# COMP0078

# Supervised Learning

Coursework 1

Student Number: 22108699

Student Number: 22197365

November 16th, 2022

# Contents

# 1 Part I: Regression

## 1.1 Question 1

### 1.1.1 1a

Figure 1 shows the fitted curves.

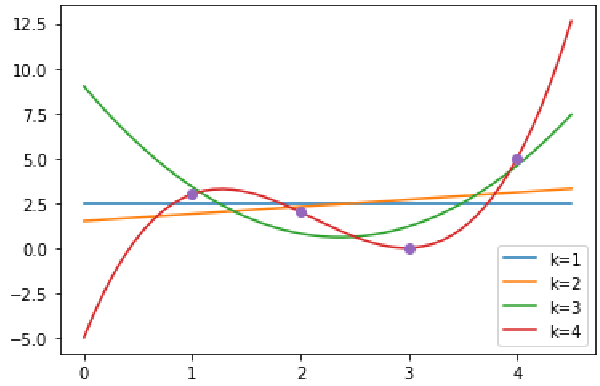


Figure 1: Fitted curves

### 1.1.2 1b

The equation corresponding to $k = 1$ is 2.5

The equation corresponding to $k = 2$ is $1.5 + 0.4x$

The equation corresponding to $k = 3$ is $9 - 7.1x + 1.5x^2$

The equation corresponding to $k = 4$ is $-5 + 15.17x - 8.5^2 + 1.33x^3$

### 1.1.3 1c

MSE of k = 1 is 3.25

MSE of k = 2 is 3.05

MSE of k = 3 is 0.80

MSE of k = 4 is $3.97 * 10^{-26}$

## 1.2 Question 2

### 1.2.1 2a

Figure 2 shows the curve $sin^2(2x),\ 0 < x < 1$ and random data generated with noise.

Figure!3 shows the fitted curves by the polynomial of dimension $k = 2, 5, 10, 14, 18$.

### 1.2.2 2b

Figure 4 shows the $ln(te_k(S))$ with the polynomial of dimension $k = 1, ..., 18$.

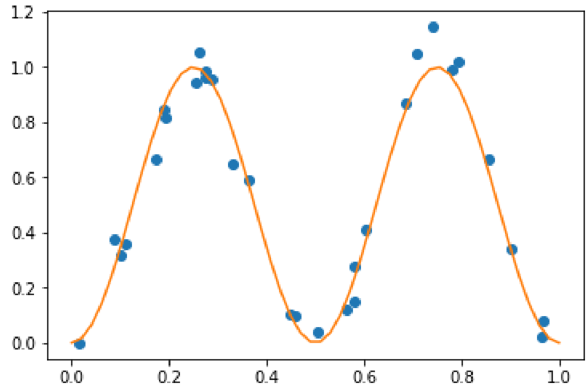


Figure 2: $sin^2(2x)$ and random data points

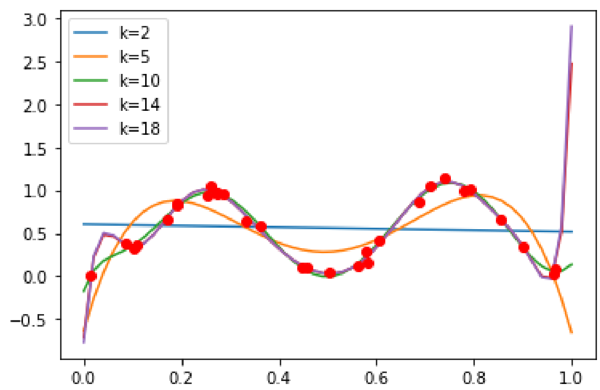


Figure 3: $sin^2(2x)$ and random data points

#### 1.2.3 2c

Figure 5 shows the $ln(tse_k(S,T))$ versus the polynomial dimension $k$. Due to the overfitting, the $ln(tse_k(S,T))$ starts to increase when $k > 10$.

