MLT: Week 5

Linear Regression: Least Squares and Kernel Regression

Sherry Thomas

Linear Regression Algorithm

Linear Regression Algorithm

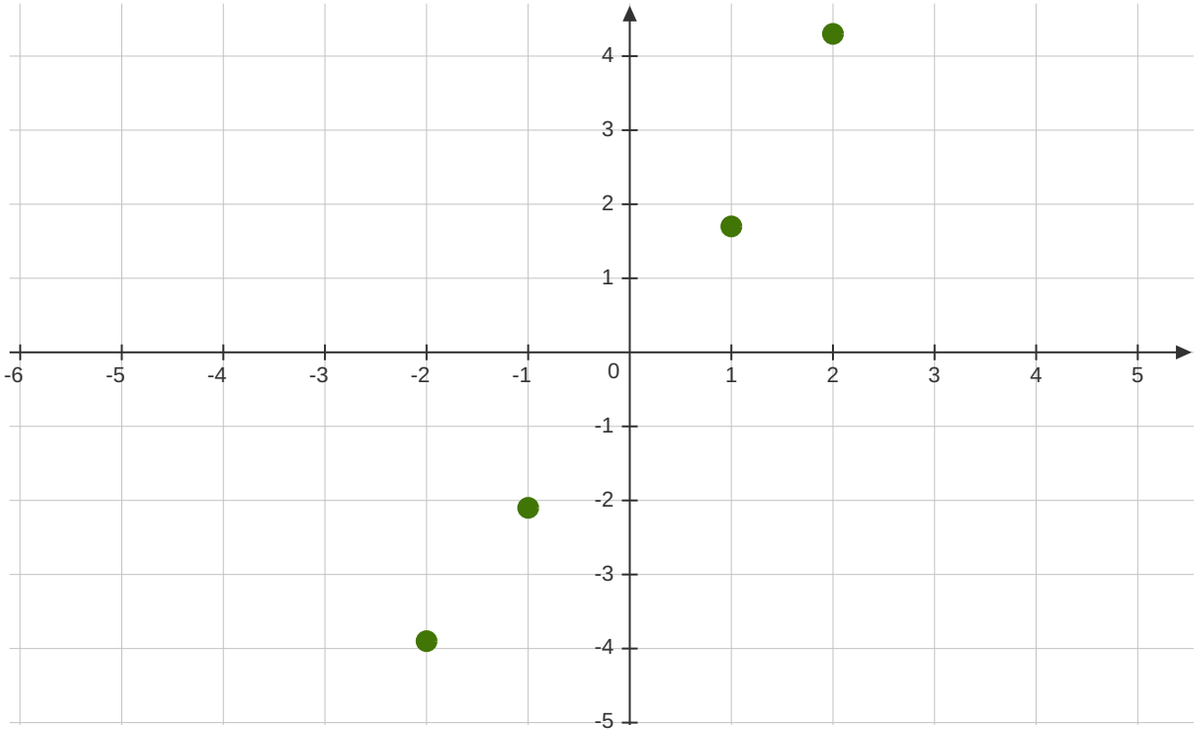
Given a dataset $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ where $\mathrm{x}_i \in \mathbb{R}^d$,
let $\{y_1, \ldots, y_n\}$ be the labels, where $y_i \in \mathbb{R}$.

$$\mathrm{X} = \begin{bmatrix} -2 & -1 & 1 & 2 \end{bmatrix} \quad \mathrm{y} = \begin{bmatrix} -3.9 \\ -2.1 \\ 1.7 \\ 4.3 \end{bmatrix}$$

Linear Regression Algorithm

Given a dataset $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ where $\mathrm{x}_i \in \mathbb{R}^d$, let $\{y_1, \ldots, y_n\}$ be the labels, where $y_i \in \mathbb{R}$.

$$\mathrm{X} = \begin{bmatrix} -2 & -1 & 1 & 2 \end{bmatrix} \quad \mathrm{y} = \begin{bmatrix} -3.9 \\ -2.1 \\ 1.7 \\ 4.3 \end{bmatrix}$$



Linear Regression Loss Function

Linear Regression Loss Function

The most commonly used loss function for Regression is the Squared Sum Error. It is given by,

$$\sum_{i=1}^{n} ||\mathrm{w}^T \mathrm{x}_i - y_i||_2^2 \qquad \forall i$$

Linear Regression Loss Function

The most commonly used loss function for Regression is the Squared Sum Error. It is given by,

$$\sum_{i=1}^{n} ||\mathrm{w}^T \mathrm{x}_i - y_i||_2^2 \qquad \forall i$$

We use the line $y = x$ as a baseline to better understand the loss function.
In the above line, w is represented as $\mathrm{w} = [1]$. Using this in the loss function, we obtain,

