Linear Regression w/ polynomial features

X = feature matrix

$$X = \begin{bmatrix} x_1^0 & x_1^1 & \cdots & x_1^M \\ x_2^0 & x_2^1 & \cdots & x_2^M \\ \vdots & \vdots & & \\ x_N^0 & x_N^1 & \cdots & x_N^M \end{bmatrix}$$ is of size $N \times (M+1)$

$$t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

$$w^* = (X^T X)^{-1} \cdot X^T \cdot t$$

what happens when $N < M$?

$$X^T X \quad \left((M+1) \times \underline{N} \right) \times \left(\underline{N} \times (M+1) \right) = \boxed{(M+1) \times (M+1)}$$

$X^T X$ is singular $\Rightarrow \det(X^T X) = 0$

Then $(X^T X)^{-1}$ does <u>not</u> exist!

$X^T X$ is not full Rank

$$X^T X + \lambda I$$

depending on value of λ,
then $(X^T X + \lambda I)^{-1}$ exists.

DATA

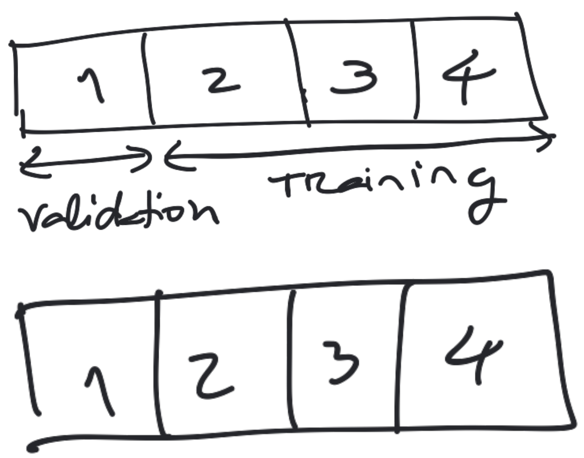
training, 80%

TEST, 20%

this is the set to perform cross-validation

STEPS:

4-fold CV:



1. Shuffle training samples.
2. Partition it in k folds.
3. For some set of parameters

   assuming R^2 metric

   3.1 Train w/ 2,3,4, METRIC: 0.9
   Validate w/ 1, METRIC: 0.7

   3.2 Train w/ 1,3,4, METRIC: 0.99
   Validate w/ 2, METRIC: 0.01

   3.3 Train w/ 1,2,4, METRIC: 0.89
   validate w/ 3, METRIC: 0.3

   3.4 Train w/ 1,2,3, METRIC: 0.9
   Validate w/ 4, METRIC: 0.2

   validation metric $= \dfrac{0.7+0.01+0.3+0.2}{4}$

4. Choose set of parameters that maximize R^2 score in validation

Ways to avoid overfitting?

1) More data
2) Cross-validation
   → K-fold CV
3) Regularization

Regularization

We saw that w^* compute very values when the model is overfitting.

$$\underbrace{J}_{\text{Mean squared ERROR}} = \frac{1}{N}\sum_{i=1}^{N} e_i^2 + \underbrace{\lambda \cdot \boxed{\sum_{i=0}^{M} w_i^2}}_{\text{penalty term that enforces weights to be small}}$$

