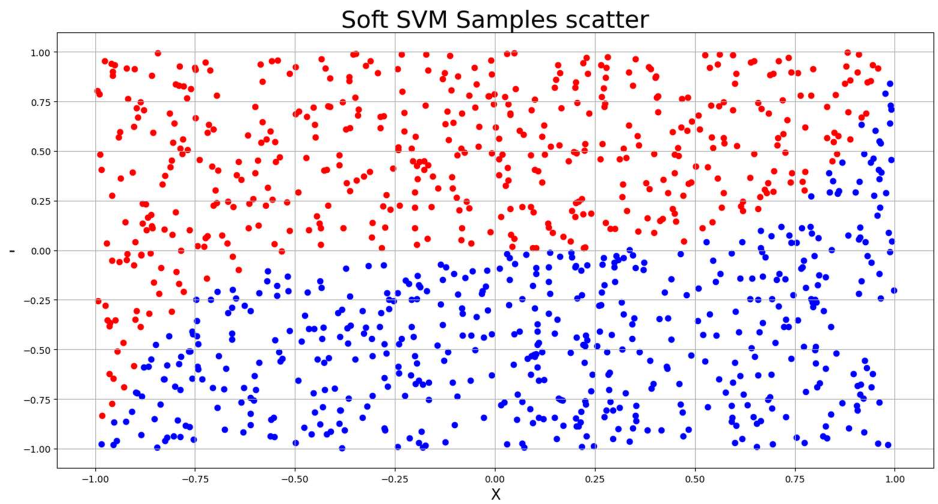
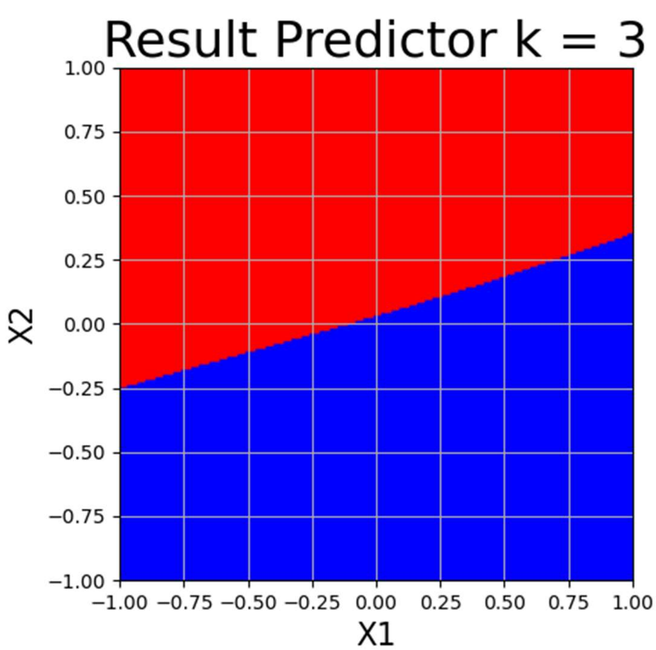
# Intro To ML – Kernels, GD and PCA

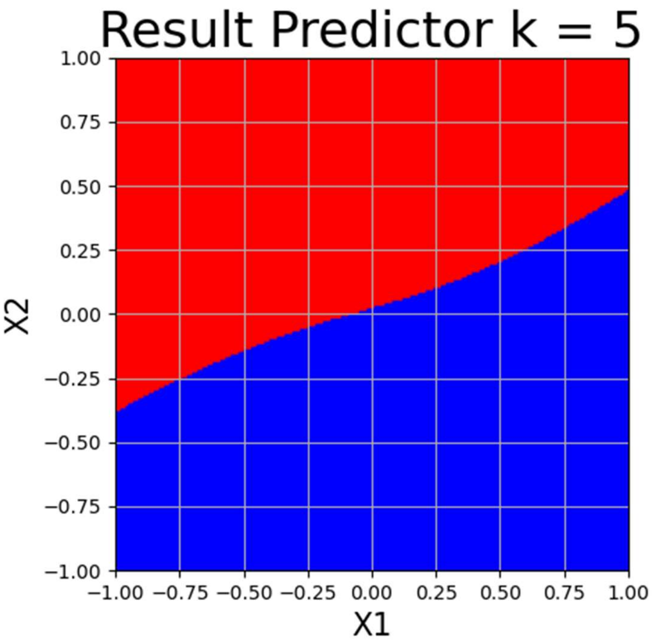
Amit Zulan – 207299033, Dvir Ben Zikri - 315409508