#### 1.2.4 2d

Figure 6 shows the average MSE of the training/testing dataset in 100 runs for each k.

### 1.3 3

Same assumptions about data generation but different basis

$$\{\sin(1\pi x), \sin(2\pi x), \sin(3\pi x), \ldots, \sin(k\pi x)\} \text{ (for } k = 1, \ldots, 18 \text{ )}$$

#### 1.3.1 3b

Figure 7 shows $ln(te_k(S))$ versus the polynomial dimension $k$.

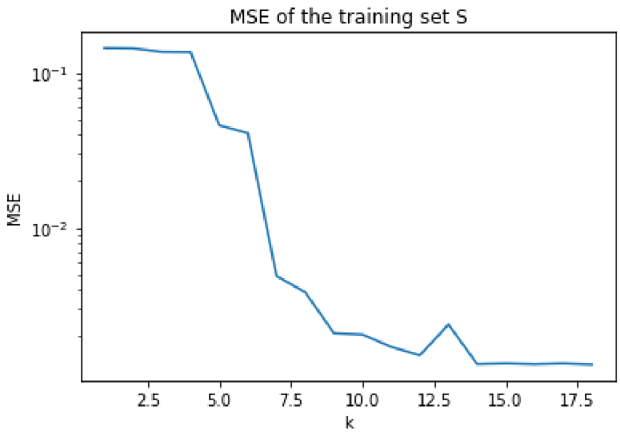


Figure 4: $ln(te_k(S))$ versus the polynomial dimension $k$

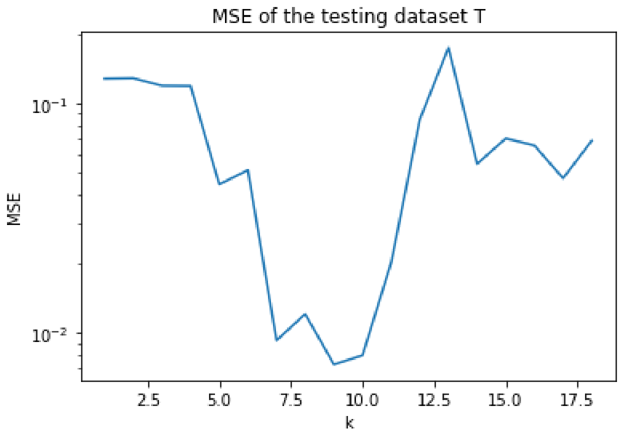


Figure 5: $ln(tse_k(S,T))$ versus the polynomial dimension $k$

#### 1.3.2 3c

Figure 8 shows $ln(te_k(S))$ versus the polynomial dimension $k$.

#### 1.3.3 3d

Figure 9 shows $ln(te_k(S))$ versus the polynomial dimension $k$.

### 1.4 Question 4

#### 1.4.1 4a

average MSE on the training dataset is 84.81

average MSE on the testing dataset is 83.88

#### 1.4.2 4b

the constant function $f = b$, $b$ is the average of the column "MEDV" in the training data.

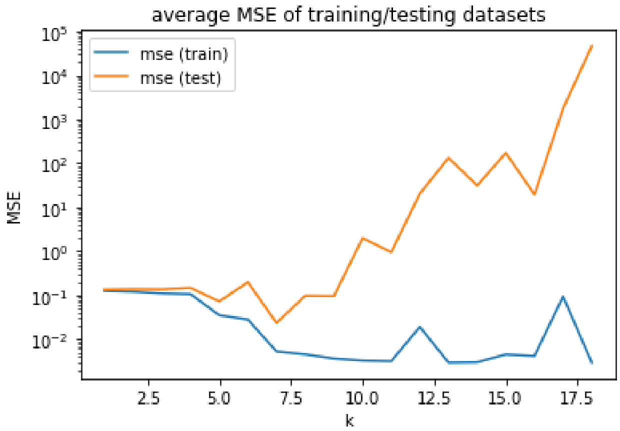


Figure 6: average $ln(tse_k(S,T))$ versus the polynomial dimension $k$ in 100 runs

