RAPID
Transforming Process Industries

UNIVERSITY OF DELAWARE

Data Driven Modeling 1

Data processing, loss function and the problem of overfitting

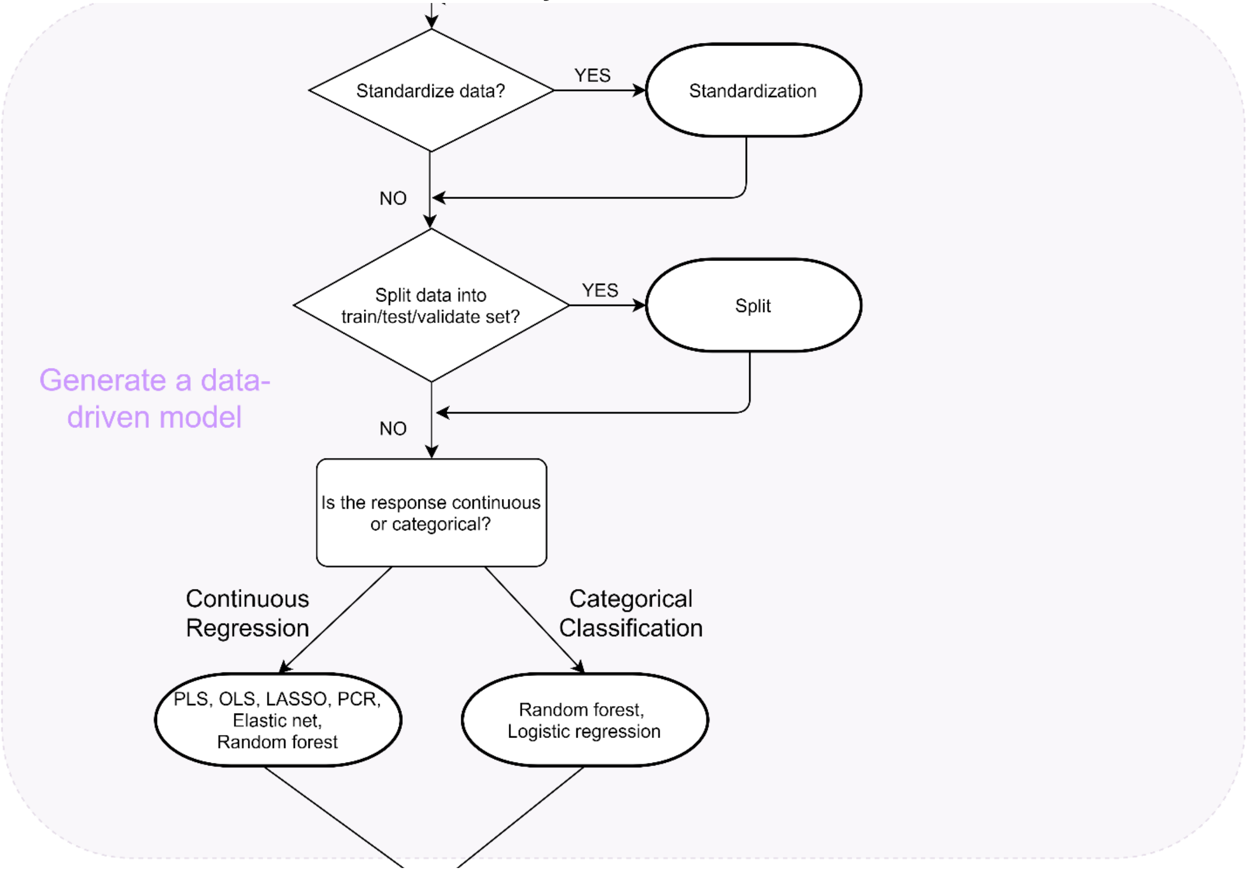
Rapid Advancement in Process Intensification Deployment

AIChE
The Global Home of Chemical Engineers

Data Driven Modeling Flowchart

Features - $\boldsymbol{X}$

Response - y



Model - $h_\theta(X)$

The problem: design the features $\boldsymbol{X}$ and choose the best set of parameters $\boldsymbol{\theta}$