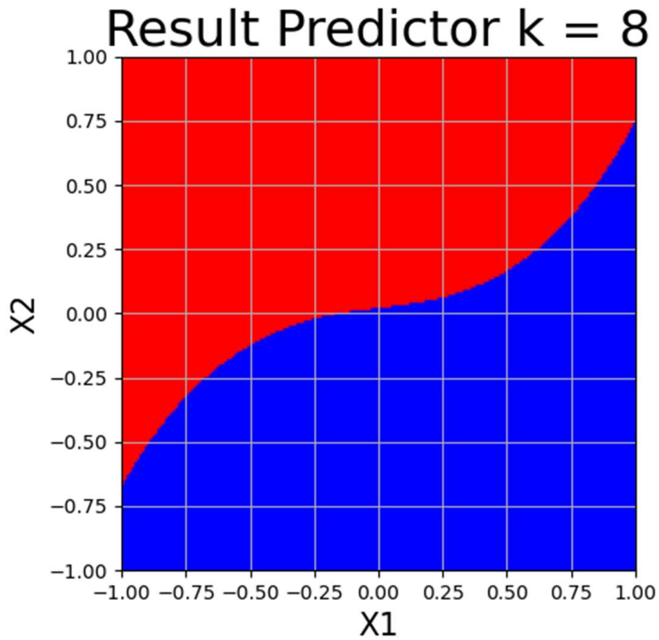
1. Code is added separately.
2. 




   a. We can see the sample set is not linearly separable therefore, it might be a good idea to increase the dimension. Moreover, it seems there would be a non-linear line from a higher dimension that would separate this sample.
   b. The resulting classifier used $\lambda = 1, k = 8$, and the resulting error is 0.01.
   c. Increasing dimensionality could make a previously non-linearly separable sample set, to be separable in higher dimension. However, increasing dimensionality could lead to overfitting and creating non-linear separators due to noise.
   d. 




Result Predictor k = 5
X1
X2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00


Result Predictor k = 8
X1
X2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00

3.

a. $K(x,x') \coloneqq \big(2x(7)+x(3)\big)\cdot x'(2)$
We need to show that this function cannot be a kernel function, to do that we will show that this function violates the commutative property of the inner product function, since a kernel function is defined as: $K\big(\psi(x),\psi(x')\big) = \langle\psi(x),\psi'(x)\rangle$.

If we switch $x$ and $x'$ in the above function, we get:

$$K(x',x) \coloneqq \big(2x'(7)+x'(3)\big)\cdot x(2) \neq \big(2x(7)+x(3)\big)\cdot x'(2) = K(x,x')$$

It can also be seen by the fact that we use different index values of the vectors x and x' in the above function.

b. $\mathrm{K(x,x')} \coloneqq 5 - (\mathrm{x}(1) - \mathrm{x}(2))(\mathrm{x}'(1) - \mathrm{x}'(2))$
We need to show that this function cannot be a kernel function.
We shall look at the case when $x = x'$:

$$\mathrm{K(x,x)} = \langle\psi(x),\psi(x)\rangle = \|\psi(x)\|_2^2 \geq 0$$
$$K(x,x) = 5 - \big(x(1)-x(2)\big)^2$$

Since $\big(x(1)-x(2)\big)^2 \geq 0$ and $\exists x\ s.t\ \big(x(1)-x(2)\big)^2 \geq 5 \rightarrow K(x,x) \leq 0$ in contradiction to definition of K as an inner product of two vector, which in this case are the same which results in $l_2$ norm.

$$i.e.\, for\ x = (80{,}60) \rightarrow \big(x(1)-x(2)\big)^2 = 20^2 \rightarrow K(x,x) = -395 < 0$$

c. $\mathrm{f(x,x')} = \big(\mathrm{x}(1)\mathrm{x}'(1)\big)^6 + \mathrm{e}^{\mathrm{x}(3)+\mathrm{x}(5)+\mathrm{x}'(3)+\mathrm{x}'(5)} + \frac{1}{\mathrm{x}(1)\mathrm{x}(1')} + \big(\mathrm{x}(4)+\mathrm{x}(6)\big)\big(\mathrm{x}'(4)+\mathrm{x}'(6)\big)$
We need to find $\psi$:

$$\mathrm{f(x,x')} = \mathrm{x}(1)^6 x'(1)^6 + e^{x(3)+x(5)}\cdot e^{x'(3)+x'(5)} + \frac{1}{\mathrm{x}(1)}\cdot\frac{1}{\mathrm{x}'(1)} + \big(\mathrm{x}(4)+\mathrm{x}(6)\big)\big(\mathrm{x}'(4)+\mathrm{x}'(6)\big)$$

