Week 02 - Part 3

Review: - Deriving $\underline{w}_{LS}$ by finding the gradient an setting it to zero

- Deriving $\underline{w}_{LS}$ by (pseudo)-solving the system of linear equations $\underline{y} = X\,\underline{w}$.

Today: - Deriving $\underline{w}_{LS}$ with Geometric interpretations

- Regularized least squares
- Non-linear transformation

Recall:

- Least square solution: $\underline{w}_{ls} = X^{\dagger} \underline{y} = (X^T X)^{-1} X^T \underline{y}$

- prediction by $\underline{w}_{ls}$: $\hat{\underline{y}}_{ls} = X \underline{w}_{ls} = X X^{\dagger} \underline{y}$

  - It's like we take $\underline{y}$ and with a projection matrix transforming it into $\hat{\underline{y}}_{ls}$.
  - $XX^{\dagger}$ is a projection matrix.

- This observation leads us into geometric interpretation of Least squares.

Geometric Interpretation of Least Squares

- Observe that $\underline{\hat{y}} = X \underline{w} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1d} \\ x_{20} & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{N0} & x_{N1} & \cdots & x_{Nd} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$

$$= w_0 \begin{bmatrix} x_{10} \\ x_{20} \\ \vdots \\ x_{N0} \end{bmatrix} + w_1 \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{bmatrix} + \cdots + w_d \begin{bmatrix} x_{1d} \\ \vdots \\ x_{Nd} \end{bmatrix}$$

- So, $\underline{\hat{y}}$ is Linear Combination of Columns of X.

- Thus, $\underline{\hat{y}}$ is in the space of all possible linear Combinations of Columns of X

- The space of all possible linear Combination of Columns of X is called Col-span$\{X\}$

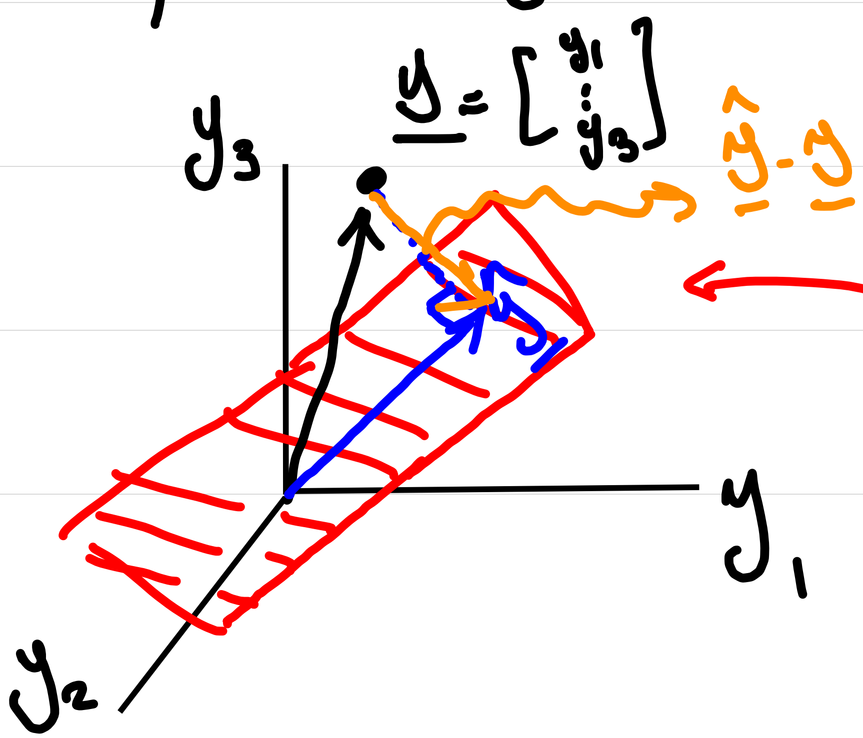
■ Let's illustrate Col-span(X) for $N=3$, $d=1$, and $X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$.