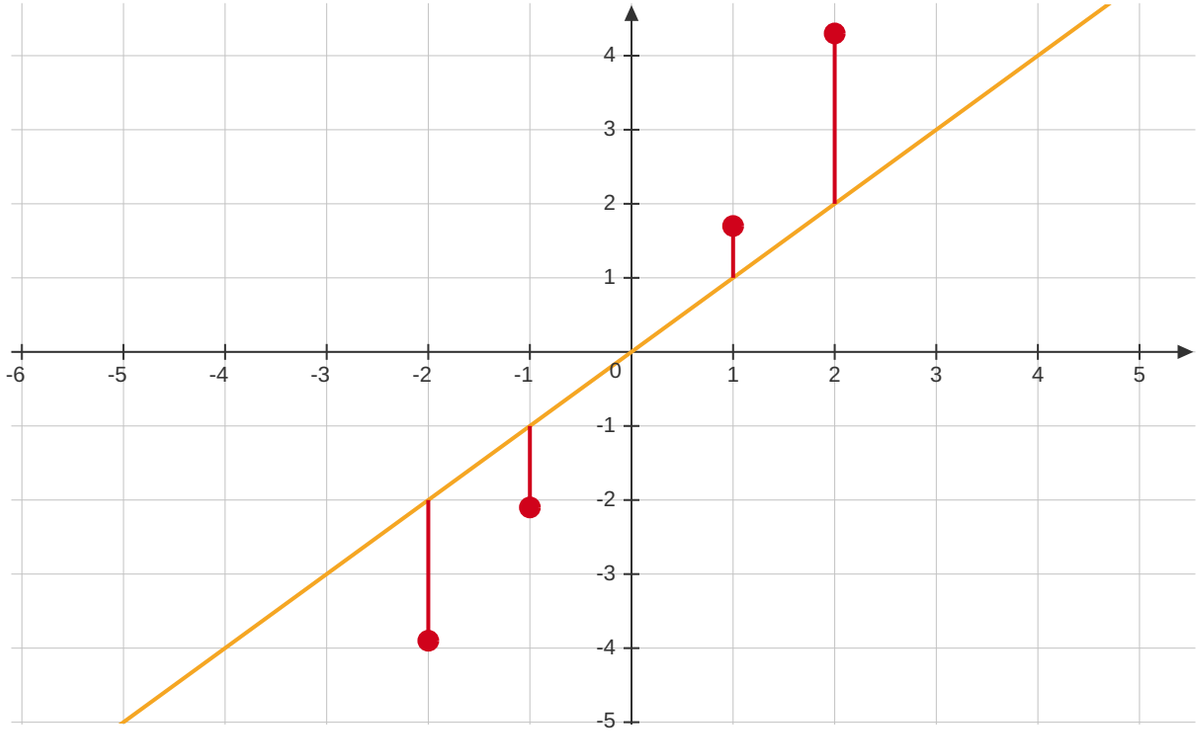
Linear Regression Loss Function

The most commonly used loss function for Regression is the Squared Sum Error. It is given by,

$$\sum_{i=1}^{n} ||\mathrm{w}^T \mathrm{x}_i - y_i||_2^2 \qquad \forall i$$

We use the line $y = x$ as a baseline to better understand the loss function.

In the above line, w is represented as $\mathrm{w} = [1]$. Using this in the loss function, we obtain,