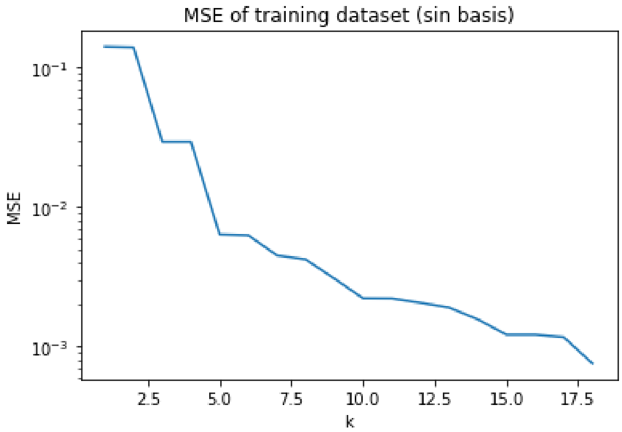


Figure 7: $ln(te_k(S))$ versus the polynomial dimension $k$

### 1.4.3 4c

MSE of the training data for each attribute is

'CRIM': 70.75276745986666,

' ZN ': 73.52147656913385,

'INDUS ': 65.05844672735891,

'CHAS': 80.86247683096228,

'NOX': 69.58952147626793,

'RM': 42.83782680571877,

'AGE': 73.6715507255629,

'DIS': 81.63960934803552,

'RAD': 70.71004870796806,

'TAX': 66.49290329447254,

'PTRATIO': 61.825393429721586,

'LSTAT': 37.90032741236634

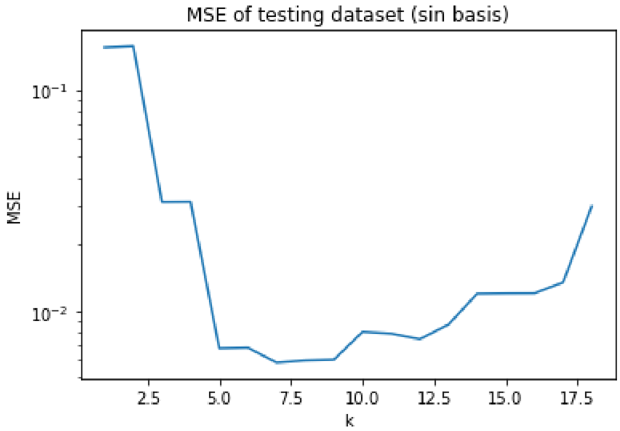


Figure 8: $ln(tse_k(S,T))$ versus the polynomial dimension $k$

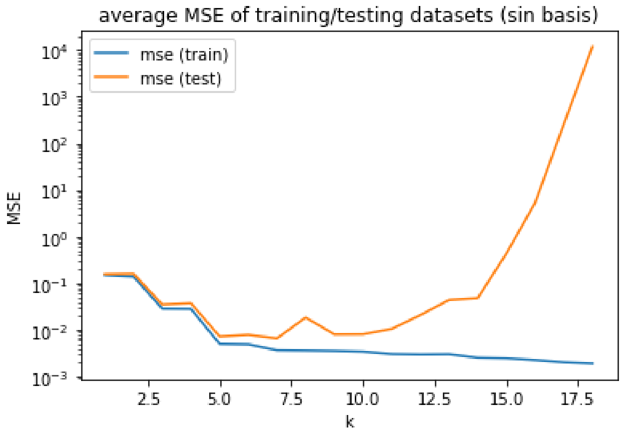


Figure 9: average $ln(tse_k(S,T))$ versus the polynomial dimension $k$ in 100 runs

