MSE491: Application of Machine Learning in Mechatronic Systems

Cross-Validation, Overfitting and Complexity

Mohammad Narimani, *Ph.D., P.Eng.*
Lecturer
School of Mechatronic Systems Engineering
Simon Fraser University

SFU SIMON FRASER UNIVERSITY SURREY

Outline

- Regression and Gradient Descent ✓

- Cross-Validation, Overfitting and Complexity, training set, validation set, test set

Cross-Validation

- You can't fit the model to your training data and hope it would accurately work for real new data. Therefore, there is always a need to validate ML model.

- To measure generalization accuracy when performing a supervised ML experiment, hold out a part of the available data as a test set.

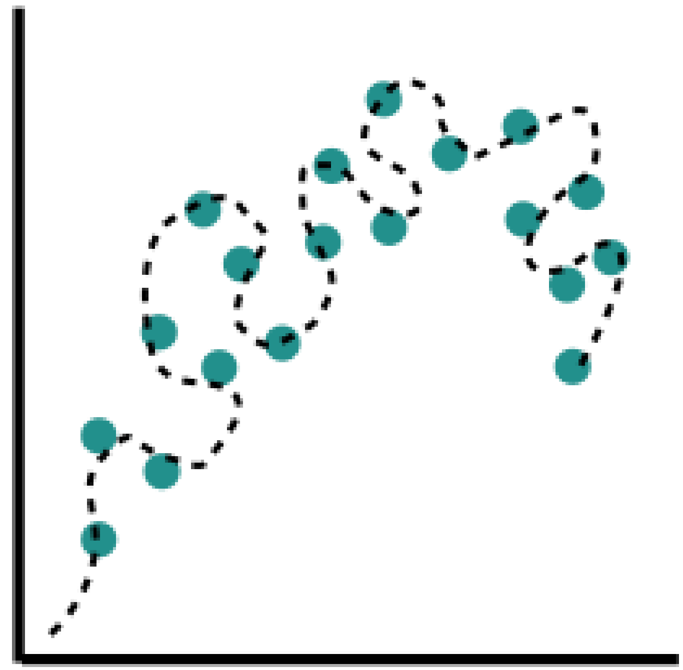
- We split the data as:
  - Training set (50-75%)
  - Test set (25-50%)

Simple Linear Regression

- In the previous example

Overfitting and Underfitting

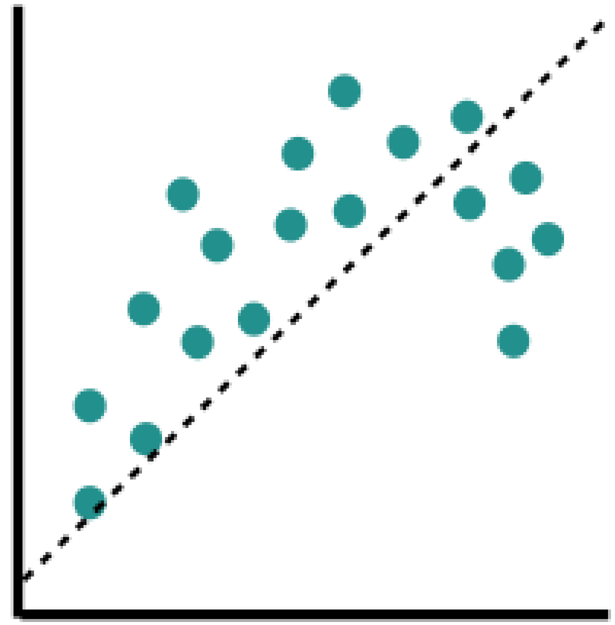
- Overfitting (high variance): happens when a trained model learns the details based on the noise in the training data which negatively impacts the performance of the model on new data.
  - It may happen when the number of features is large, and it may lead a serious problem if the number of training data is small.



$$\boxed{h_\theta = \theta_0 + \theta_1 x + \theta_1 x^2 + \cdots + \theta_9 x^9}$$

