Regularization and Model Tuning

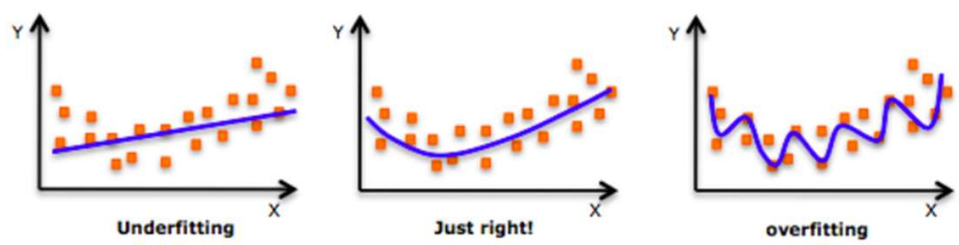
Week 5

Mohammad Esmalifalak

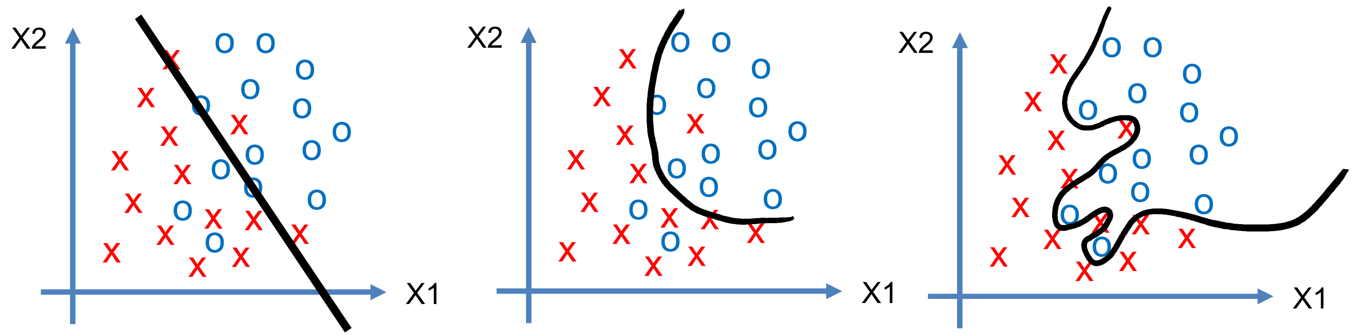
McMaster University
CENTRE FOR CONTINUING EDUCATION

Model fitting issues

- One of the goals of machine learning is generalizability.
- If your model works only on your training data, the model is effectively useless.
- Depending on the training process your model might be
  - Underfitted
  - Overfitted
  - Well fitted



Over Fitting vs Under Fitting



Model performs well on the training data but generalizes poorly on test → Over fitting