MSE of the testing data for each attribute is

'CRIM': 74.37228480113156,

' ZN ': 73.75857512716416,

'INDUS ': 64.53283018377144,

'CHAS': 84.23612457474943,

'NOX': 68.21252910202332,

'RM': 45.518639076003026,

'AGE': 70.38497960286736,

'DIS': 74.6391037972427,

'RAD': 75.39307851180165,

'TAX': 65.21728225351555,

'PTRATIO': 64.68264104968723,

'LSTAT': 39.914718537251744

#### 1.4.4 4d

MSE of the training data for all attributes is 22.757

MSE of the testing data for all attributes is 23.424

## 1.5 Question 5

### 1.5.1 5a

$(\gamma, \sigma)_{best} = (2^{-35}, 2^{10.5})$ with the 5-folds cross-validation error=12.62.

### 1.5.2 5b

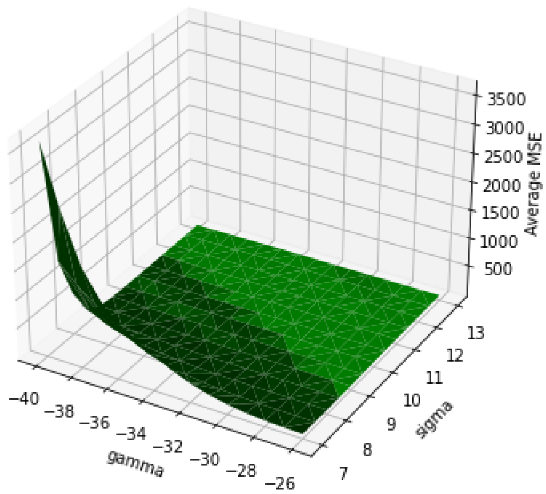


Figure 10: 5-folds cross-validation error with the order of $(\gamma, \sigma)$

### 1.5.3 5c

testing error is 16.42 and training error is 6.84 with$(\gamma, \sigma)_{best} = (2^{-35}, 2^{10.5})$

### 1.5.4 5d

| Method | MSE train | MSE test |
|---|---|---|
| Naive Regression | $83.87 \pm \sigma' 4.704$ | $85.74 \pm \sigma' 9.236$ |
| LR (attribute 1) | $70.95 \pm \sigma' 4.083$ | $74.26 \pm \sigma' 8.130$ |
| LR (attribute 2) | $73.21 \pm \sigma' 4.649$ | $74.26 \pm \sigma' 9.137$ |
| LR (attribute 3) | $64.41 \pm \sigma' 4.589$ | $65.52 \pm \sigma' 9.107$ |
| LR (attribute 4) | $81.44 \pm \sigma' 4.356$ | $83.10 \pm \sigma' 8.630$ |
| LR (attribute 5) | $68.42 \pm \sigma' 4.577$ | $70.47 \pm \sigma' 9.059$ |
| LR (attribute 6) | $42.70 \pm \sigma' 3.567$ | $45.76 \pm \sigma' 7.254$ |
| LR (attribute 7) | $71.92 \pm \sigma' 4.934$ | $73.76 \pm \sigma' 9.736$ |
| LR (attribute 8) | $78.57 \pm \sigma' 5.099$ | $80.68 \pm \sigma' 10.046$ |
| LR (attribute 9) | $71.64 \pm \sigma' 4.426$ | $73.48 \pm \sigma' 8.732$ |
| LR (attribute 10) | $65.38 \pm \sigma' 4.421$ | $67.23 \pm \sigma' 8.731$ |
| LR (attribute 11) | $63.23 \pm \sigma' 4.076$ | $61.91 \pm \sigma' 8.025$ |
| LR (attribute 12) | $37.92 \pm \sigma' 2.348$ | $39.95 \pm \sigma' 4.690$ |
| LR (attribute all) | $22.00 \pm \sigma' 1.655$ | $24.36 \pm \sigma' 3.550$ |
| Kernel Ridge Regression | $13.50 \pm \sigma' 5.009$ | $15.90 \pm \sigma' 6.230$ |

# 2 KNN

## 2.1 6

Figure 11 shows the decision region of a hypothesis $h_{S,v}$ visualized with $|S| = 100$ and $v = 3$.