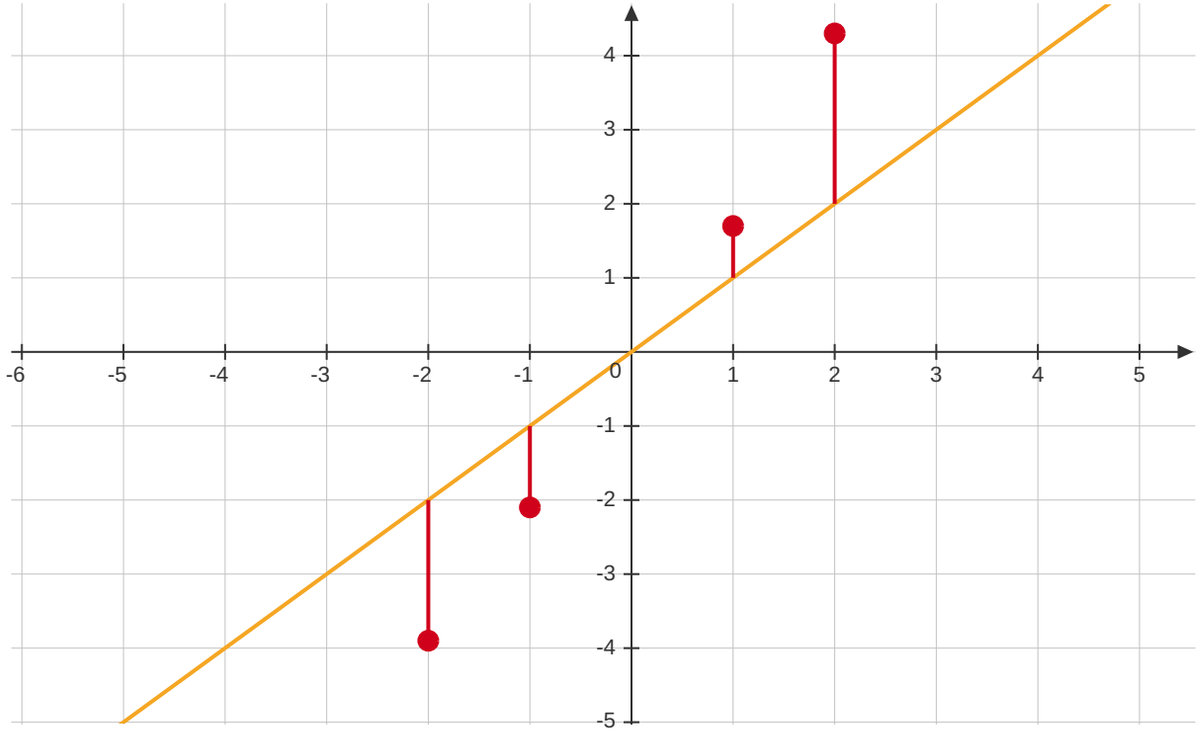


Linear Regression Loss Function

The most commonly used loss function for Regression is the Squared Sum Error. It is given by,

$$\sum_{i=1}^{n} ||\mathrm{w}^T \mathrm{x}_i - y_i||_2^2 \qquad \forall i$$

We use the line $y = x$ as a baseline to better understand the loss function.
In the above line, w is represented as $\mathrm{w} = [1]$. Using this in the loss function, we obtain,



$$||\mathrm{X}^T\mathrm{w} - \mathrm{y}||_2^2 = ||[1 \;\; 2 \;\; -1 \;\; -2]^T[1] - [1.7 \;\; 4.3 \;\; -2.1 \;\; -3.9]^T||_2^2 = ||[0.7 \;\; -2.3 \;\; 1.1 \;\; 1.9]^T||_2^2 = 3.25576^2 = 10.6$$

Optimizing the Error Function

Optimizing the Error Function

The minimization equation can be rewritten in the vectorized form as,

$$\min_{\mathrm{w}\in\mathbb{R}^d} \frac{1}{2}||\mathrm{X}^T\mathrm{w}-\mathrm{y}||_2^2$$

Let this be a function of w as follows:

$$f(\mathrm{w}) = \frac{1}{2}||\mathrm{X}^T\mathrm{w}-\mathrm{y}||_2^2$$

$$f(\mathrm{w}) = \frac{1}{2}(\mathrm{X}^T\mathrm{w}-\mathrm{y})^T(\mathrm{X}^T\mathrm{w}-\mathrm{y})$$

$$\therefore \bigtriangledown f(\mathrm{w}) = (\mathrm{X}\mathrm{X}^T)\mathrm{w}-(\mathrm{X}\mathrm{y})$$

Optimizing the Error Function

The minimization equation can be rewritten in the vectorized form as,

$$\min_{\mathrm{w}\in\mathbb{R}^d} \frac{1}{2}||\mathrm{X}^T\mathrm{w}-\mathrm{y}||_2^2$$

Let this be a function of w as follows:

$$f(\mathrm{w}) = \frac{1}{2}||\mathrm{X}^T\mathrm{w}-\mathrm{y}||_2^2$$
$$f(\mathrm{w}) = \frac{1}{2}(\mathrm{X}^T\mathrm{w}-\mathrm{y})^T(\mathrm{X}^T\mathrm{w}-\mathrm{y})$$
$$\therefore \bigtriangledown f(\mathrm{w}) = (\mathrm{X}\mathrm{X}^T)\mathrm{w}-(\mathrm{X}\mathrm{y})$$

Setting the above equation to zero, we get

$$(\mathrm{X}\mathrm{X}^T)\mathrm{w}-(\mathrm{X}\mathrm{y}) = 0$$
$$(\mathrm{X}\mathrm{X}^T)\mathrm{w}^* = \mathrm{X}\mathrm{y}$$
$$\therefore \mathrm{w}^* = (\mathrm{X}\mathrm{X}^T)^+\mathrm{X}\mathrm{y}$$

where $(\mathrm{X}\mathrm{X}^T)^+$ represents the pseudo-inverse of $\mathrm{X}\mathrm{X}^T$.