Model performs poorly on both training and test → Under fitting

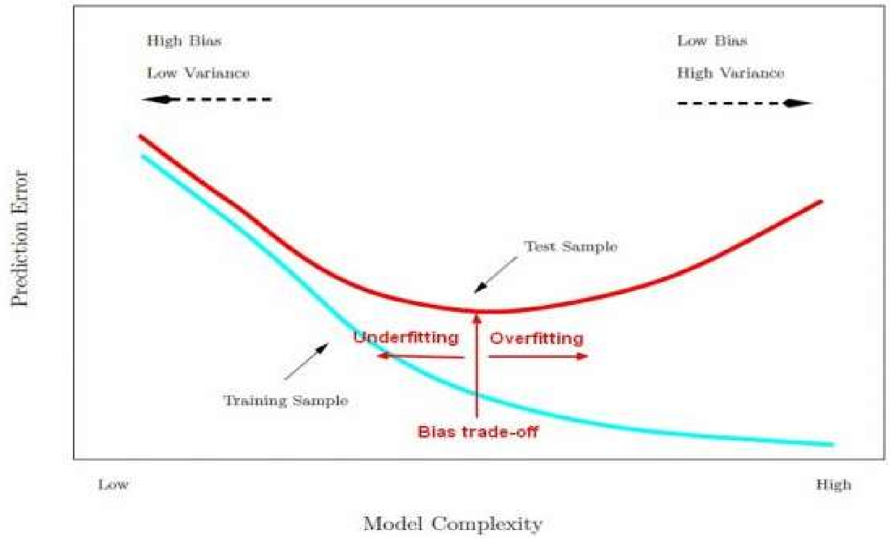
Overfitting vs Underfitting

Over-fitting

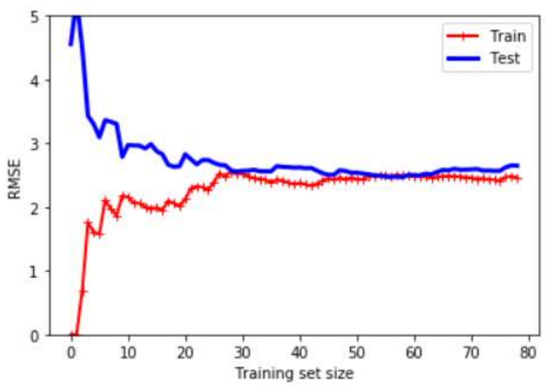
- Generally occurs when a model is excessively complex.
- Sensitive to noise and other parameters which were not modeled during the training.

Under-fitting

- Occurs when a model is over simplified
- Number of training samples are not enough



Over Fitting vs Under Fitting

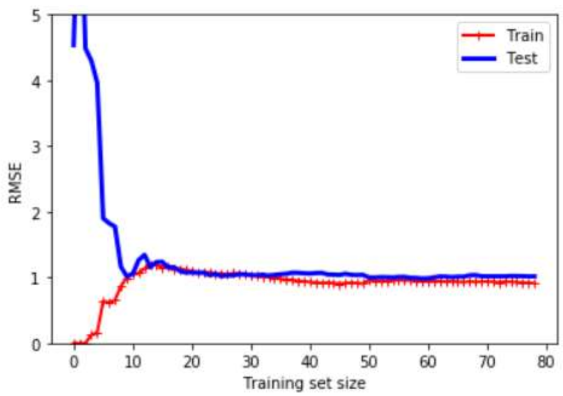


1- Low training RMSE and high testing RMS at beginning

2- Training data goes up until it reaches a **plateau**

3- Both curves have reached a plateau; they are close and fairly high (sign of under-fitting).

4- If model is under-fitting, adding more training data will not help.

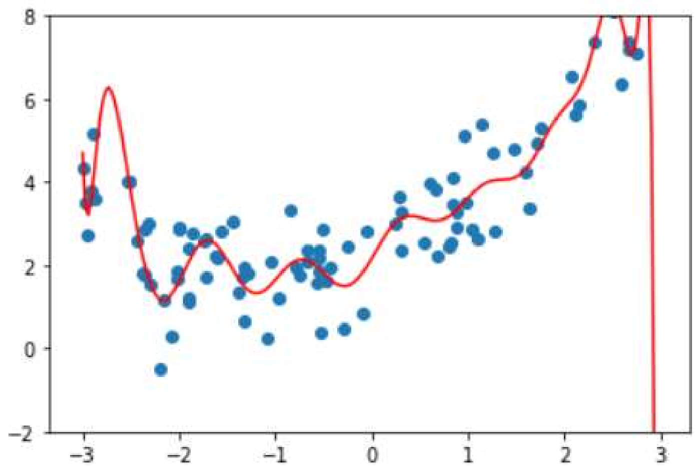


1- Low training RMSE and high testing RMS at beginning

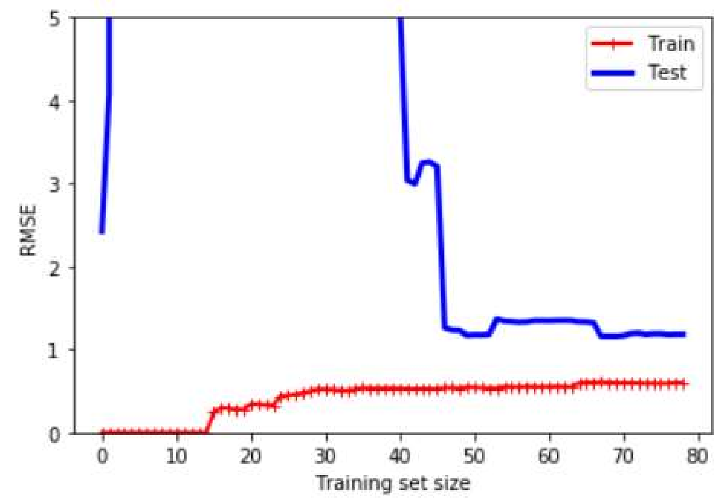
2- Training data goes up until it reaches a **plateau**

3- Both curves have reached a plateau; they are close and fairly low (sign of fitting right).