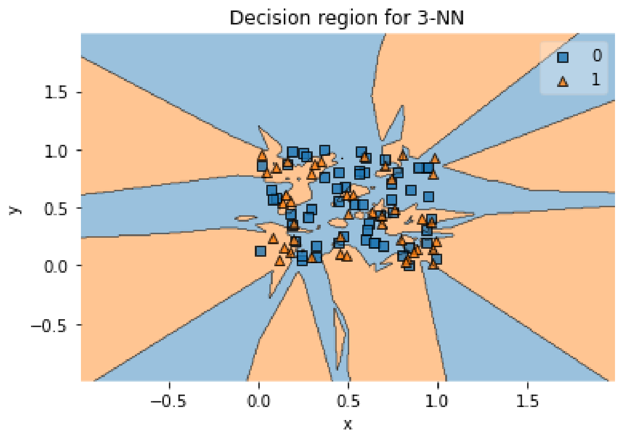


Figure 11: A hypothesis $h_{S,v}$ visualized with $|S| = 100$ and $v = 3$.

## 2.2 7

Figure 12 shows the generalization error of $k$-NN as a function of $k$. Explanation:

1. The figure is expected as a U-shape curve and our simulation is closed to the expectation.
2. The error first decreases as k grows (in this case, $k \in (0, 10)$) due to under-fitting. As $k$ is big enough, the error increases as k grows due to over-fitting. the optimal k is around 10.
3. The error attaches its maximum when $k = 1$, this is because at that time, we predict the label uniformly. (as it seek the closed point and we generated our training points with its label uniformly.)
4. The error is up and down as k grows one by one.

## 2.3 8

Figure 13 shows the optimal $k$ for a group of m during 100 runs. Explanation:

1. The figure is expected as an increasing curve and our simulation is closed to the expectation.
2. As the number of training points increasing, we need more points (information) to locate the test points at the hypothesis space and thus the optimal k is getting bigger.

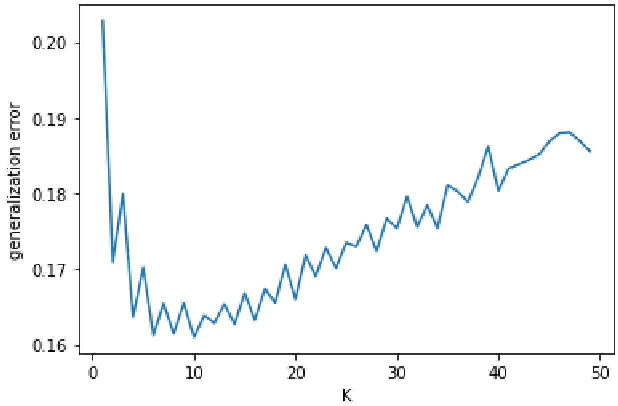


Figure 12: generalization error of $k$-NN as a function of $k$.

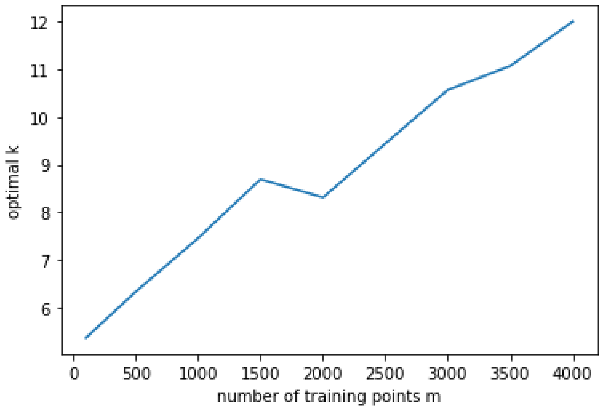


Figure 13: the optimal $k$ for a group of m during 100 runs.