Using Least Squares Regression

Using Least Squares Regression

To account for the bias term, we augment our feature matrix with an additional row of ones.
Hence, our current dataset is,

$$\mathrm{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix}$$

Using least squares regression to obtain the weight vector, we get,

$$\begin{aligned} \mathrm{w} &= (\mathrm{X}\mathrm{X}^T)^+ \mathrm{X}\mathrm{y} \\ &= \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \right)^+ \begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} -3.9 \\ -2.1 \\ 1.7 \\ 4.3 \end{bmatrix} \\ \mathrm{w} &= \left(\begin{bmatrix} 4 & 0 \\ 0 & 10 \end{bmatrix} \right)^+ \begin{bmatrix} 0 \\ 20.2 \end{bmatrix} = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.1 \end{bmatrix} \begin{bmatrix} 0 \\ 20.2 \end{bmatrix} = \begin{bmatrix} 0 \\ 2.02 \end{bmatrix} \end{aligned}$$

Using Gradient Descent

Using Gradient Descent

Given that w^* is the solution of an unconstrained optimization problem, we can solve it using gradient descent, where:

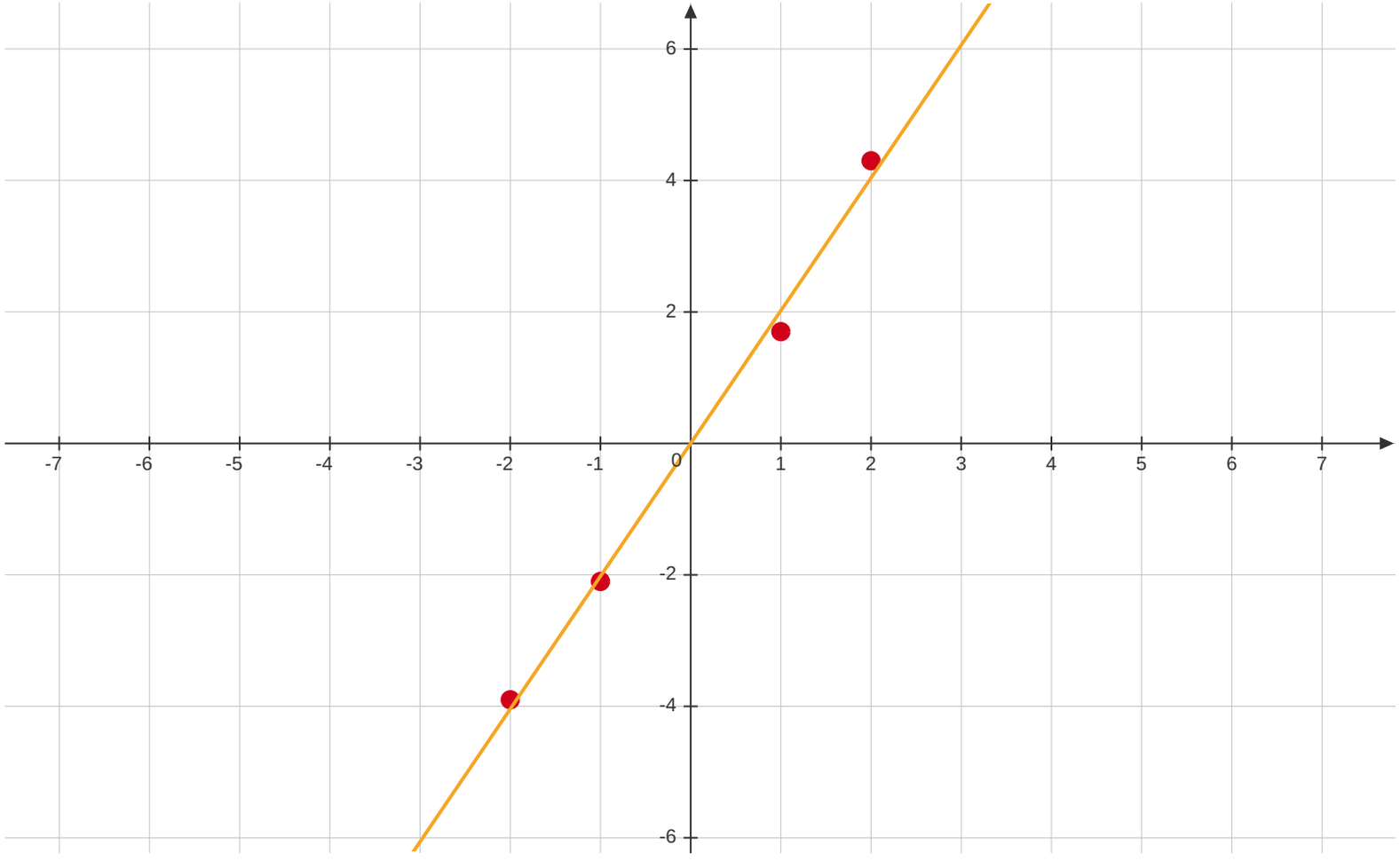
$$\mathrm{w}^{t+1} = \mathrm{w}^t - \eta^t \bigtriangledown f(\mathrm{w}^t)$$
$$\therefore \mathrm{w}^{t+1} = \mathrm{w}^t - \eta^t \big[(\mathrm{X}\mathrm{X}^T)\mathrm{w}^t - (\mathrm{X}\mathrm{y})\big]$$