Overfitting and Underfitting

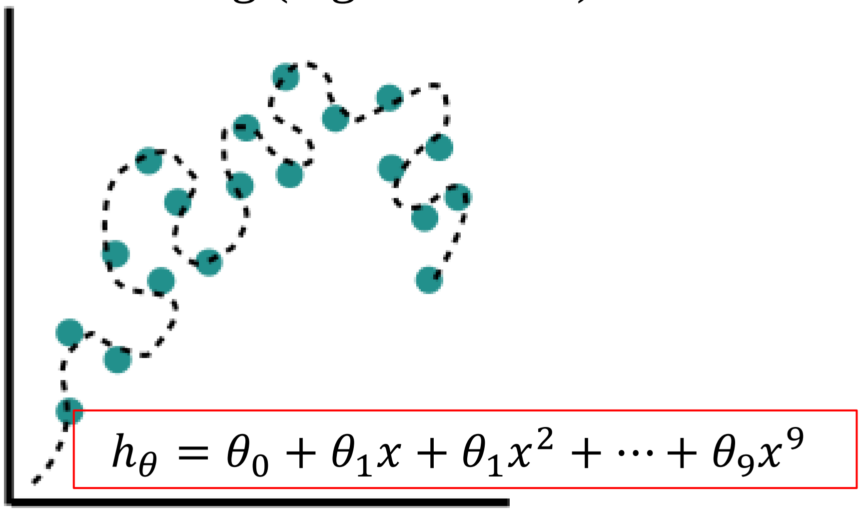
- Underfitting (high bias): happens when a trained model can neither re-produce the training data nor generalize to new data



$$\boxed{h_\theta = \theta_0 + \theta_1 x}$$

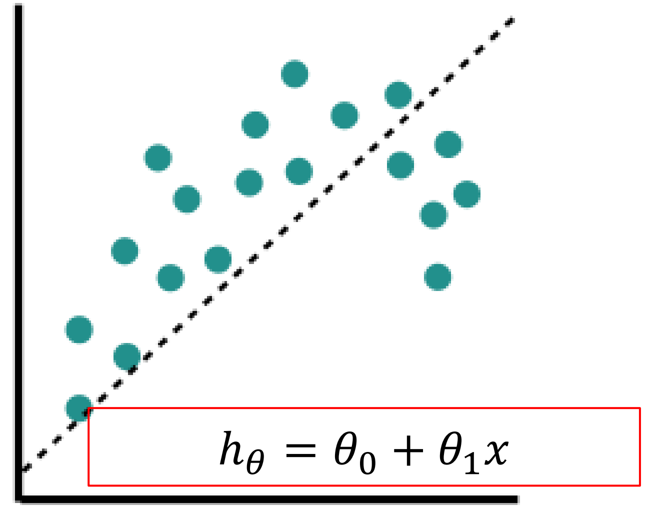
Overfitting and Underfitting

Overfitting (high variance):



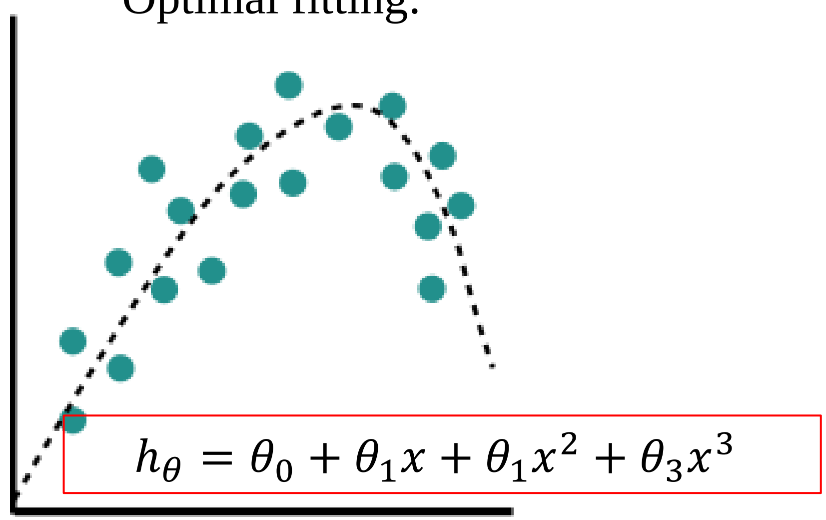
$$h_\theta = \theta_0 + \theta_1 x + \theta_1 x^2 + \cdots + \theta_9 x^9$$