Thus, we can see that $\psi$ should be:

$$\psi(x) = \left(\mathrm{x}(1)^6, e^{x(3)+x(5)}, \frac{1}{x(1)}, x(4)+x(6)\right)$$

$$K\big(\psi(x),\psi'(x)\big) = \langle\psi(x),\psi'(x)\rangle$$
$$= \langle\left(\mathrm{x}(1)^6, e^{x(3)+x(5)}, \frac{1}{x(1)}, x(4)+x(6)\right), \left(\mathrm{x}'(1)^6, e^{x'(3)+x'(5)}, \frac{1}{x'(1)}, x'(4)+x'(6)\right)\rangle =$$
$$= \mathrm{x}(1)^6 x'(1)^6 + e^{x(3)+x(5)}\cdot e^{x'(3)+x'(5)} + \frac{1}{\mathrm{x}(1)}\cdot\frac{1}{\mathrm{x}'(1)} + \big(\mathrm{x}(4)+\mathrm{x}(6)\big)\big(\mathrm{x}'(4)+\mathrm{x}'(6)\big)$$

4. Consider a linear combination of $k \in \{1,2,...\}$ convex functions $f_i : \mathbb{R}^d \rightarrow \mathbb{R}\ for\ i \in \{1,\dots,k\}$:

$$g(u) = \sum_{i=1}^{k} a_i f_i(u)$$
$$where\ a_1,\dots,a_k \in \mathbb{R}\ and\ u \in \mathbb{R}^d$$

a. $\forall i \neq 1\ a_i = 0\ and\ a_1 = -1$

For those $a_i$ we get g:

$$g(u) = -f_1(u)$$

To prove that this function is not convex we need to show that for all two vectors $u, v \in \mathbb{R}^d$

$$g(\alpha \cdot u + (1-\alpha) \cdot v) > \alpha \cdot g(u) + (1-\alpha) \cdot g(v)$$

➔

$$\begin{aligned} &g(\alpha \cdot u + (1-\alpha) \cdot v) \\ &= -f(\alpha \cdot u + (1-\alpha) \cdot v) \overset{*}{\gneqq} -\big(\alpha \cdot f(u) + (1-\alpha) \cdot f(v)\big) \\ &= \alpha \cdot \big(-f(u)\big) + (1-\alpha) \cdot \big(-f(v)\big) = \alpha \cdot g(u) + (1-\alpha) \cdot g(v) \end{aligned}$$

* Since f is convex function.

So, we got:
$g(\alpha \cdot u + (1-\alpha) \cdot v) \geq \alpha \cdot g(u) + (1-\alpha) \cdot g(v)$

Since equality is only for linear or constant functions which two are always convex, if $f_1$ is not either, we get:

$$g(\alpha \cdot u + (1-\alpha) \cdot v) > \alpha \cdot g(u) + (1-\alpha) \cdot g(v)$$

If $f_1$ is indeed a constant or linear function, we can choose a different function $f_j$ which is not a constant or linear function, and if such a function does not exist than any linear combination of linear or constant functions is convex -> g will be convex.

b. $\forall a_1, \dots, a_k \geq 0$
We need to show that for any two vectors $u, v \in \mathbb{R}^d$

$$g(\alpha \cdot u + (1-\alpha) \cdot v) \leq \alpha \cdot g(u) + (1-\alpha) \cdot g(v)$$

$$\begin{aligned} &g(\alpha \cdot u + (1-\alpha) \cdot v) \\ &= \sum_{i=1}^{k} a_i f_i(\alpha \cdot u + (1-\alpha) \cdot v) \overset{*}{\lneqq} \sum_{i=1}^{k} a_i \Big(\alpha \cdot f_i(u) + (1-\alpha) \\ &\cdot f_i(v)\Big) = \alpha \cdot \sum_{i=1}^{k} a_i \cdot f_i(u) + (1-\alpha) \cdot \sum_{i=1}^{k} a_i \cdot f_i(v) \\ &= \alpha \cdot g(u) + (1-\alpha) \cdot g(v) \end{aligned}$$

* This inequality comes from the fact that each $f_i$ is convex and that each $a_i \geq 0$ which does not change the inequality.

$$a_i \cdot f_i(\alpha \cdot u + (1-\alpha) \cdot v) \leq a_i \cdot \Big(\alpha \cdot f_i(u) + (1-\alpha) \cdot f_i(v)\Big) \; for \; a_i \geq 0$$

And for $a_i < 0 \rightarrow a_i \cdot f_i(\alpha \cdot u + (1-\alpha) \cdot v) > a_i \cdot \Big(\alpha \cdot f_i(u) + (1-\alpha) \cdot f_i(v)\Big)$

So, we got that for $\forall a_1, \dots, a_k \geq 0$ g is convex.

5. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where the input domain is $X = \mathbb{R}^d$ and the labels are from $Y = \mathbb{R}$. Consider learning of a linear predictor using the following optimization for $\lambda > 0$:

$$\underset{w \in \mathbb{R}^d}{\text{minimize}}\ \lambda \|w - v\|_2^2 + \sum_{i=1}^{m} \left(\langle w, x_i \rangle - y_i\right)^2$$

Where $v \in \mathbb{R}^d$ *is a given vector*.

a. First, we'll write this in matrix form.

$$\underset{w \in \mathbb{R}^d}{\text{minimize}}\ \lambda \|w - v\|_2^2 + \|X^T w - y\|_2^2$$

Where $X = [x_1, \ldots, x_m]$ $d \times m$ *matrix*, $y = [y_1, \ldots, y_m]^T \in \mathbb{R}^m$ *is a column vector*.