Here, η is a scalar used to control the step size of the descent, and t is the current iteration.

Using gradient descent to obtain the weight vector, we have:

$$\mathrm{w}^1 = \mathrm{w}^0 - \eta^0 \big[(\mathrm{X}\mathrm{X}^T)\mathrm{w}^0 - (\mathrm{X}\mathrm{y})\big]$$
$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 \times \left[\left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \right) \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} -3.9 \\ -2.1 \\ 1.7 \\ 4.3 \end{bmatrix} \right]$$
$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 \times \left[- \begin{bmatrix} 0 \\ 20.2 \end{bmatrix} \right]$$
$$\mathrm{w}^1 = \begin{bmatrix} 0 \\ 2.02 \end{bmatrix}$$

Regression Line using Gradient Descent



Kernel Regression Algorithm

Kernel Regression Algorithm

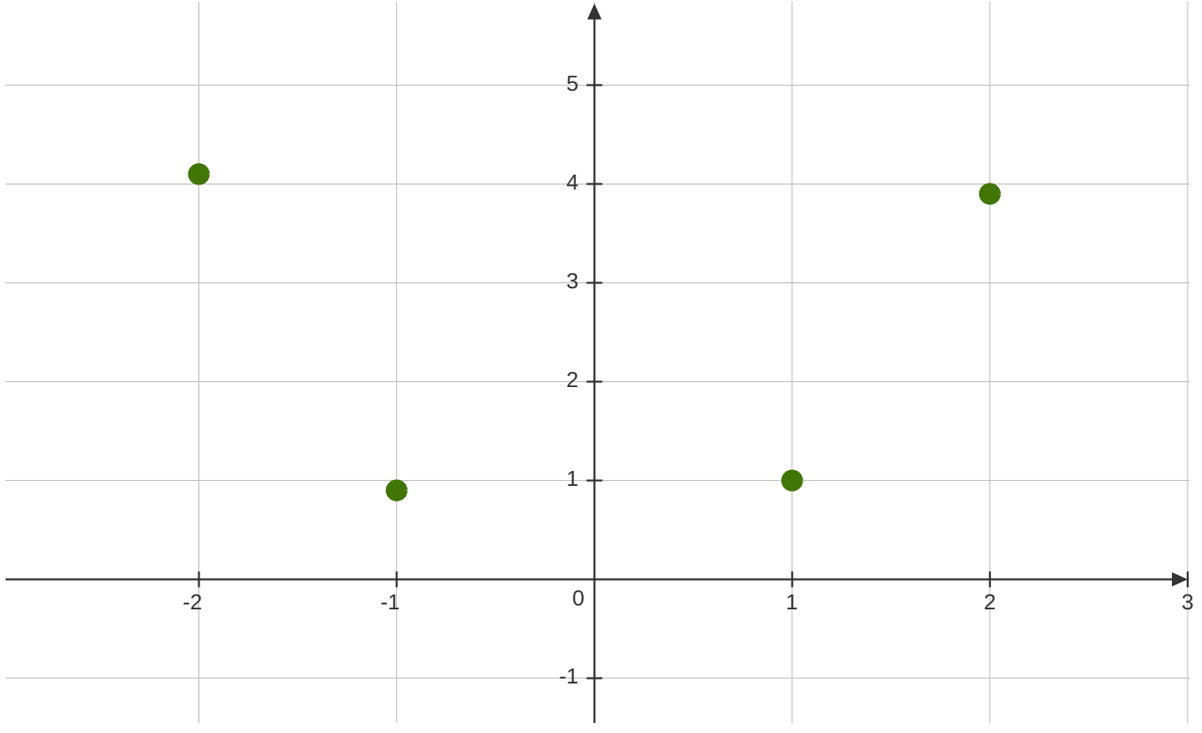
Given a dataset $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ where $\mathrm{x}_i \in \mathbb{R}^d$,
let $\{y_1, \ldots, y_n\}$ be the labels, where $y_i \in \mathbb{R}$.