# 3 Part III: Kernel with Regression

## 3.1 Question 9

### 3.1.1 9a

The given function is $K_c(x, z) := c + \sum_{i=1}^{n} x_i z_i$. If the $K_c$ is a positive semidefinite kernel, it is symmetric and the matrix $(K(x_i, x_j) : i, j = 1, \ldots, k)$ is positive semidefinite for every $k \in \mathbb{N}$ and every $x_1, \ldots, x_k \in \mathbb{R}^n$.

If the matrix $K(x,x)$ is a positive semidefinite matrix, then we have

$$\begin{aligned}
\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j) &\geq 0 \\
\sum_{i,j=1}^{m} c_i c_j (c + \sum_{k=1}^{n} x_{ki} x_{kj}) &\geq 0 \\
\sum_{i,j=1}^{m} c_i c_j c &\geq - \sum_{i,j=1}^{m} c_i c_j \sum_{k=1}^{n} x_{ki} x_{kj} \\
(\sum_{i=1}^{m} c_i)^2 c &\geq - \sum_{i,j=1}^{m} \sum_{k=1}^{n} c_i c_j x_{ki} x_{kj} \\
(\sum_{i=1}^{m} c_i)^2 c &\geq - \sum_{i,j=1}^{m} \sum_{k=1}^{n} (c_i x_{ki})(c_j x_{kj}) \\
(\sum_{i=1}^{m} c_i)^2 c &\geq -(\sum_{i}^{m} \sum_{k=1}^{n} c_i x_{ki})^2 \\
c &\geq - \frac{(\sum_i^m \sum_{k=1}^n c_i x_{ki})^2}{(\sum_{i=1}^m c_i)^2},
\end{aligned} \tag{1}$$

where $m \in \mathbb{N}, c_i, c_j \in \mathbb{R}, i, j = 1, \ldots, m$.

Since the RHS is only with two values of square, the maximum of RHS is 0. Therefore the condition that $K(x,x)$ is a positive semidefinite matrix is $c \geq 0$.

If $K_c$ is symmetric matrix, we can rewrite $K_c(x,z)$ as

$$\begin{aligned}
K_c(x,z) &= c + \sum_{i=1}^{n} x_i z_i \\
&= (\sqrt{c}, x_1, x_2, \ldots, x_i)(\sqrt{c}, z_1, z_2, \ldots, z_i)^T \\
&= \phi(x)^T \phi(z),
\end{aligned} \tag{2}$$

where $\phi : \mathbb{R}^n \to \mathscr{W}$ and Hilbert space $\mathscr{W}$ and for all $c \geq 0$.

Hence, $K_c$ is a positive semidefinite kernel if $c \geq 0$.

### 3.1.2 9b

By Representer Theorem, we have $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}_i, \mathbf{x})$, when consider the linear regression with kernel under square error lost, we have the error

$$\begin{aligned}
\varepsilon &= \sum_{i=1}^{m} \left( \sum_{j=1}^{m} \alpha_j K_c(\mathbf{x}_i, \mathbf{x}) - y_i \right)^2 \\
&= \sum_{i=1}^{m} \left( \sum_{j=1}^{m} \alpha_j c + \sum_{j=1}^{m} \alpha_j \sum_{i=1}^{n} (x_i)^2 - y_i \right)^2 \\
&= \sum_{i=1}^{m} \left( \sum_{j=1}^{m} \alpha_j c + \sum_{i=1}^{n} (\sum_{j=1}^{m} \alpha_j x_i^2 - y_i) \right)^2 \\
&= \sum_{i=1}^{m} (\sum_{j=1}^{m} \alpha_j c)^2 + \sum_{i=1}^{n} (\sum_{j=1}^{m} \alpha_j x_i^2 - y_i)^2 + 2c \sum_{i=1}^{m} (\sum_{j=1}^{m} \alpha_j \sum_{i=1}^{n} (\sum_{j=1}^{m} \alpha_j x_i^2 - y_i))
\end{aligned}$$