Notations Recap

- Number of features – n
- Number of data points – m
- Target/response/dependent variable - y
- Features/descriptors/independent variables – X $(x_0, \ldots, x_{\mathrm{j}}, \ldots, x_n)$
- Feature value of the i^{th} data point $x_j^{(i)}$
- The parameters - $\theta_0, \ldots, \theta_j, \ldots, \theta_n$
- The model (e.g. polynomial regression)-

$$h_\theta(X) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n$$

The problem: design the features $\boldsymbol{X}$ and choose the best set of parameters $\boldsymbol{\theta}$

Preprocessing Data

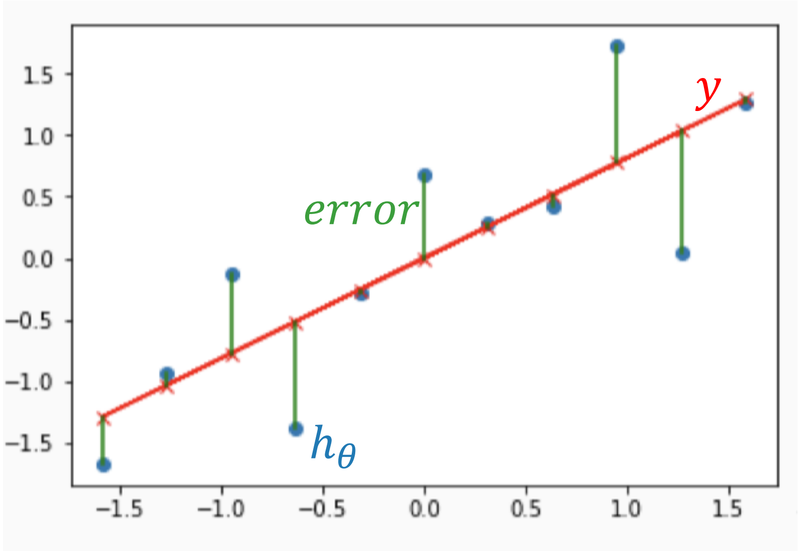
- Normalization – rescale data into the range of [0,1]
- Standardization – rescale data to have a mean of 0 and standard deviation of 1

When to preprocess the data?

- Features are of different scales
- Depends on the sensitivity of the model
- Experiment with and without preprocessing data

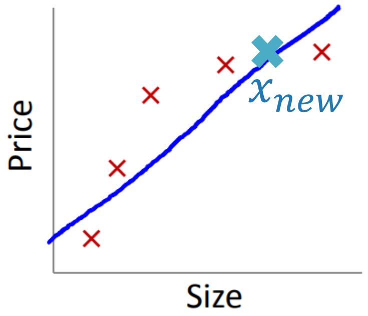
Loss Function in regression – $J(\theta)$

- Loss function
  - the difference between prediction and real values (errors)
- Mean absolute error (MAE) $MAE = \frac{1}{n}\sum_{j=1}^{n}|y_i - h_\theta(x)|$
- Root mean square error (RMSE) $RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_i - h_\theta(x))^2}$
- MAE <= RMSE
- RMSE penalize large errors (outliers) more



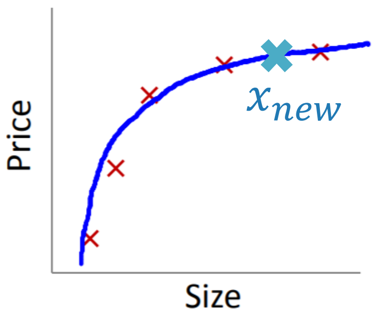
Example: polynomial regression (housing price)

- Housing price \$ (y) is a function of size ft^2(x)



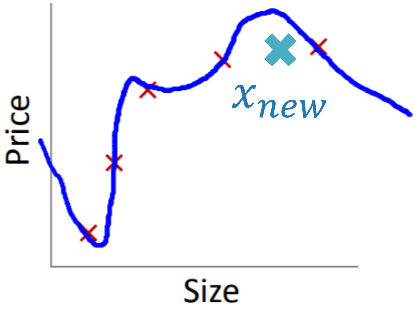
$\theta_0 + \theta_1 x$

Underfit
High bias



$\theta_0 + \theta_1 x + \theta_2 x^2$

Just right



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Overfit
High variance

- Overfitting: if we have too many features, the model fit the training set very well, but fail to generalize to new examples (predict prices on new examples)

How to prevent overfitting?

1) Reduce the number of features:

- Manually select which features to keep.
- Use a model selection algorithm (studied later in the course).

2) Regularization

- Keep all the features but reduce the magnitude of parameters θ_j.
- Regularization works well when we have a lot of slightly useful features.