$$\mathrm{X} = \begin{bmatrix} 1 & 2 & -1 & -2 \end{bmatrix} \quad \mathrm{y} = \begin{bmatrix} 1 \\ 3.9 \\ 0.9 \\ 4.1 \end{bmatrix}$$

Kernel Regression Algorithm

Given a dataset $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ where $\mathrm{x}_i \in \mathbb{R}^d$, let $\{y_1, \ldots, y_n\}$ be the labels, where $y_i \in \mathbb{R}$.

$$\mathrm{X} = \begin{bmatrix} 1 & 2 & -1 & -2 \end{bmatrix} \quad \mathrm{y} = \begin{bmatrix} 1 \\ 3.9 \\ 0.9 \\ 4.1 \end{bmatrix}$$



Using Kernel Regression

Using Kernel Regression

Let $\mathrm{w}^* = \mathrm{X}\alpha^*$ for some $\alpha^* \in \mathbb{R}^n$.

$$
\begin{aligned}
\mathrm{X}\alpha^* &= \mathrm{w}^* \\
\therefore \mathrm{X}\alpha^* &= (\mathrm{X}\mathrm{X}^T)^+\mathrm{X}\mathrm{y} \\
(\mathrm{X}\mathrm{X}^T)\mathrm{X}\alpha^* &= (\mathrm{X}\mathrm{X}^T)(\mathrm{X}\mathrm{X}^T)^+\mathrm{X}\mathrm{y} \\
(\mathrm{X}\mathrm{X}^T)\mathrm{X}\alpha^* &= \mathrm{X}\mathrm{y} \\
\mathrm{X}^T(\mathrm{X}\mathrm{X}^T)\mathrm{X}\alpha^* &= \mathrm{X}^T\mathrm{X}\mathrm{y} \\
(\mathrm{X}^T\mathrm{X})^2\alpha^* &= \mathrm{X}^T\mathrm{X}\mathrm{y} \\
\mathrm{K}^2\alpha^* &= \mathrm{K}\mathrm{y} \\
\therefore \alpha^* &= \mathrm{K}^{-1}\mathrm{y}
\end{aligned}
$$

where $\mathrm{K} \in \mathbb{R}^{n \times n}$ and K can be also obtained using a kernel function like the Polynomial Kernel or RBF Kernel

Let's use the polynomial kernel of degree of two. By applying the kernel function to the dataset, we obtain,

Let's use the polynomial kernel of degree of two. By applying the kernel function to the dataset, we obtain,

$$k(\mathrm{x}_i, \mathrm{x}_j) = \left(\mathrm{x}_i^T \mathrm{x}_j \ + \ 1\right)^2$$

$$k(\mathrm{x}_1, \mathrm{x}_1) = \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 1 \right)^2 = (\ 2 + 1\ )^2 = 9$$

$$k(\mathrm{x}_1, \mathrm{x}_2) = \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 1 \right)^2 = (\ 6 + 1\ )^2 = 16$$

$$\therefore \ \mathrm{K} = \begin{bmatrix} 9 & 16 & 1 & 16 \\ 16 & 36 & 0 & 36 \\ 1 & 0 & 9 & 0 \\ 16 & 36 & 0 & 36 \end{bmatrix}$$

Let's use the polynomial kernel of degree of two. By applying the kernel function to the dataset, we obtain,

$$k(\mathrm{x}_i, \mathrm{x}_j) = \left(\mathrm{x}_i^T \mathrm{x}_j \ + \ 1\right)^2$$