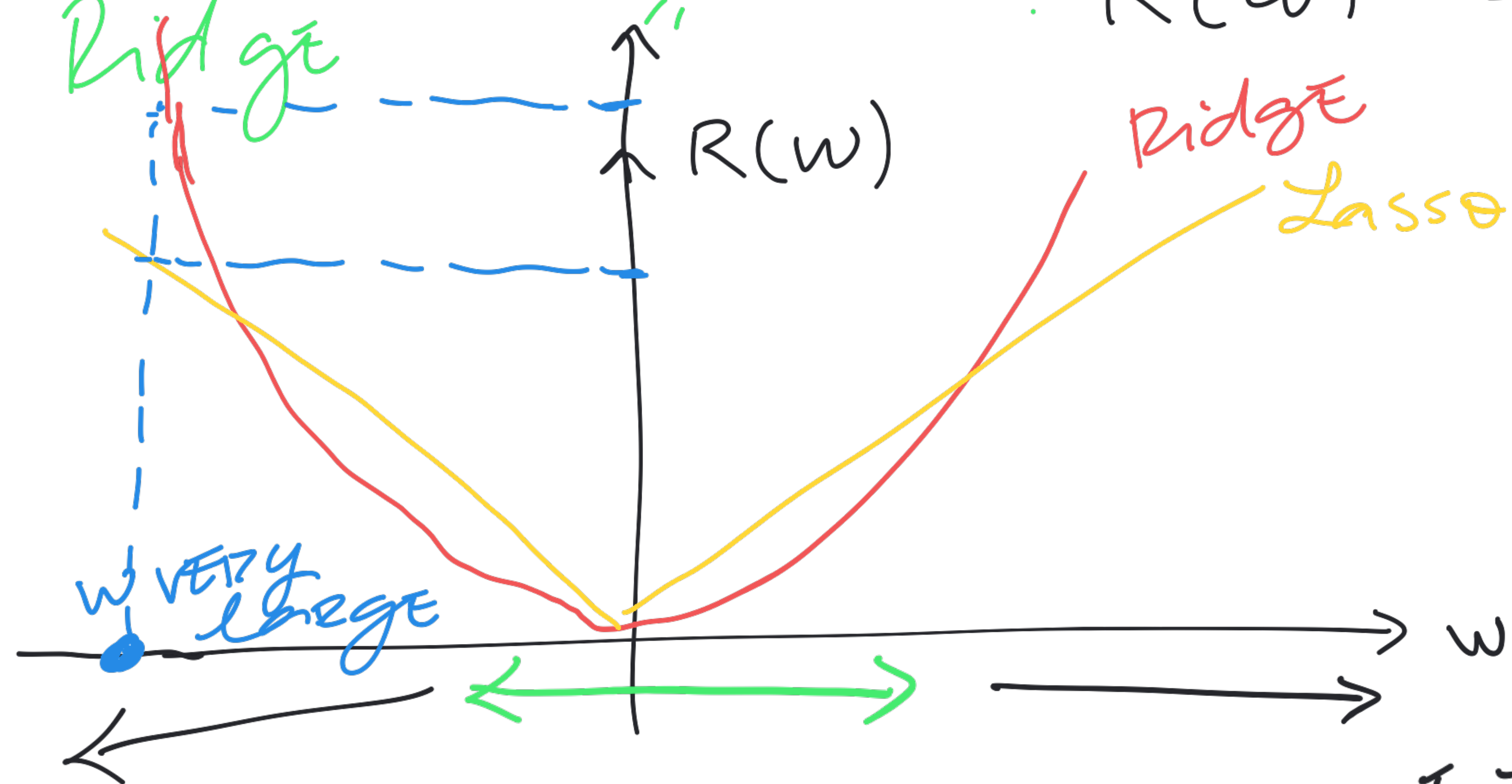
$$= w^T w = \|w\|_2^2$$

Ridge Regularizer

$$\boxed{\arg\min_{w} J}$$

Lasso regularizer : $R(w)=\sum_i |w_i|$

Ridge : $R(w)=\sum_i w_i^2$



Observations:

① Ridge will penalize more than Lasso for large parameter values

② ridge is highly affected by outliers

③ Lasso is going to force the weight values to zero, i.e., promotes sparsity

Some elements of w are exactly zero.

$w = [0.5, 0.5, 1]$, $\|w\|_2^2 = 0.5^2 + 0.5^2 + 1^2 = 1.5$

, $\|w\|_1 = 0.5 + 0.5 + 1 = 2$

$w = [0, 0, 2]$, $\|w\|_2^2 = 4$

$\|w\|_1 = 2$

features

So, the model:

$$y = \underline{\underline{w_0}} \cdot x^0 + \underline{\underline{w_1}} x^1 + \dots + w_m x^m$$

if I use Lasso regularizer

$\Rightarrow$ some weights will go to zero

$\Rightarrow$ feature selection