$$\text{Col-span}(X) = \left\{ w_0 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + w_1 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} : w_0, w_1 \in \mathbb{R} \right\} = \left\{ \begin{bmatrix} w_0 \\ w_0 + w_1 \\ w_0 + 2w_1 \end{bmatrix} : w_0, w_1 \in \mathbb{R} \right\}$$

$$= \left\{ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} : y_1 = w_0,\ y_2 = w_0 + w_1,\ y_3 = w_0 + 2w_1 \text{ and } w_0, w_1 \in \mathbb{R} \right\}$$

$$= \left\{ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} : 2y_2 - y_3 = y_1, \text{ and } y_1, y_2, y_3 \in \mathbb{R} \right\} = \left\{ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} : \underbrace{-y_1 + 2y_2 - y_3 = 0}_{\text{a subspace of 3-dim space.}} \right\}$$

(it is in fact a plane)

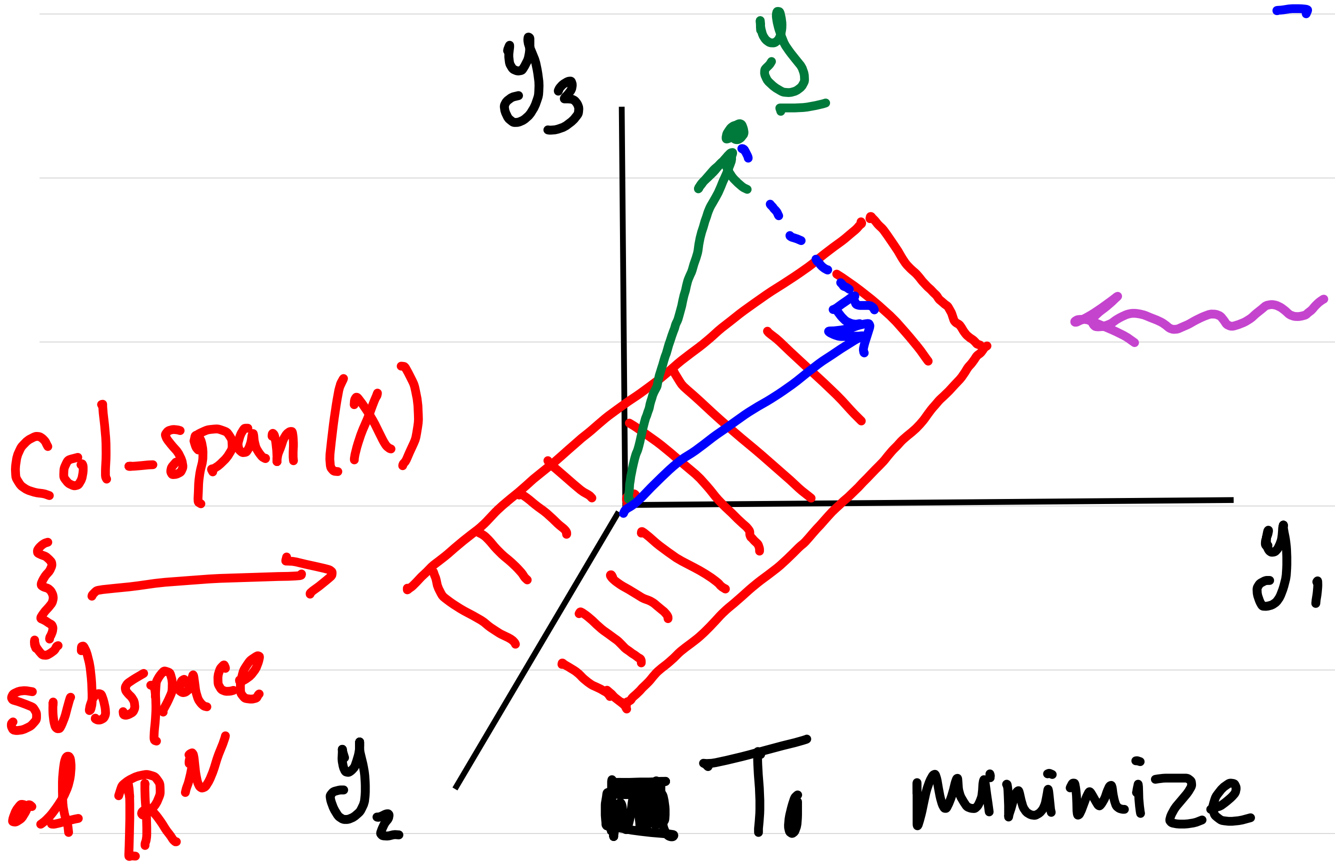


Col-span(X) = Space of $\underline{\hat{y}}$

(is subspace of $\mathbb{R}^N$)