Over Fitting vs Under Fitting



Higher complexity = better fit to training data
(But less generalizable)



Low training RMSE but high testing RMS at steady state (sign of over-fitting).

One way to improve an overfitting model is to feed it more training data until the validation error reaches the training error.

How to avoid overfitting?

- Have a very large training data
- Reduce model complexity
- Regularization

Regularization

- **Regularization (in General)**

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

penalty='l2'

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

penalty='l1'

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

L2 regularization	L1 regularization
Computational efficient due to having analytical solutions	Computational inefficient on non-sparse cases
Non-sparse outputs	Sparse outputs
No feature selection	Built-in feature selection

Regularization

- **Regularization (in linear regression)**

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$$

Ridge Regression

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

Lasso Regression

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+\alpha\sum_{i=1}^{n}|\theta_i|$$

Elastic Net

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+r\alpha\sum_{i=1}^{n}|\theta_i| + \frac{(1-r)\alpha}{2}\sum_{i=1}^{n}\theta_i^2$$

Regularization

- **Gradient Descent**

No Regularization:

Repeat, {

$$\theta_k := \theta_k - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_k^{(i)}$$

}

With Regularization:

Repeat, {

$$\theta_k := \theta_k\left(\mathbf{1} - \boldsymbol{\alpha}\frac{\boldsymbol{\lambda}}{\boldsymbol{m}}\right) - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_k^{(i)}$$

}

Regularization

- **Linear Regression**

No Regularization:

$$\theta = (X^T X)^{-1} X^T y$$

With Regularization:

(Ridge Regression)

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}\right)^{-1} X^T y$$

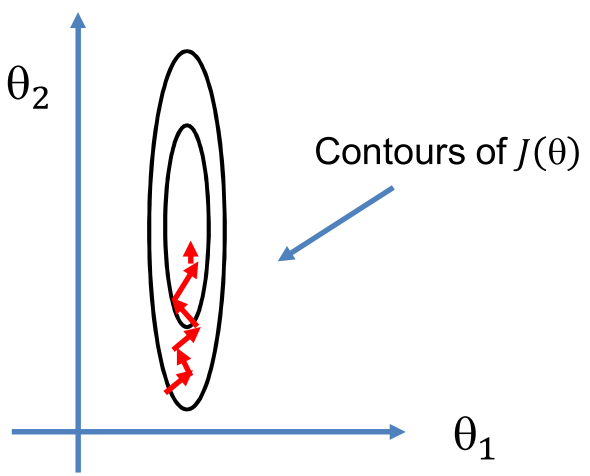
1

Feature Normalization

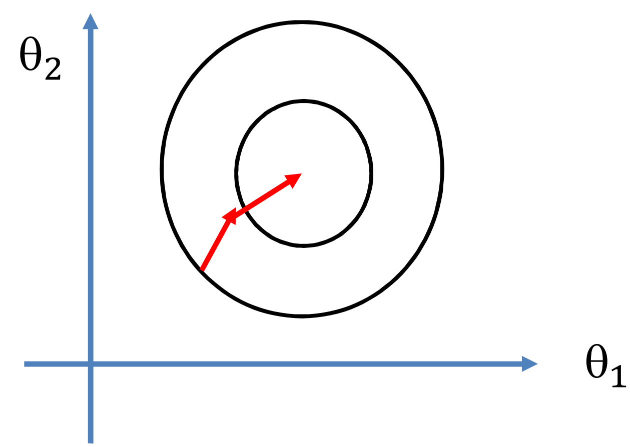
x1= size of house: (1000-3000)
x2= # of bedrooms: (0-5)

x1= (x1-mean)/range: [-0.5,0.5]
x2= # (x1-mean)/range: [-0.5,0.5]

$$\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$



Converging to global minimum is harder



Converging to global minimum is easier

McMaster University
CENTRE FOR CONTINUING EDUCATION