1.1
Nearest Neighbor Classifier is using labeled data as reference to classify a data without label by choosing the nearest labeled cluster. K-means clustering is a process to cluster unlabeled datas into cluster by iterations of choosing random centroid as a clustering reference. So Nearest Neighbor Classifier can be viewed as K means clustering after choosing centroids to cluster to. Yes, finding a cluster from the existing ones for a new data point in the test set K-Means clustering can be viewed as Nearest Neighbor Classifier because the K means centroids have been chosen.

1.2
Yes it is possible. We can at first preprocess the data by taking the natural logarithm of all y, so the N datapoint will become N data points $\{(x_{1\text{-}n}, \ln(y_{1\text{-}n})\}$, then we can use LinearRegression model from sklearn to fit this processed data. If the score of the model (which models the loss) is within a reasonable range, then we can this data set has a exponential relationship.

2.

① $L = \frac{1}{n}\sum_{i}^{n}[y^i - (\theta_0 + \theta_1 x^i)]^2 + \beta(e^{\theta_0} + e^{\theta_1})$

② $\frac{\partial L}{\partial \theta_0} = -\frac{2}{n}\sum_{i}^{n}(y^i - \theta_0 - \theta_1 x^i) + \beta e^{\theta_0}$

③ $\frac{\partial L}{\partial \theta_1} = -\frac{2}{n}\sum_{i}^{n} x_i (y^i - \theta_0 - \theta_1 x^i) + \beta e^{\theta_1}$

④ $\theta_0 \leftarrow \theta_0 - \alpha \frac{\partial L}{\partial \theta_0}$

⑤ I think it won't work well because if β is large, which means a big regularization panalty, for $\beta e^{\theta_0} / \beta e^{\theta_1}$ to minimize, $\theta_0 / \theta_1 \to -\infty$, which against the first purpose of trying to reduce θ_0 / θ_1 to 0 to simplify the model.

3.

(1) The 3 neareast neighbours are

x^7 : Euclidean distance 1

x^4, x^5, Euclidean distance $\sqrt{2}$,

Label should be x_4 x_5 's label, $y=+1$

(2) The 4th and 5th neareast neighbour is x^2, x^6, distance = 2

Label should be $y=-1$

(3) The five nearest points are:

x^7, $d=1$

x_4 $d=2$

x_5 $d=2$

x_2 $d=4$

x_6 $d=4$, label should be $y=-1$

4.1 $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ $A^T = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} -1 & -2 \\ 2 & 1 \end{bmatrix} != \begin{bmatrix} -2 & 1 \\ -1 & 2 \end{bmatrix}$$

4.2 $(||X||_2)^2 = X^T X$

$$X^T X \; X^T X \; X^T X \; X^T X = 2 \times 2 \times 2 \times 2 = 16$$

5 (10 points) Clustering

Assume there is a dataset $S = \{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}\}$ where:

$$x^{(1)} = -8, \quad x^{(2)} = -10, \quad x^{(3)} = 0,$$
$$x^{(4)} = 11, \quad x^{(5)} = 12.$$

1. If we have initial centroids $c_1 = 0$, $c_2 = 11$, $c_3 = 12$ and perform ***K*-means** clustering algorithm on dataset S using Euclidean distance. Fill in the table until the algorithm converges. To show the algorithm has converged, it is necessary to fill in one more iteration that duplicates the previous one. You may not use all the rows. Some example entries are filled. For each of the iteration, find the positions of the centroids, assign the data points to the centroids, and compute the loss.

	centroid positions			data points assignment					
iter	c_1	c_2	c_3	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	SSE
1	0	11	12	c_1	c_1	c_1	c_2	c_3	164
2	−6	11	12	c_1	c_1	c_1	c_2	c_3	56
3	−6	11	12	c_1	c_1	c_1	c_2	c_3	56
4									
5									

5.2

Centroids positions : $-9, 0, 11.5$

SSE: $1+1+0.25+0.25 = 2.5$

5.3

As mentioned above, $-9, 0, 11.5$

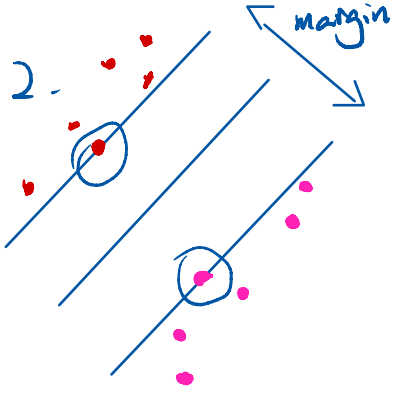
(6)