- $\text{Col-span}(X) = \left\{ \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} : \underbrace{-y_1 + 2y_2 - y_3 = 0} \right\}$

- A plane that goes through origin.
- It is a subspace of $\mathbb{R}^N$.



⟵ the space of possible $\hat{\underline{y}}$.

- To minimize $\|\underline{y} - \hat{\underline{y}}\|$: Euclidean distance b/w $\underline{y}$ & $\hat{\underline{y}}$

- must find $\hat{\underline{y}}$ on Col-span{X} that is closest to $\underline{y}$.

- The best $\underline{\hat{y}}$ (i.e. $\hat{y}_{LS}$) is the Projection of $\underline{y}$ onto Col-span$\{X\}$.

- That means $(\underline{y} - \underline{\hat{y}}_{LS})$ must be orthogonal to any vector in Col-span$\{X\}$.

- Thus, $(\underline{y} - \underline{\hat{y}}_{ls})$ is orthogonal to every column of X.

Reminder: $\underline{a} \perp \underline{b} \Longleftrightarrow \underline{a}^T \underline{b} = 0$

- Thus, $X^T(\underline{y} - \underline{\hat{y}}_{ls}) = \underline{0} \Longrightarrow X^T(\underline{y} - X\underline{w}_{ls}) = 0$

$$\Rightarrow X^T X \underline{w}_{ls} = X^T \underline{y} \Longrightarrow \underline{w}_{ls} = (X^T X)^{-1} X^T \underline{y}.$$

Regularized Linear Regression / Least squares

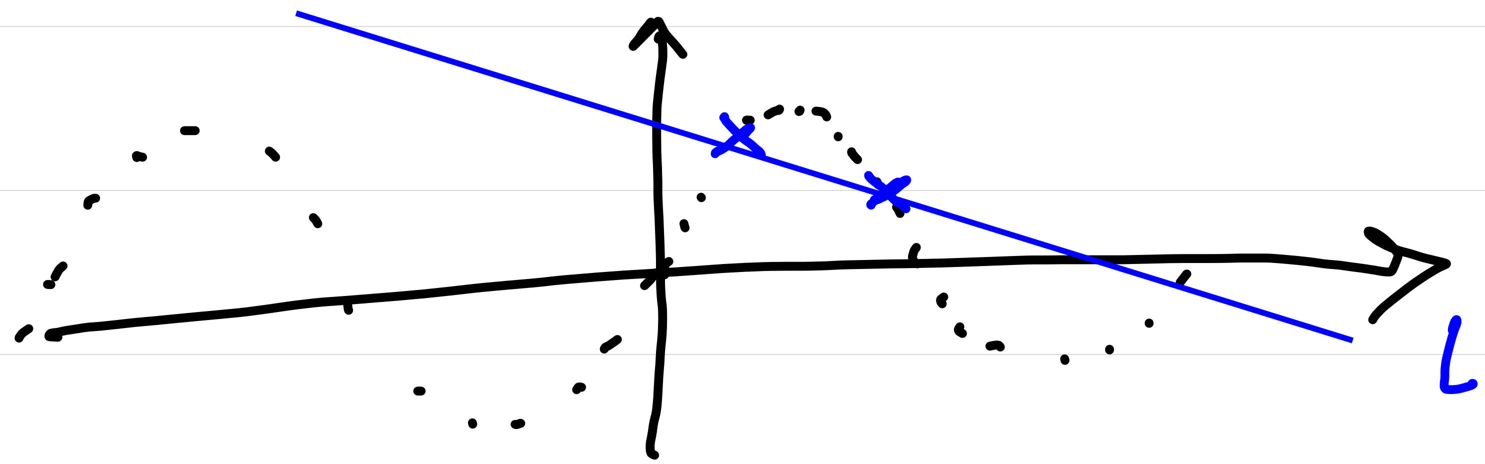
- Previously, we tried to minimize $\|X\underline{w} - \underline{y}\|^2$