Regularization

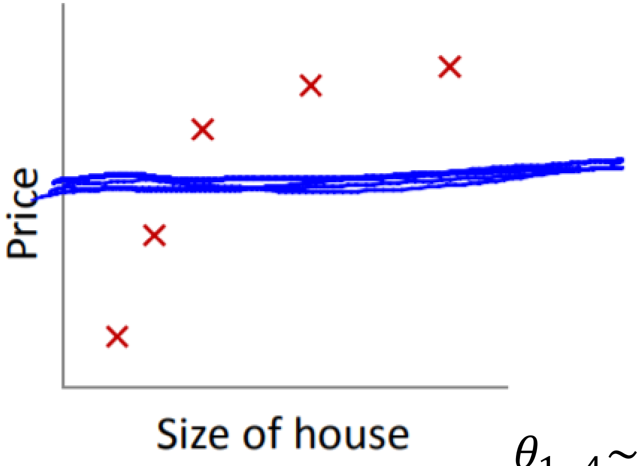
- Small values for parameters $\theta_0, \dots, \theta_j, \dots, \theta_n$
- The updated loss function

$$J_{new}(\theta) = J_{old}(\theta) + \lambda \sum_{j=1}^{n} \theta_j^2$$

- Regularization parameter - λ; the larger λ; the smaller the θs

$\lambda = 10^{10}$

(underfit)

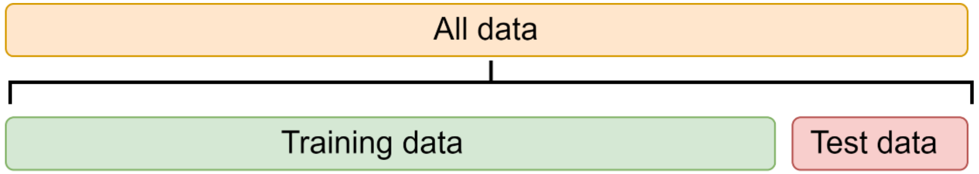


$\theta_{1-4} \sim 0$

$\theta_0 + \cancel{\theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4}$

Training Set – Model Building; Test set – Model Evaluation

Question 1: how do we know the model can be generalized to a new dataset?