We find α using the following equation,

$$k(\mathrm{x}_1, \mathrm{x}_1) = \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 1 \right)^2 = (\ 2 + 1\ )^2 = 9$$

$$k(\mathrm{x}_1, \mathrm{x}_2) = \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 1 \right)^2 = (\ 6 + 1\ )^2 = 16$$

$$\therefore \ \mathrm{K} = \begin{bmatrix} 9 & 16 & 1 & 16 \\ 16 & 36 & 0 & 36 \\ 1 & 0 & 9 & 0 \\ 16 & 36 & 0 & 36 \end{bmatrix}$$

Let's use the polynomial kernel of degree of two. By applying the kernel function to the dataset, we obtain,

$$k(\mathrm{x}_i, \mathrm{x}_j) = \left(\mathrm{x}_i^T \mathrm{x}_j \ + \ 1\right)^2$$

$$k(\mathrm{x}_1, \mathrm{x}_1) = \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 1 \right)^2 = (\ 2 + 1\ )^2 = 9$$

$$k(\mathrm{x}_1, \mathrm{x}_2) = \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 1 \right)^2 = (\ 6 + 1\ )^2 = 16$$

$$\therefore \ \mathrm{K} = \begin{bmatrix} 9 & 16 & 1 & 16 \\ 16 & 36 & 0 & 36 \\ 1 & 0 & 9 & 0 \\ 16 & 36 & 0 & 36 \end{bmatrix}$$

We find α using the following equation,

$$\alpha = \mathrm{K}^{-1}\mathrm{y}$$

$$= \begin{bmatrix} 9 & 16 & 1 & 16 \\ 16 & 36 & 0 & 36 \\ 1 & 0 & 9 & 0 \\ 16 & 36 & 0 & 36 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 3.9 \\ 0.9 \\ 4.1 \end{bmatrix}$$

$$\alpha = \begin{bmatrix} -0.1813 \\ 0.1707 \\ -0.1798 \\ 0.1737 \end{bmatrix}$$

Finding the Predictions using α

Finding the Predictions using α

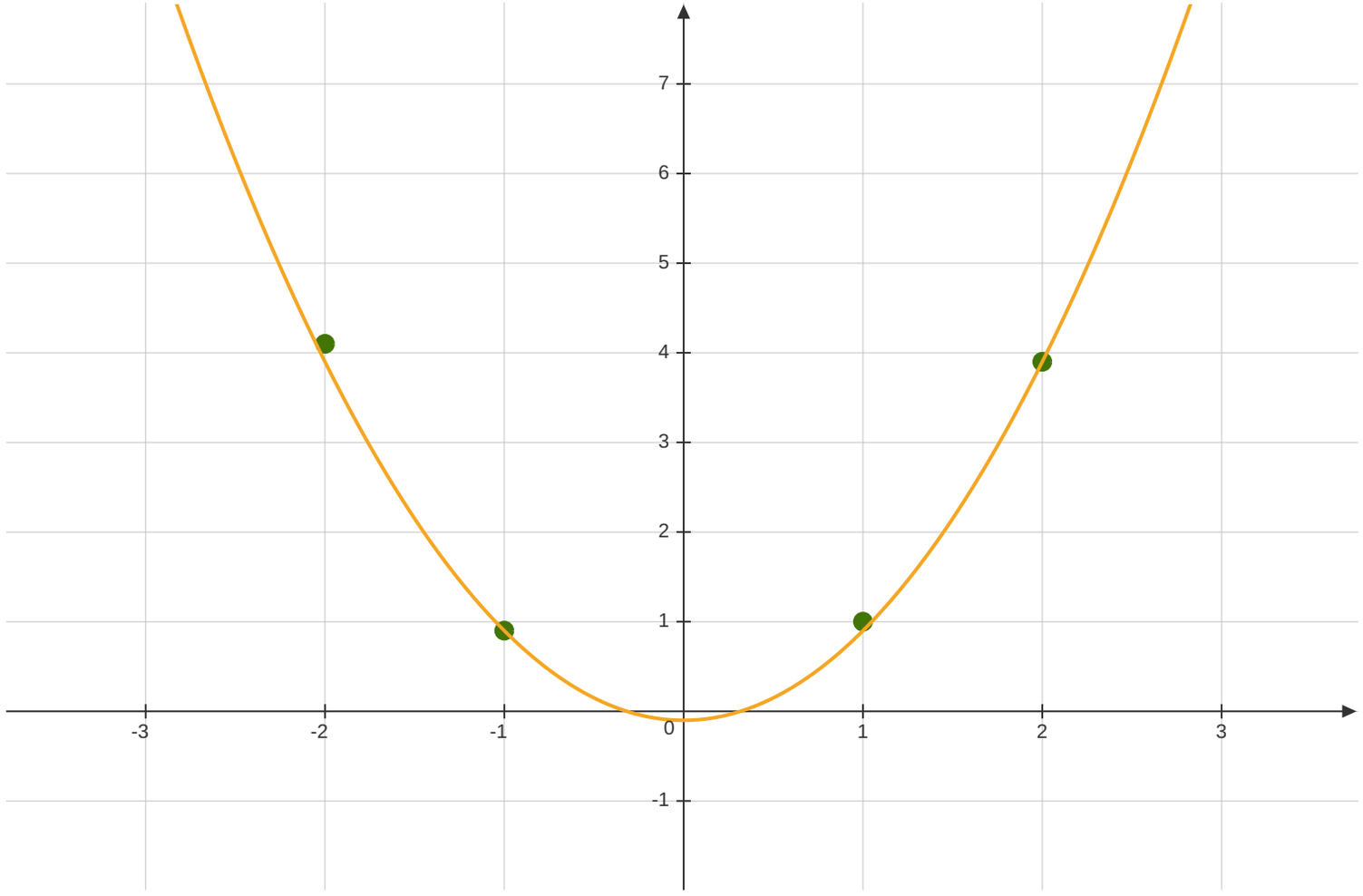
We use the following equation to get y_{pred}