- In regularized version, we minimize $\|X\underline{w} - \underline{y}\|^2 + \underbrace{\lambda \|\underline{w}\|^2}_{\text{penalty function}}$

  penalty function (against large weight)

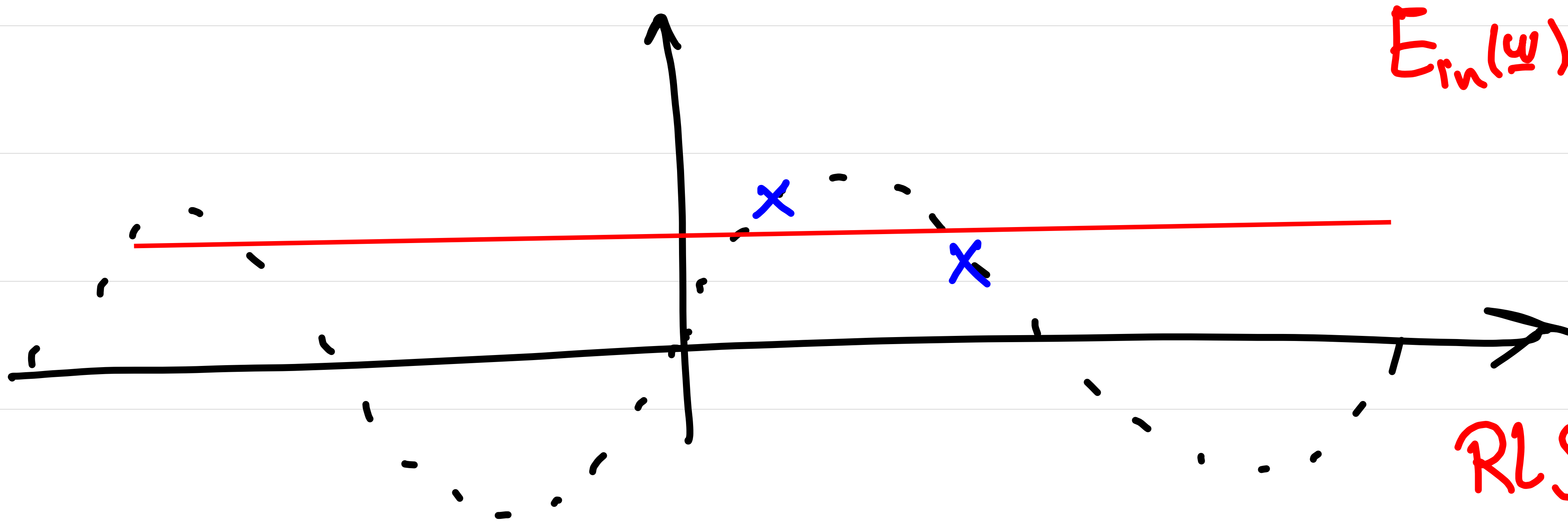
- The motivation is to avoid overfitting
  - your data is noisy
  - you do not have enough data (compared to the complexity of the target function)

■ E.g. target: $f(x) = \sin(\pi x)$



LS Solution

$$E_{in}(\underline{w}) = \|\underline{y} - X\underline{w}\|^2 + \lambda \|\underline{w}\|_2^2$$



RLS Solution
(smaller slope)

Note: ① $\lambda = 0 \Longrightarrow LS$

② How to choose λ? validation

How do we Solve this Problem?

