FoDA L13

Cross-Validation

(for Regression)

Oct 4, 2022

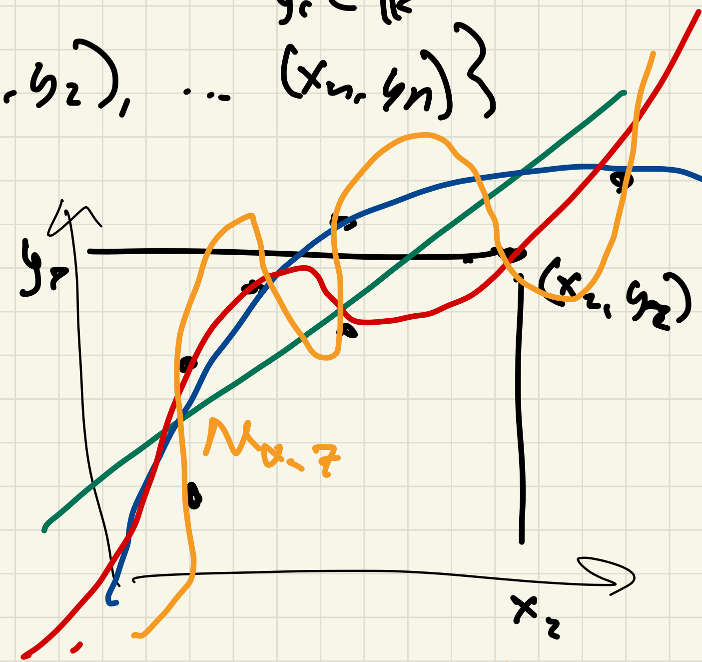
Revisit Polynomial Regression

$(x-2)^2 + 3x^2(2-x^3)$

Input $(X, y) \subset \mathbb{R} \times \mathbb{R}$ $\quad x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$

$= \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$

$$\tilde{X}_p = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & & x_2^p \\ 1 & & & & \\ 1 & x_n & x_n^2 & & x_n^p \end{bmatrix}$$



$$M_{\alpha, p}(x) = \sum_{j=0}^{p} \alpha_j x^j$$

$\alpha = (\alpha_0, \ldots \alpha_p)$

$$\alpha^* = (\tilde{X}_p^T \tilde{X}_p)^{-1} \tilde{X}_p^T y$$

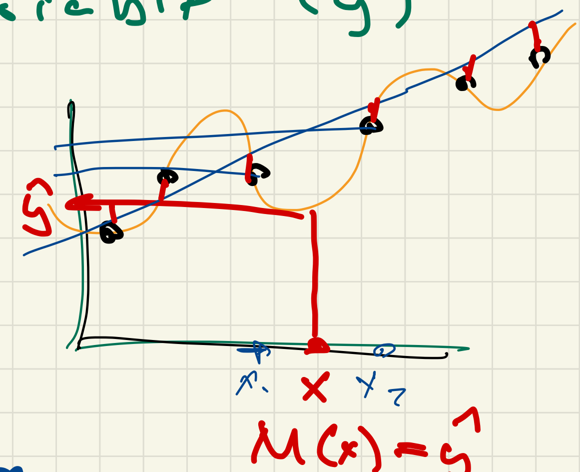
$SSE((x, y), M_{\alpha, 7}) = 0$

What makes a good Model?

$M_\alpha : \mathbb{R} \rightarrow \mathbb{R}$ (domain x → y)

