbb{R}$ are all negative, the inequality mark is reversed because we get a bigger negative value, hence the condition for convexity does not hold.

b. Let $a_1 \dots a_k < 0$, $b \in [0,1]$ and $u, v \in \mathbb{R}^d$ then:
$g(bu + (1-b)v) = \sum_{i=1}^{k} a_i f_i(bu + (1-b)v) \underset{f\ is\ convex}{\leq} \sum_{i=1}^{k} a_i b f_i(u) + a_i(1-b)f_i(v)) = \sum_{i=1}^{k} a_i b f_i(u) + \sum_{i=1}^{k} a_i(1-b)f_i(v)) = bg(u) + (1-b)g(b)$

since the coefficient $a_1 \dots a_k \geq 0$ we can keep the inequality in place and the condition for convexity holds.

Question 5

a. To formulate the closed-form expression for w that minimizes the given objective function, we need to solve the optimization problem. Now, the objective function in the question is convex based because it's a sum of 2 convex functions (as we saw in the class for similar cases). To find the minimum, we need to set the gradient of the function with respect to w to 0 : $\nabla f(w) = 0$.
Now, lets define:

$$f(w) = \lambda \|w - v\|_2^2 + \sum_{i=1}^{m} (\langle w, x_i \rangle - y_i)^2$$

$$g(w) = \lambda \|w - v\|_2^2$$

$$h_i(w) = (\langle w, x_i \rangle - y_i)^2$$

We can re-write f to be:

$$f(w) = g(w) + \sum_{i=1}^{m} h_i(w)$$

Now we can calculate the gradient of each component separately:

$$\boldsymbol{\nabla g(w)} = \nabla(\lambda \|w - v\|_2^2) =$$

$$= \nabla\left(\lambda \sum_{i=1}^{k} (w_i - v_i)^2\right) = 2\lambda\left(\sum_{i=1}^{k} (w_i - v_i)\right)\nabla\left(\sum_{i=1}^{k} (w_i - v_i)\right) =$$

$$= 2\lambda\left(\sum_{i=1}^{k}(w_i - v_i)\right) = \ \boldsymbol{2\lambda(w - v)}$$

$$\boldsymbol{\nabla h_i(w)} = \nabla(\langle w, x_i\rangle - y_i)^2 = \ \boldsymbol{2(\langle w, x_i\rangle - y_i)x_i}$$

Now, lets assign the values again and solve $\nabla f(w) = 0$

$$\nabla f(w) = \ \nabla g(w) + \sum_{i=1}^{m}\nabla h_i(w) = \ \ 2\lambda(w - v) + \sum_{i=1}^{m} 2(\langle w, x_i\rangle - y_i)x_i$$

$$\nabla f(w) = 0 \longrightarrow \ 2\lambda(w - v) + \sum_{i=1}^{m} 2(\langle w, x_i\rangle - y_i)x_i = 0$$

$$\longrightarrow \ 2\lambda w - 2\lambda v + 2X^T Xw - 2X^T y = 0 \longrightarrow w(2\lambda I + 2X^T X)$$

$$= 2\lambda v + 2X^T y \ \longrightarrow \boldsymbol{w = (\lambda I + X^T X)^{-1}(\lambda v + X^T y)}$$

b. As we learned in the gradient decent algorithm

$$w^{(t+1)} = w^{(t)} - \ \eta\nabla f\left(w^{(t)}\right)$$

Hence:

$$w^{(t+1)}\left(w^{(t)}, \eta\right) = w^{(t)} - \ \eta\left(\ 2\lambda\left(w^{(t)} - v\right) + \sum_{i=1}^{m} 2\left(\langle w^{(t)}, x_i\rangle - y_i\right)x_i\right)$$

c. In stochastic gradient decent the step is performed on a random i value uniformly selected from $\{1 \dots m\}$, and the gradient of $\nabla\ell(w, (x_i, y_i))$.

converting our target function into the updated form:

$$w^{(t+1)}\left(w^{(t)}, \eta\right) = w^{(t)} - \ \eta\left(\ \nabla R\left(w^{(t)}\right) + \nabla\ell\left(w^{(t)}, (x_i, y_i)\right)\right)$$

Let's mark:

$$R(w) = \ g(w) = \lambda\|w - v\|_2^2\ ,$$

$$\ell(w, (x_i, y_i)) = \ h_i(w) = (\langle w, x_i\rangle - y_i)^2$$

and based on our calculations above we will get:

$$w^{(t+1)}\left(w^{(t)}, \eta\right) = w^{(t)} - \ \eta\left(2\lambda\left(w^{(t)} - v\right) + 2\left(\langle w^{(t)}, x_i\rangle - y_i\right)x_i\right)$$

Question 6

a. In the question it was given that PCA was performed over the dataset to reduce the dimensionality from 4 to 2, so the distortion in that case would be the lowest 2 eigenvalues of $X^T X$.

We can notice that $x_t(3) = 3\boldsymbol{x_t(1)} + \boldsymbol{x_t(2)}$ and $x_t(4) = 2x_t - 4\,x_t(3) = \boldsymbol{-2x_t(2) - 12\,x_t(1)}$

Hence both $x_t(3)$ and $x_t(4)$ are linear combination of $x_t(1), x_t(2)$ which means they are linearly dependent and the rank of $A = X^T X$ will be 2. This in turn leads to the fact that the matrix A would have 2 eigenvalues equal to 0.

Since A is positive semi-definitive matrix, its eigenvalues are all ≥ 0 hence we can conclude from the above that the 2 lowest eigenvalues of A are both 0, hence the distortion is 0.

b. First, we will choose samples: $x_1 = (0,1,1,1)$, $x_2 = (1,0,1,0)$, $x_3 = (2,0,4,4)$, $x_4 = (1,1,2,1)$ that satisfy the equations: $x_t(3) = \big(x_t(1)\big)^2 + \big(x_t(2)\big)^3$ and $x_t(4) = \big(x_t(3) - x_t(1)\big)^2$.

So $X = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 2 & 0 & 4 & 4 \\ 1 & 1 & 2 & 1 \end{pmatrix}$

Lets compute matrix $A = X^T X = \begin{pmatrix} 6 & 1 & 11 & 9 \\ 1 & 2 & 3 & 2 \\ 11 & 3 & 22 & 19 \\ 9 & 2 & 19 & 18 \end{pmatrix}$

The eigenvalues for this matrix are: $\lambda_1 \approx 44.68,\ \lambda_2 \approx 1.92, \lambda_3 \approx 1.39, \lambda_4 = 0$.

This mean that the distortion is the sum of the 2 lowest eigenvalues $\approx 1.39 > 0$, which is the distortion in (a).