- We want to $\min_{\underline{w}} \|X\underline{w} - \underline{y}\|^2 + \lambda \|\underline{w}\|^2$

- Let $f(\underline{w}) = \|X\underline{w} - \underline{y}\|^2 + \lambda \|\underline{w}\|^2$

- Observe that $\nabla_{\underline{w}} f(\underline{w}) = 2X^T(X\underline{w} - \underline{y}) + 2\lambda \underline{w}$

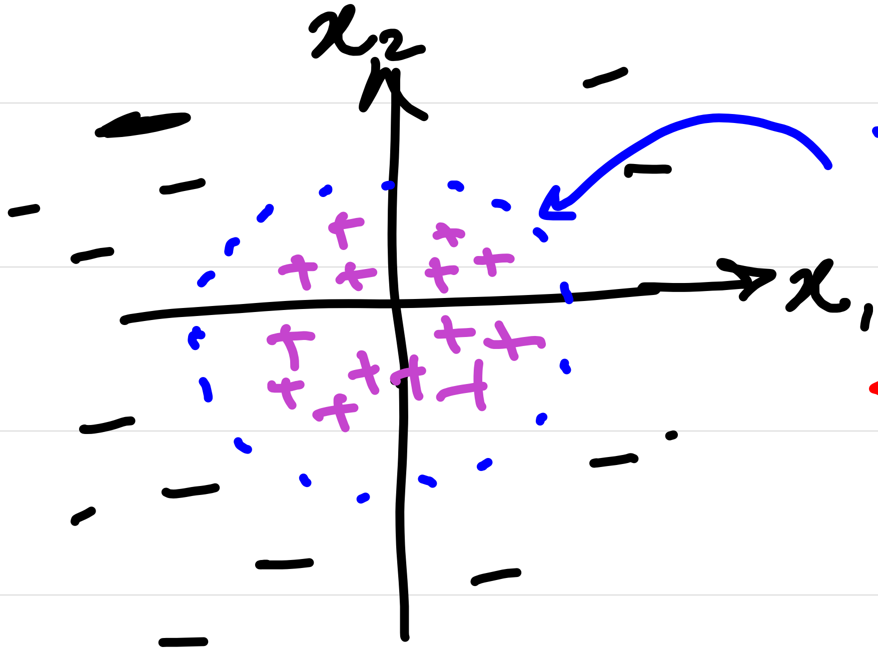
- We want $\nabla_{\underline{w}} f(\underline{w}) = 0 \Rightarrow (X^TX + \lambda I)\underline{w} = X^T\underline{y}$

$$\Rightarrow \underline{w}_{RLS} = (X^TX + \lambda I)^{-1} X^T \underline{y}$$

So far: We studied linear Models

- But in many cases linear Models are not good enough.

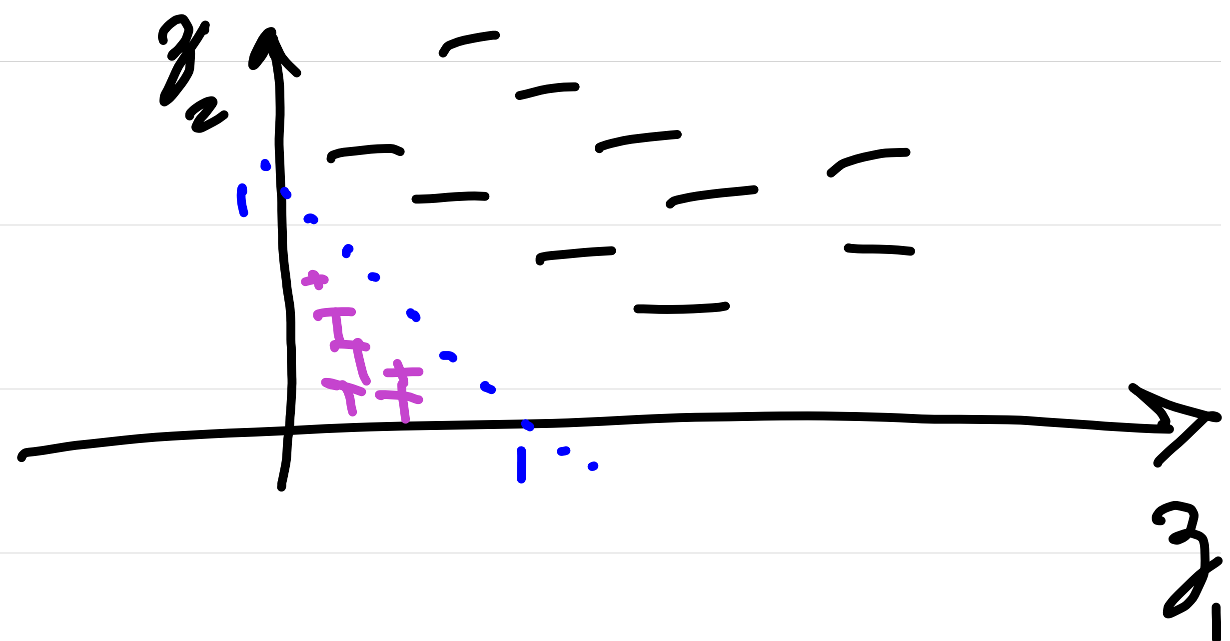
E.g.