Underfitting (high bias):



$$h_\theta = \theta_0 + \theta_1 x$$

Optimal fitting:



$$h_\theta = \theta_0 + \theta_1 x + \theta_1 x^2 + \theta_3 x^3$$

How to solve overfitting

- Increase the number of training data (real or augmentation data)

- Reduce the number of features

- K-fold cross-validation

- Regularization

How to solve overfitting: Regularization

- **Regularization**: is a technique to solve overfitting problem (complexity of a ML model) without eliminating features.

- This complexity is minimized by penalizing the cost function for having large magnitudes of parameters θ_i's.

- This approach works well when the model has a lot of features with low contribution in predicting y.

How to solve overfitting

- **Regularization**:

Cost function:

$$J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2\right.$$

$$\min_{\theta_0, \theta_1, \dots, \theta_n} J(\theta_0, \theta_1, \dots, \theta_n)$$

λ is regularization parameter

How to solve overfitting

- **Intuition of Regularization**:

Consider:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{i=1}^{n} \theta_j^2\right]$$

$$\min_\theta J(\theta)$$

- If $\lambda >> \Rightarrow \theta_i's \approx 0 \Rightarrow h_\theta(x) = \theta_0$

- Therefore, choosing λ automatically, is another part of ML algorithm

Regularization

- **How to tune the value of λ?**

The brief answer is Cross-validation with different values of λ!

Step 1: λ is set and the model is trained using training set. Then the cost function is evaluated on the test set and the result is saved.

Step 2: The value of λ is updated and step 1 repeated.

Step 3: The results of cost function vs λ is plotted and the best value of λ is selected.

Regularized Linear Regression

Knowing

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{i=1}^{n}\theta_j^2\right]$$

and using gradient descent to minimize $J(\theta)$:

$$\begin{cases} \theta_0 = \theta_0 - \alpha.\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_0^{(i)} \\ \theta_j = \theta_j - \alpha.\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)} + \alpha\frac{\lambda}{m}\theta_j \quad j = 1,2,3\ldots,n \end{cases}$$

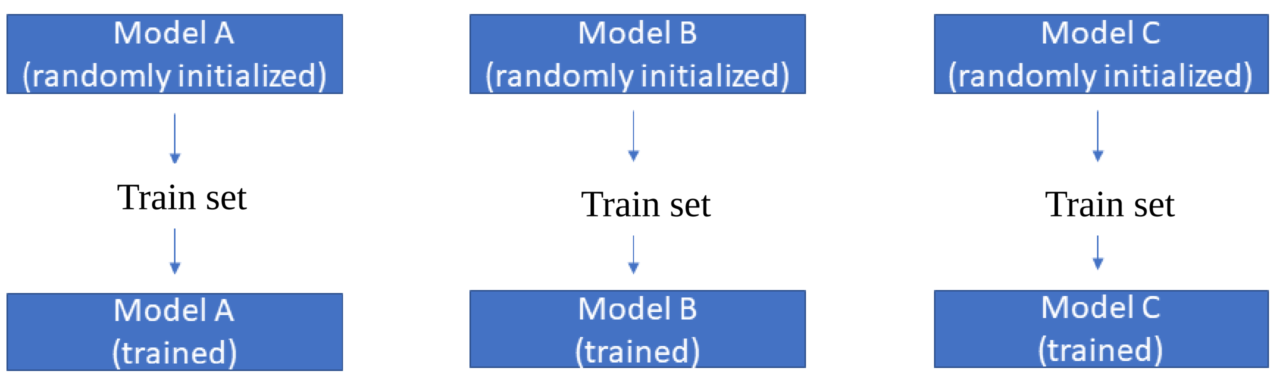
Regularization

- Regularization for Normal Equation:

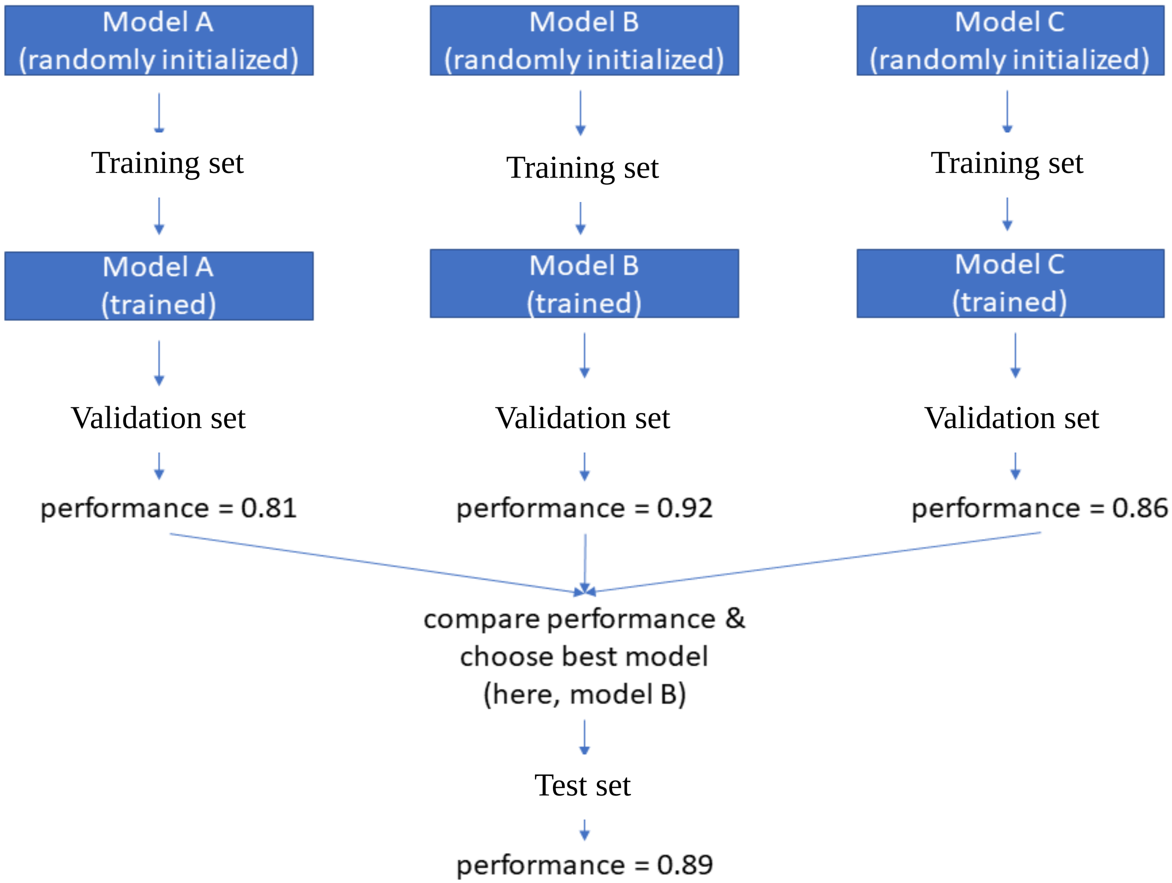
Cross-Validation with validation set

- Model A, B, and C can be different architectures (like NN, SVM or logistic regression), or the same model with different hyperparameters (linear regression with different regularization parameter)

- We split the data as:
  - Training set (60-80%)
  - Validation set (10-20%)
  - Test set (10-20%)



Cross-Validation with validation set



- Note: if there aren't many hyperparameters you can shift some of data into the test set.