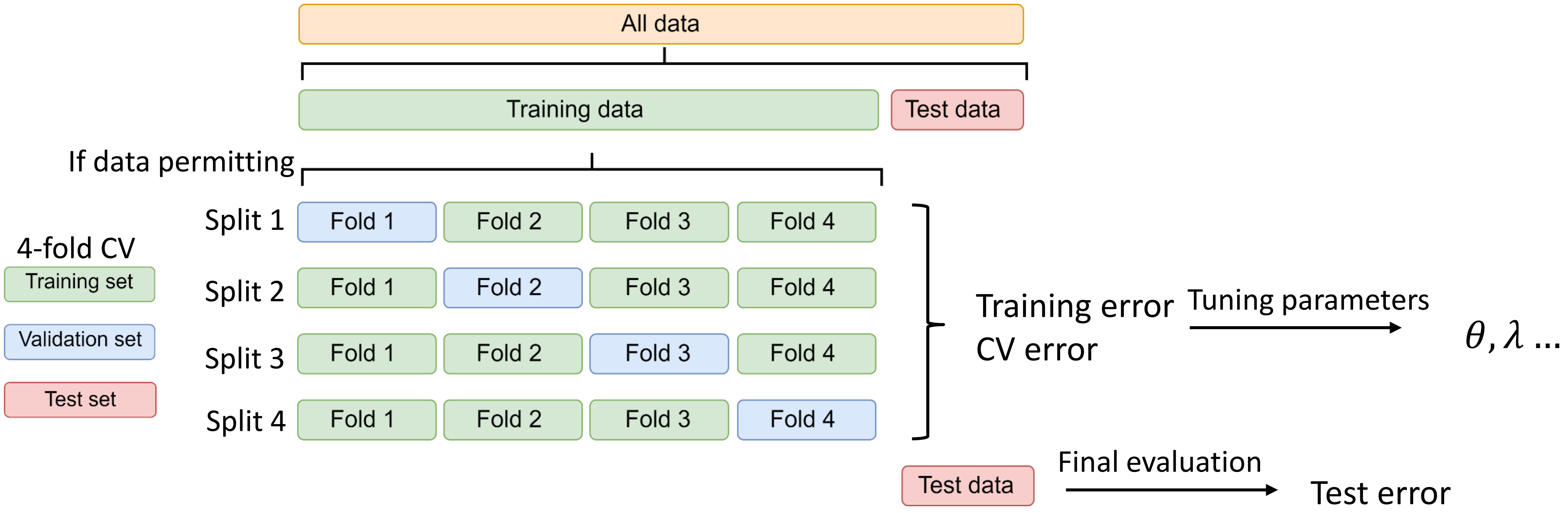


- Test set: a subset of the data not used during the training but used to evaluate the final model after the training, selected randomly
- Typical train/test split 90/10, 80/20, 75/25
- General rule: test/training set distributions should be similar to the target distribution

Question 2: how to pick the best model during training?

Cross Validation (CV) – Model Checking and Selection

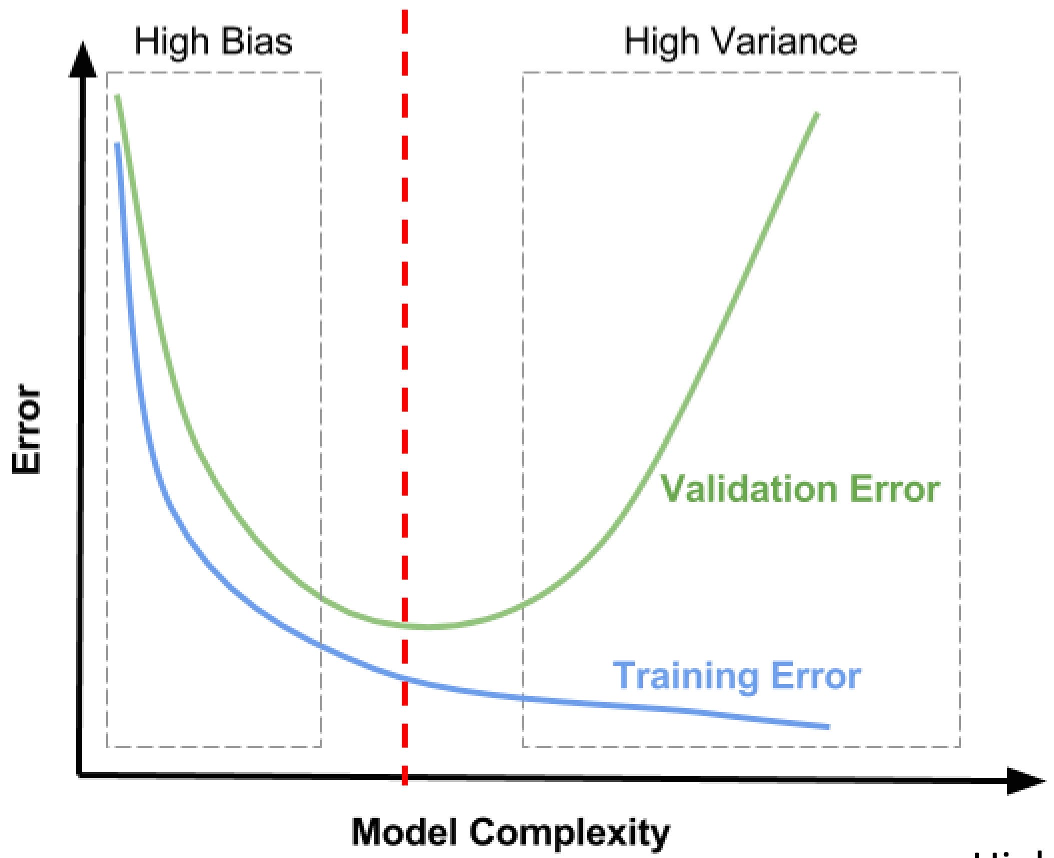
- K-fold CV
  - Typical k values = 3, 4, 5, 10



- LOOCV (leave one out cross validation) k = n (dataset size)
  - Use when limited data is available (<100)

Bias–variance tradeoff

The optimal model



High bias (underfit)

- High training error
- Validation error is similar in magnitude to the training error

High variance (overfit)

- Low training error
- Very high validation error

Improving model performance

To fix high variance (complex model, less data)

- Get more data points
- Try smaller set of features
- Try increasing regularization λ

To fix high bias (simple model, sufficient data)

- Try getting additional features
- Try decreasing regularization λ

Practice Time!