true decision boundary $x_1^2+x_2^2=1$

linear decision boundaries won't perform well.

Then define $z_1 = x_1^2$ and $z_2 = x_2^2$



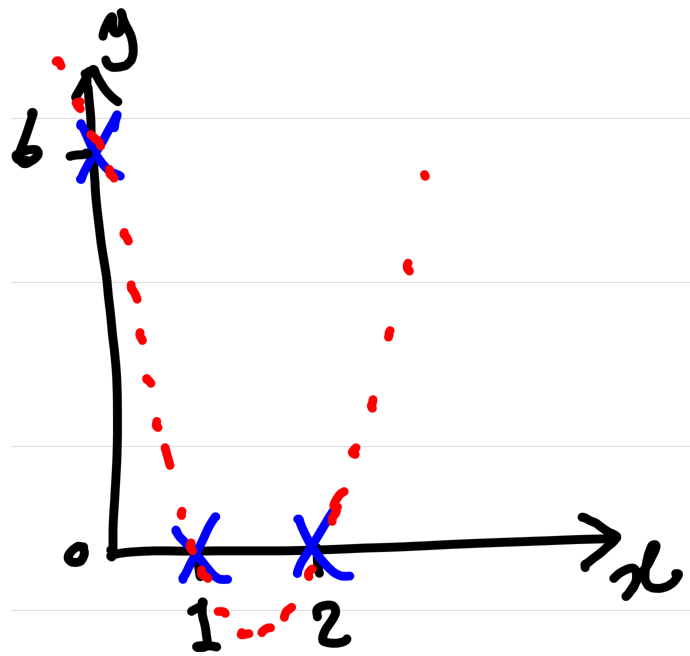
The points are linearly separable in Z-space.

Suppose PLA gives you $h(\underline{z}) = \text{Sign}(z_1 + z_2 - 1)$. Then, we know $g(\underline{x}) = \text{Sign}(x_1^2 + x_2^2 - 1)$

In general:

- Let $\underline{z} = \Phi(\underline{x})$ be non-linear transformation ("feature transformation")
- Let $h(\underline{z})$ be a linear classifier / regression function in $\underline{z}$ space $\left(h(\underline{z}) = \text{Sign}(\underline{w}^T \underline{z}) \text{ or } h(\underline{z}) = \underline{w}^T \underline{z}\right)$
- Then $g(\underline{x}) = h(\Phi(\underline{x}))$ is non-linear classifier in $\underline{x}$ space

E.g. Quadratic Regression



- Define $\underline{z} = (z_0 = 1,\ z_1 = x,\ z_2 = x^2)$

- $\underline{y} = \underline{w}^T \underline{z} = w_0 + w_1 z_1 + w_2 z_2$ ← linear in $\underline{z}$

$$= w_0 + w_1 x + w_2 x^2$$ ← Quadratic in $\underline{x}$

x_{n_0} is always 1 Since augmented

- Let's find the $\underline{w}_{LS}$: $\underline{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\underline{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\underline{x}_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$$\underline{z}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},\quad \underline{z}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},\quad \underline{z}_3 = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \rightarrow Z = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

$$\underline{y} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$$

$$\underline{w}_{LS} = (Z^T Z)^{-1} Z^T \underline{y} = \cdots = \begin{bmatrix} 6 \\ -9 \\ 3 \end{bmatrix} \Longrightarrow$$

$$\hat{y} = \underline{w}_{LS}^T \underline{z} = 6 - 9 z_1 + 3 z_2 = 6 - 9x + 3x^2$$