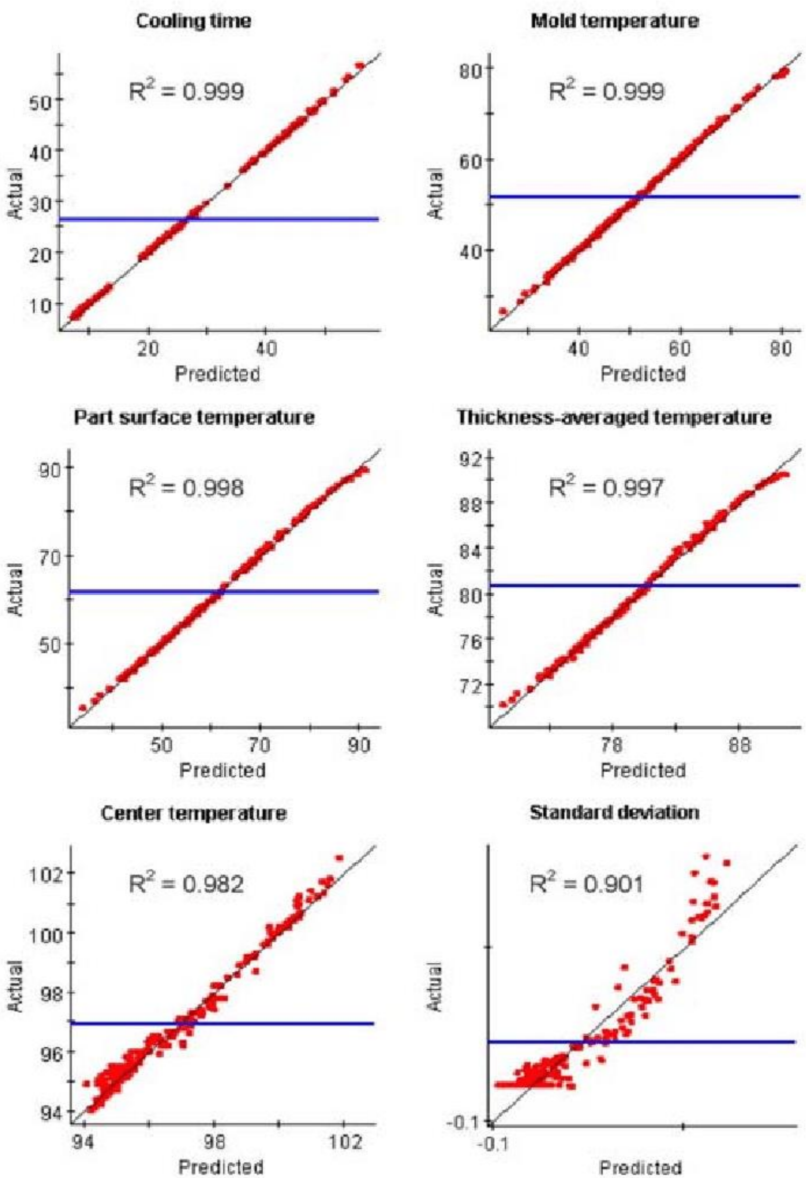
R2: Coefficient of Determination

- R-squared ($R2$ or R^2): is a statistical measure of how close the data are to the fitted regression line.

$$R^2 = 1 - \frac{SS_r}{SS_t}$$

$$SS_r = \sum_i \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$SS_t = \sum_i \left(y_{mean} - y^{(i)}\right)^2$$



Pearson's Correlation Coefficient r (Correlation Coefficient)

- **Pearson's Correlation Coefficient is** used to measure how strong a relationship is between two variables.

- It can be an important tool for feature engineering in building machine learning models.
  - For example: shoe size is not a useful predictor for salary!!

$$r = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2 \ \sum_{i=1}^{m}(y_i - \overline{y})^2}}, \qquad -1 \le r \le 1$$

Data Visualization

- Scatter Diagram:When an investigator collects two series of observations and wishes to see whether there is a relationship between them, a scatter diagram is the first and simplest tool.