1.

$$\frac{\partial f}{\partial w} \begin{cases} 2\lambda w, & 1 - y_i(w^T x_i - b) \le 0 \\ 2\lambda w + \frac{1}{m}\sum_{i}^{m} -y_i x_i, & 1 - y_i(w^T x_i - b) > 0 \end{cases}$$

note, only sum the $-y_i x_i$ if that x_i, y_i satisfies $1 - y_i(w^T x_i - b) > 0$, in total there is m points satisfied the condition, $0 \le m \le n$

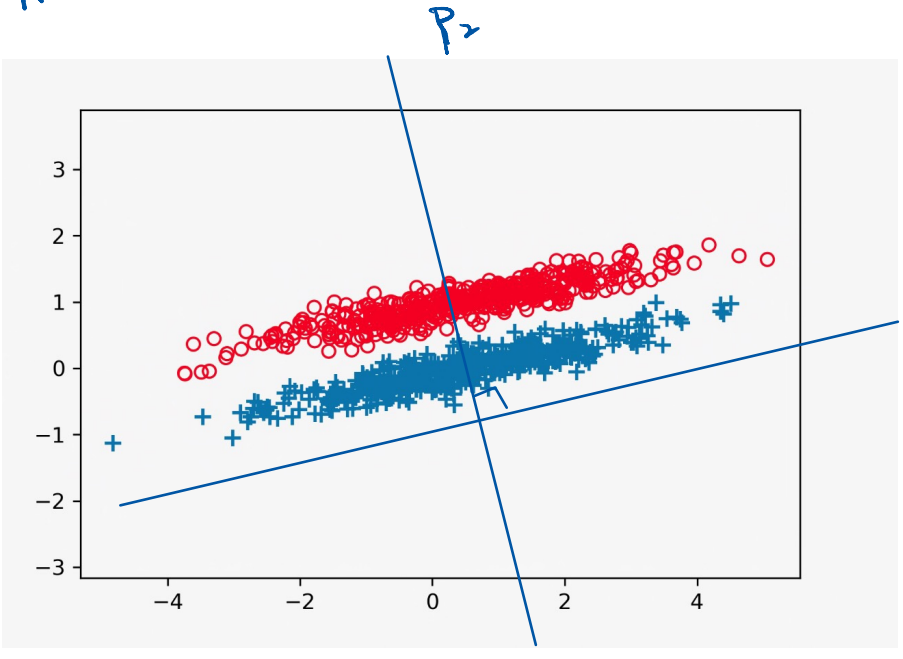
2.



For a binary-label dataset that has minimized support vectors, the minimized vector is 2, as showed in figure

7.

(1)



(2) P2

8.

if $\hat{y} - y = \underbrace{[1, 1, 1 \ldots\ldots 1]}_{n}$

that is, each prediction $\hat{y}_i$ is 1 more than $\hat{y}$.

then:

$$MAE = \frac{1}{n} n \times |1| = 1$$

$$RMSE = \sqrt{\frac{1}{n} \times n (1)^2} = 1$$

$$ME = \max |1| = 1$$

9. ① $k(x,y) = (x \cdot y + 2)^2$

$\Phi(x_i) = \langle x_1^2, x_2^2 \cdots x_d^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3 \cdots \sqrt{2}x_1x_d, \sqrt{2}x_2x_3$
$\sqrt{2}x_2x_d \cdots \sqrt{2}x_{d-1}d, 2x_1, \cdots 2x_d, 2 \rangle$

$\Phi(y_i) = \langle y_1^2, y_2^2 \cdots y_d^2, \sqrt{2}y_1y_2, \sqrt{2}y_1y_3 \cdots \sqrt{2}y_1y_d,$
$\cdots \sqrt{2}y_{d-1}d, 2y_1 \cdots 2y_d, 2 \rangle$

② $k(x,y) = (xy-1)^3$

$d=2 \quad x=[1,0] \quad y=[0,1]$

$$K = \begin{bmatrix} k(x,x), & k(x,y) \\ k(y,x), & k(y,y) \end{bmatrix} = \begin{bmatrix} 0, & -1 \\ -1, & 0 \end{bmatrix}$$

$a=[1,1] \quad aka^T = [1,1] \begin{bmatrix} 0,-1 \\ -1,0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = -2 < 0$

K is not positive semidefinite matrix, so not a valid kernal function