L13: Cross-Validation

Mon, Feb 23

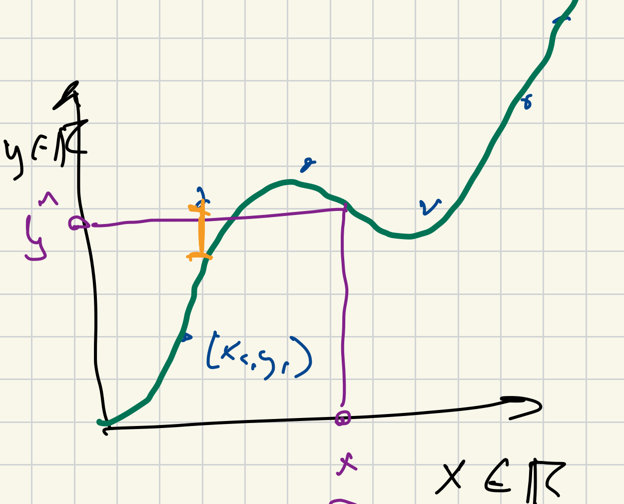
FoDA

Jeff M. Phillips

Polynomial Regression

Input (x, y) $\qquad \{(x_i, y_i)\}$ $\qquad y_i \in \mathbb{R}$, $\; x_i \in \mathbb{R}$

$$\tilde{X}_p = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & & x_2^p \\ & \vdots & & & \\ & \vdots & & & \\ 1 & x_n & x_n^2 & & x_n^p \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$$



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

$$\hat{y} = m_\alpha(x) = \sum_{j=0}^{p} \alpha_j x^p$$

$$\langle \alpha, (1, x, x^2, \dots, x^p) \rangle$$

What makes a good model? M_α

Supervised setting (X, y) $M_\alpha(x_i) \approx y_i$

1. predict on new data not trained on.
from some distribution $\mathcal{U}$

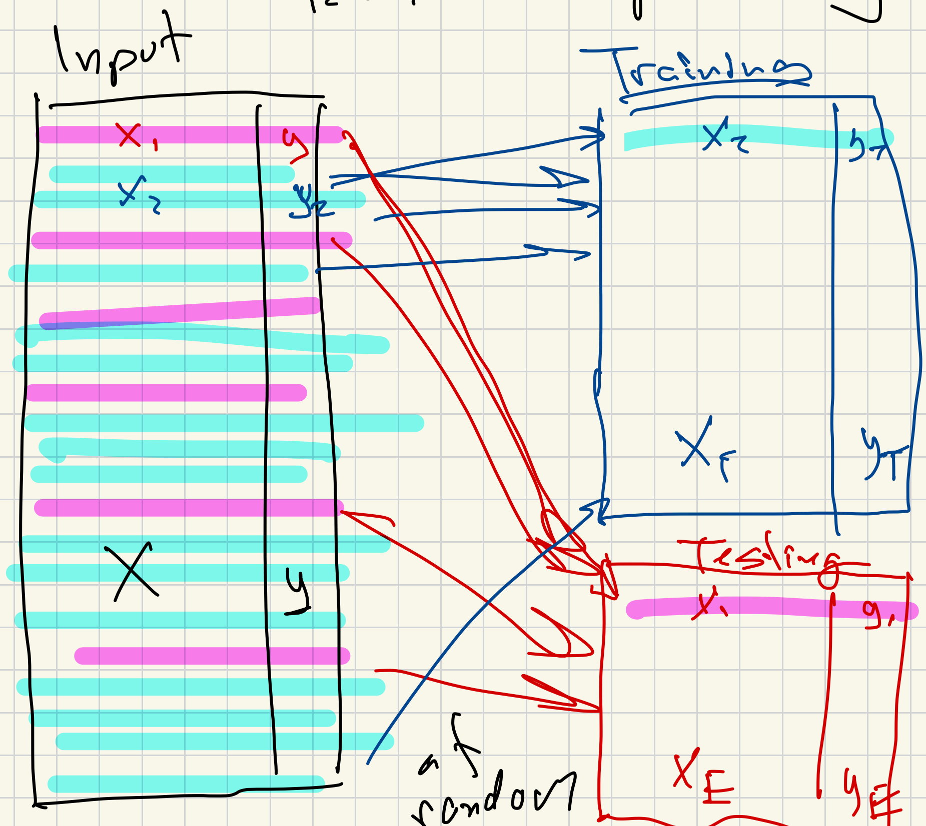
2. consistency: if $\|x - x'\|$ small $\Rightarrow \|M_\alpha(x) - M_\alpha(y')\|$ small.