$$\mathrm{w}^{*}\phi(\mathrm{X}) = \sum_{i=1}^{n} \alpha_i^{*} k(\mathrm{x}_i, \mathrm{x}_i)$$

where α_i^{*} gives the importance of the i^{th} datapoint towards w^{*}.

$$\begin{aligned}
\mathrm{y}_{pred} &= \mathrm{K}^{T}\alpha \\
&= \begin{bmatrix} 9 & 16 & 1 & 16 \\ 16 & 36 & 0 & 36 \\ 1 & 0 & 9 & 0 \\ 16 & 36 & 0 & 36 \end{bmatrix} \begin{bmatrix} -0.1813 \\ 0.1707 \\ -0.1798 \\ 0.1737 \end{bmatrix} \\
\mathrm{y}_{pred} &= \begin{bmatrix} 0.92 \\ 3.94 \\ 0.98 \\ 4.06 \end{bmatrix}
\end{aligned}$$

Regression Line using Kernel Regression



Question 1:

Q. Gaussian kernel regression with parameter $\sigma^2 = 1/2$ was applied to the following dataset with two features:

$$\mathrm{X} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad \mathrm{y} = \begin{bmatrix} 2.1 & 1 & 2 & 1.2 \end{bmatrix}^T$$

The weight vector can be written as $\mathrm{w} = \phi(\mathrm{X})\,\alpha$ where $\phi(\mathrm{X})$ is the transformation mapping corresponding to the kernel. The vector α is given by $\begin{bmatrix} 2.1 & -2.1 & 3 & 0 \end{bmatrix}^T$ which is obtained as $\mathrm{K}^{-1}\mathrm{y}$ where K is the kernel matrix. What will be the prediction for point $\begin{bmatrix} 1 & 1 \end{bmatrix}^T$?

A. The kernel function is given by,

$$k(\mathrm{x}_i, \mathrm{x}_j) = exp\left(\frac{-||\mathrm{x}_i - \mathrm{x}_j||^2}{2\sigma^2}\right) = exp\left(\frac{-||\mathrm{x}_i - \mathrm{x}_j||^2}{2\left(\frac{1}{2}\right)}\right) = exp\left(-||\mathrm{x}_i - \mathrm{x}_j||^2\right)$$

The prediction is given by,

$$\begin{aligned}
\mathrm{K}^T\alpha &= \begin{bmatrix} k(\mathrm{x}_1, \mathrm{x}_{test}) & k(\mathrm{x}_2, \mathrm{x}_{test}) & k(\mathrm{x}_3, \mathrm{x}_{test}) & k(\mathrm{x}_4, \mathrm{x}_{test}) \end{bmatrix} \begin{bmatrix} 2.1 \\ -2.1 \\ 3 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} k\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) & k\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) & k\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) & k\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \end{bmatrix} \begin{bmatrix} 2.1 \\ -2.1 \\ 3 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} e^{-1} & e^{-1} & e^{0} & e^{-2} \end{bmatrix} \begin{bmatrix} 2.1 \\ -2.1 \\ 3 \\ 0 \end{bmatrix} \\
&= 2.1e^{-1} - 2.1e^{-1} + 3e^{0} + 0e^{-2} \\
\mathrm{K}^T\alpha &= 3
\end{aligned}$$