Due to $\sum_{i=1}^{n}(\sum_{j=1}^{m}\alpha_j x_i^2 - y_i)^2$ is the error for linear regression under square error lost, the rest terms $f(c,\alpha) = \sum_{i=1}^{m}(\sum_{j=1}^{m}\alpha_j c)^2 + 2c\sum_{i=1}^{m}(\sum_{j=1}^{m}\alpha_j \sum_{i=1}^{n}(\sum_{j=1}^{m}\alpha_j x_i^2 - y_i))$ can be regarded as regularization terms. Hence, $\alpha$ is inverse proportion to $c$ and if $c$ is big enough, $\alpha$ will tend to zero.

## 3.2 Question 10

If we want our trained linear classifier to simulate a 1-NN on the same dataset, that means, for the test point $\mathbf{t}$, $f(\mathbf{t})$ is determined by one point $(\mathbf{x}_n, y_n)$ only, where $(\mathbf{x}_n, y) \in \Re^n \times \{-1, 1\}$ and

$$\|\mathbf{x}_n - \mathbf{t}\|^2 \leq \|\mathbf{x}_i - \mathbf{t}\|^2 \text{ for } i = 1, ..., m, i \neq n$$

For a fixed $\beta$, we have

$$\exp\left(-\beta\|\mathbf{x}_n - \mathbf{t}\|^2\right) \geq \exp\left(-\beta\|\mathbf{x}_i - \mathbf{t}\|^2\right) \text{ for } i = 1, ..., m, i \neq n$$

For a fixed pair $(\boldsymbol{x}, \mathbf{t})$, the function $K_\beta$ is a decreasing function on $\beta \in \Re$. $K_{\beta=0} = 1$ and

$$\lim_{\beta \to \infty} K_\beta = 0$$

Hence,

$$\forall \epsilon > 0, \exists \beta \in [0, \infty], \text{s.t. } \exp\left(-\beta\|\mathbf{x}_{i\neq n} - \mathbf{t}\|^2\right) < \epsilon, \exp\left(-\beta\|\mathbf{x}_n - \mathbf{t}\|^2\right) = c,\ c > 0$$

This is, there exist a $\beta$ such that all kernel $K_\beta(\boldsymbol{x}, \mathbf{t}) = \exp\left(-\beta\|\boldsymbol{x} - \mathbf{t}\|^2\right)$ tends to zero except $\boldsymbol{x} = \boldsymbol{x}_n$.
Let $\beta$ satisfied the above condition, consider $f(\mathbf{t}) = \sum_{i=1}^{m}\alpha_i K_\beta(\boldsymbol{x}_i, \mathbf{t})$ and classifier of kernel, we can rewrite into

$$\begin{aligned}
\operatorname{sign}(f(\mathbf{t})) &= \operatorname{sign}(\sum_{i=1}^{m}\alpha_i K_\beta(\mathbf{x}_i, \mathbf{t})) \\
&= \operatorname{sign}(\alpha_n K_\beta(\mathbf{x}_n, \mathbf{t})) \\
&= \operatorname{sign}(c\alpha_n)
\end{aligned} \tag{3}$$

By linear regression model, we have $\mathbf{y} = \mathbf{K}\alpha^*$ That is,

$$y_n = \sum_{i=1}^{m}\alpha_i K_\beta(\mathbf{x}_i, \mathbf{x}_n) = a_n$$

Hence, we have

$$\operatorname{sign}(f(\mathbf{t})) = \operatorname{sign}(cy_n),\ c > 0$$

where the corresponding training point $(\mathbf{x}_n, y_n)$ satisfied $\|\mathbf{x}_n - \mathbf{t}\|^2 \leq \|\mathbf{x}_i - \mathbf{t}\|^2$ for $i = 1, ..., m, i \neq n$
Proved.