3. not fooled by outliers, noise

Assume $(X, y) \underset{iid}{\sim} \mathcal{U}$ (unknown dist)

(X_T, y_T), (X_E, y_E) both iid $\sim \mathcal{U}$

Data Splitting for C-V



$(X_T, y_T) \sim_{iid} \mathcal{U}$ $\quad$ $(X_E, y_E) \sim_{iid} \mathcal{U}$

① Train

$$\alpha_T = (X_T^T X_T)^{-1} X_T^T y_T$$

M_α ← only from train data

② Testing

$$SSE((X_E, y_E), M_\alpha)$$

$$RMSE((X_E, y_E), M_\alpha) = \sqrt{\frac{1}{n_E}\sum_{i}^{E} \left(M_\alpha(x_i) - y_i\right)^2}$$

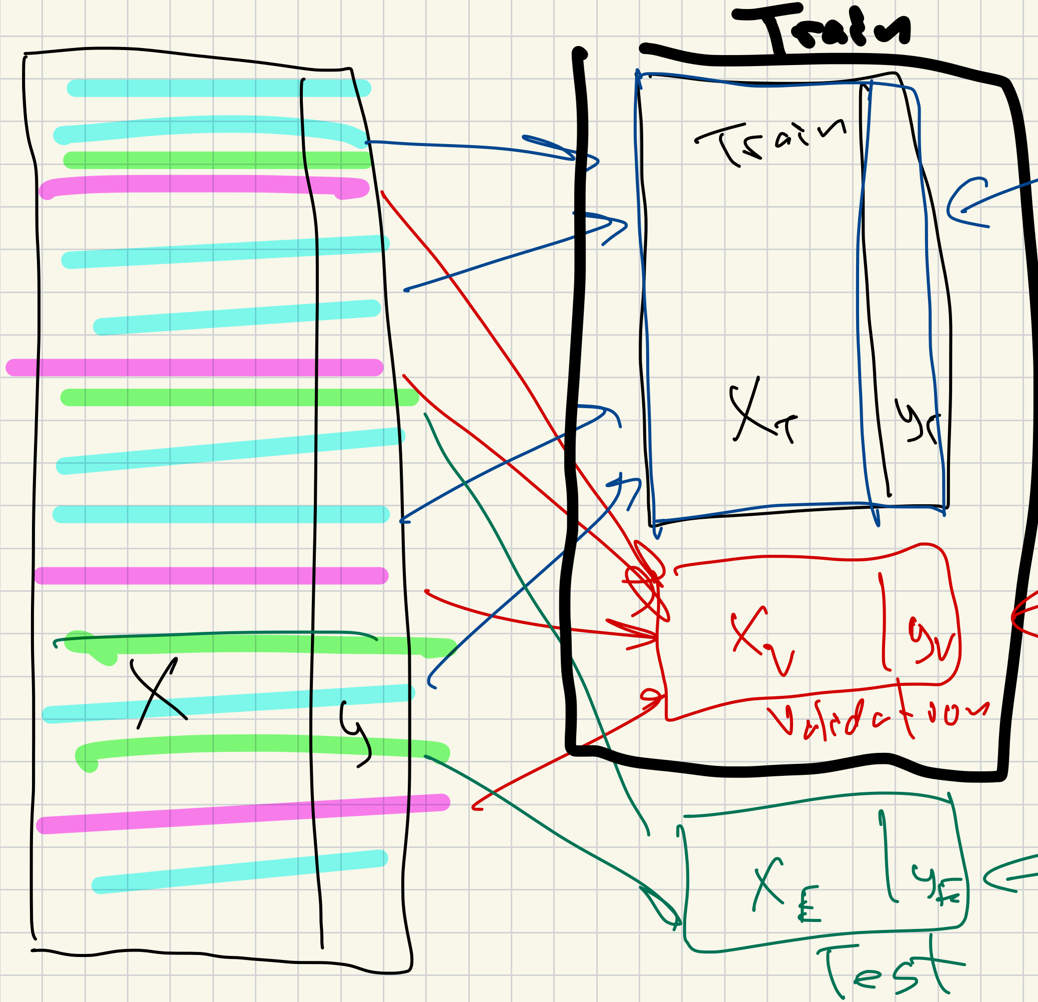
Evaluate

① How well do we expect to do on new data RMSE

generalize

② How do we choose best hyperparameter p = degree polynomial

Do not do both at the same time!



optimise core parameters

$$\alpha^* = \mathrm{Alg}(X_T, y_T)$$

Validate

- choose model
- hyperparameter

Test

evaluation

generalization

How large to make Train, Valid, Test?

Train / Test: 80/20 or 90/10

66/33

Test evaluation 1-dim quantities $\simeq 10{,}000$

Training building high-dim model $\alpha \in \mathbb{R}^{d+1}$

Tr / Val / Te $\rightarrow$ 80/10/10

60/20/20