Input $(x, y) \sim \mathcal{U}$ iid

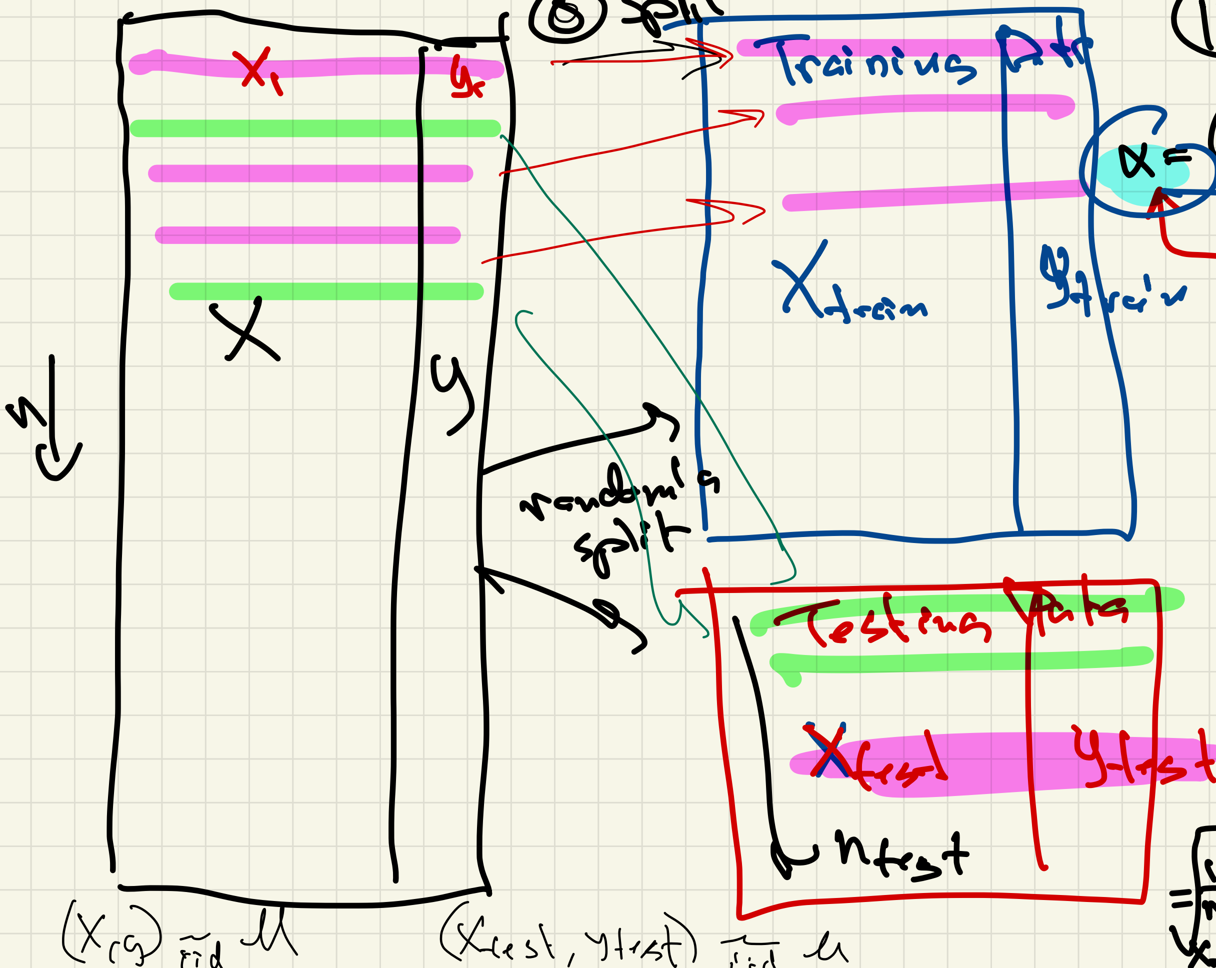
- predict dependent variable (y)
- don't overfit
- robust to noise



- For $M(x)$ perform well on new data x

Data Splitting for C-V

⓪ Split



① Train

$$\alpha = (X_{train}^T X_{train})^{-1} X_{train}^T y_{train}$$

Model M_α from training data

② Evaluate

$$SSE(\{X_{test}, y_{test}\}, M_\alpha)$$

$$RMSE(M_\alpha, X_{test}, y_{test}) = \sqrt{\frac{1}{n} \sum_{x_i \in X_{test}} (M_\alpha(x_i) - y_i)^2}$$