To find the w that minimizes the above formula we shall find the formula's gradient with respect to w.

$$\nabla_w f(w) = 2\lambda(w - v) + 2X(X^T w - y)$$

The gradient is 0 when:

$$(XX^T + \lambda I)w = Xy + \lambda v$$

Since $\lambda > 0$ $XX^T + \lambda I$ is invertible:

$$w = (XX^T + \lambda I)^{-1}(Xy + \lambda v)$$

b. To calculate the step of the GD algorithm we need to find the gradient of $f(w)$ Which we did above so:

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \cdot 2 \cdot \left(\lambda \cdot \left(w^{(t)} - v\right) + X\left(X^T w^{(t)} - y\right)\right)$$

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \cdot 2 \cdot \left((XX^T + \lambda I) \cdot w^{(t)} - Xy - \lambda v\right)$$

$$w^{(t+1)} \leftarrow \left(I - 2 \cdot \eta \cdot (XX^T + \lambda I)\right)w^{(t)} + 2 \cdot \eta \cdot (Xy + \lambda v)$$

c. To calculate the step of the SGD we need to separate $f(w)$ to $R(w)$ *and* $l\left(w, (x, y)\right)$:

$$R(w) = \lambda \|w - v\|_2^2,\ \ l\left(w, (x, y)\right) = \frac{1}{m} \sum_{i=1}^{m} m\left(\langle w, x_i \rangle - y_i\right)^2$$

The step of SGD is the gradient of $R(w)$ and $l\left(w, (x_i, y_i)\right)$ for a uniformly randomly selected $i$. The gradients are:

$$\nabla_{\mathrm{w}} R(w) = 2\lambda(w - v),\ \ \nabla_{\mathrm{w}} l\left(w, (x_i, y_i)\right) = 2m x_i\left(\langle w, x_i \rangle - y_i\right)$$

So, the SGD step is:

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

6.

a. In the 1st experiment it turned out that in all times t, $x_t(3) = 3x_t(1) + x_t(2), x_t(4) = 2x_t(2) - 4x_t(3)$
And thus $x_t(4) = 2x_t(2) - 12x_t(1) - 4x_t(2) = -12x_t(1) - 2x_t(2)$
We got that $\{x_t(1), x_t(2), x_t(3)\}$ and $\{x_t(1), x_t(2), x_t(4)\}$ are linearly dependent. All in all, we get that $rank(X) \leq 2, A = XX^T \rightarrow rank(A) \leq 2$.
Since $d = 4$ at least two of A's eigenvalues are equal to 0.
Since $k = 2\ and\ d - k = 2$ the best distortion is equal to the 2 lowest eigenvalues thus $best\_distortion = 0 + 0 = 0$.

b. In another experiment it turned out that in all times t,
$$x_t(3) = x_t^2(1) + x_t^3(2), \quad x_t(4) = \left(x_t(3) - x_t(1)\right)^2$$
Since now the $\{x_t(1), x_t(2), x_t(3), x_t(4)\}$ are not necessarily dependent, if we would pick at least $m \geq 3$, then for at least 3 linearly independent samples we would get at least 3 eigenvalues different than 0.
Since A is positive semi definite all 3 will be positive thus the best distortion must be larger than 0 and thus larger than section a.

For example, Let:
$$x_1 = \begin{pmatrix} -1 \\ 2 \\ 9 \\ 100 \end{pmatrix}, x_2 = \begin{pmatrix} 3 \\ 1 \\ 10 \\ 49 \end{pmatrix}, x_3 = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 1 \end{pmatrix}$$
$$X = \begin{pmatrix} -1 & 3 & 1 \\ 2 & 1 & 1 \\ 9 & 10 & 2 \\ 100 & 49 & 1 \end{pmatrix}, X^T = \begin{pmatrix} -1 & 2 & 9 & 100 \\ 3 & 1 & 10 & 49 \\ 1 & 1 & 2 & 1 \end{pmatrix}$$
$$A = XX^T = \begin{pmatrix} 11 & 2 & 24 & 48 \\ 2 & 6 & 28 & 250 \\ 24 & 28 & 168 & 1292 \\ 48 & 250 & 1292 & 12402 \end{pmatrix}$$
The eigenvalues of A are:
$$\lambda_1 \cong 12542, \lambda_2 \cong 44, \lambda_3 \cong 0.8, \lambda_4 = 0$$
In this case the best distortion is 0.8.