Evaluate

Lasso

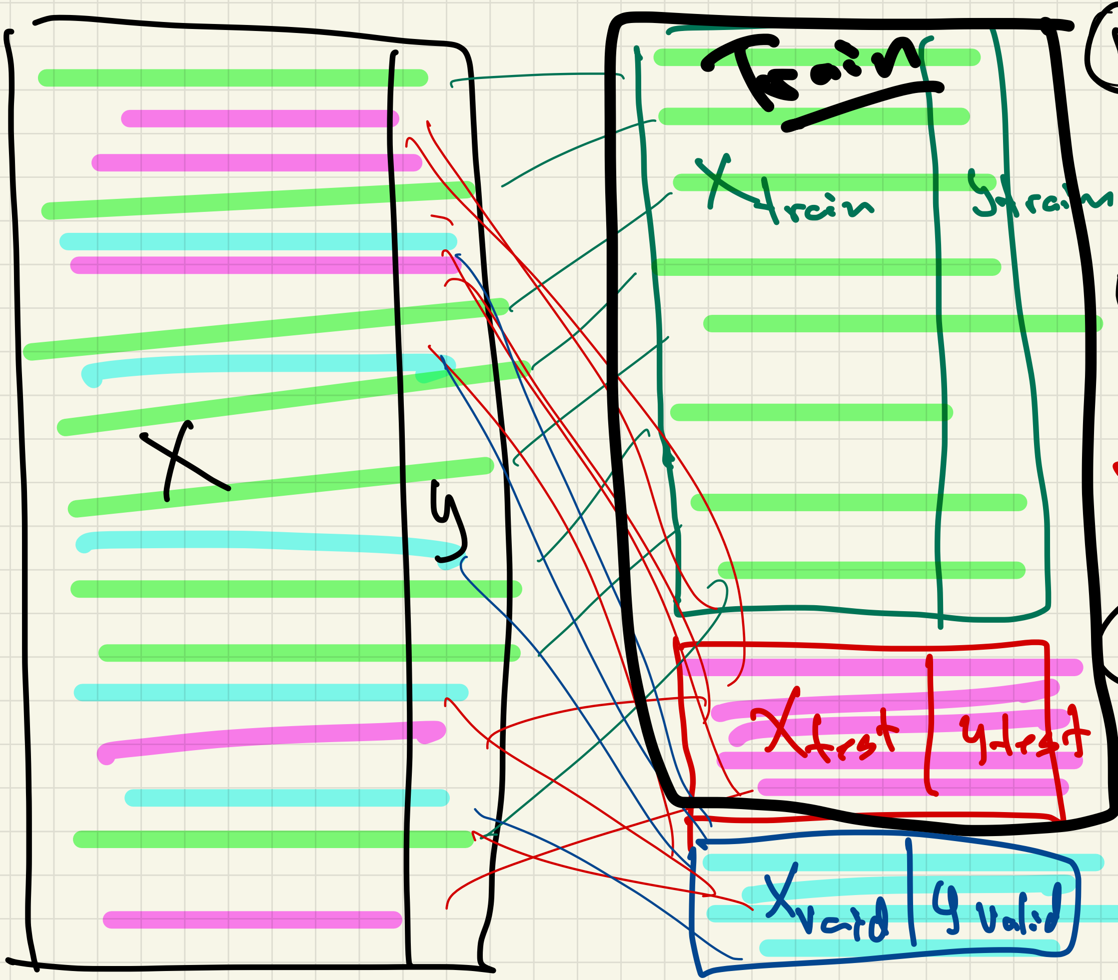
① How well do we expect Mx to work on new data?

↳ RMSE

② How do we choose best (hyper)parameters?

P = degree polynomial

Do not do both at once.



1. Train

$\alpha \leftarrow X_{train}$

2. Choose hyperparam

$p \leftarrow X_{test}$

3. Evaluate

X_{valid}

How to split?

→ Random (safer)

80/20   70/30   90/10

→ Need enough training $x \in \mathbb{R}^d$

→ Need enough test to predict. SSE $\in \mathbb{R}$

What if data is small?

$n = 10, 20$

Leave-one-out LOO-CV

n splits $\left[(n-1) \text{ train} \quad 1 \text{ test.} \right] \to$ Average