ME 491 Lecture 9.

Ch.4 Regression and Model Selection

- Most basic regression techniques is curve fitting. With Polynomial and exponential fitting from solving the linear equation

$$Ax = b$$

- When the model is not prescribed, often, our goal is to select the best model, so we will normally have

  - over-determined optimization problem

$$\underset{x}{\operatorname{argmin}} \left(\|Ax - b\|_2 + \lambda g(x) \right)$$

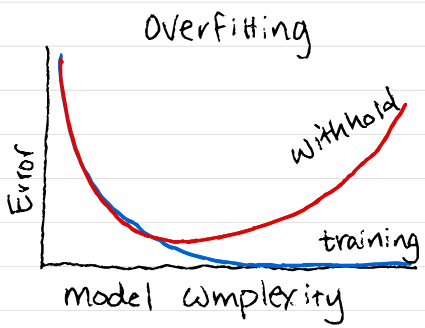
  - under-determined optimization problem

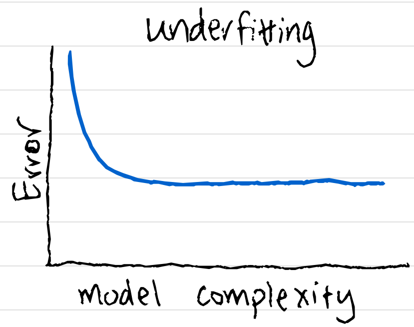
$$\underset{x}{\operatorname{argmin}}\, g(x) \quad \text{subject to} \quad \|Ax - b\|_2 \leq \epsilon$$

$g(x)$: given penalization

- Cross-validation

  if a proposed model has overfit or under fit the data.





model is constructed on training and validation data, and tested on withhold.

Ch 4.1 Classic Curve Fitting

- Regression: attempts to estimate the relationship among variables.

$$Y = f(X, \beta)$$

X: independent variables, Y: dependent variables
β: unknown parameters.

- Least-Squares Fitting Methods

use a simple function to describe a trend by minimizing the sum-square error between the selected function $f(\cdot)$ and its fit to data.

consider a set of n data points
$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$
we want to approximate the line using

$$f(x) = \beta_1 x + \beta_2$$

This gives us the linear regression model

$$Y = f(X, \beta) = \beta_1 X + \beta_2$$

the error at each point is: $f(x_k) = y_k + E_k$

Many different error metrics can be used to solve for $f(x)$.
three standard ones are

1. maximum error ℓ_∞: $E_\infty(f) = \max\limits_{1<k<n} |f(x_k) - y_k|$

2. mean absolute error ℓ_1: $E_1(f) = \frac{1}{n}\sum_{k=1}^{n} |f(x_k) - y_k|$

3. least-square error ℓ_2: $E_2(f) = \left(\frac{1}{n}\sum_{k=1}^{n} |f(x_k) - y_k|^2\right)^{1/2}$

We can generalized to the l_p norm

$$E_p(f) = \left(\frac{1}{n} \sum_{k=1}^{n} |f(x_k) - y_k|^p \right)^{1/p}$$

normally, the difference to different-norm fit is small, unless there are outliers in data.

- Least-Squares Line

Least square fitting is preferred as it is inexpensive and can be solved analytically.

We want to fit to the curve: $f(x) = \beta_1 x + \beta_2$

the error is: $E_2(f) = \sum_{k=1}^{n} |f(x_k) - y_k|^2 = \sum_{k=1}^{n} (\beta_1 x_k + \beta_2 - y_k)^2$

To minimize E_2, we want to solve

$$\frac{\partial E_2}{\beta_1} = 0 : \quad \sum_{k=1}^{n} 2(\beta_1 x_k + \beta_2 - y_k) x_k = 0$$

$$\frac{\partial E_2}{\beta_2} = 0 : \quad \sum_{k=1}^{n} 2(\beta_1 x_k + \beta_2 - y_k) = 0$$

Rearrange the two equations, we get

$$\cancel{2}\beta_1 \sum_{k=1}^{n} x_k^2 + \cancel{2}\beta_2 \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} \cancel{2} x_k y_k$$

$$\cancel{2}\beta_1 \sum_{k=1}^{n} x_k + \cancel{2} n \beta_2 = \sum_{k=1}^{n} \cancel{2} y_k$$

we can write them in a 2x2 system of linear equations

$$\begin{pmatrix} \sum_{k=1}^{n} x_k^2 & \sum_{k=1}^{n} x_k \\ \sum_{k=1}^{n} x_k & n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} x_k y_k \\ \sum_{k=1}^{n} y_k \end{pmatrix} \Rightarrow Ax = b$$

This method can be easily generated to higher polynomial fit.
For example: $f(x) = \beta_1 x^2 + \beta_2 x + \beta_3$

we will need to solve a 3x3 system to minimize the error $E_2(\beta_1, \beta_2, \beta_3)$ by taking:

$$\frac{\partial E_2}{\beta_1} = 0 \; ; \quad \frac{\partial E_2}{\beta_2} = 0 \; ; \quad \frac{\partial E_2}{\beta_3} = 0$$

Generally, we can find a degree-k fit by solving $(k+1)(k+1)$ linear system of equations $Ax = b$.

- Data Linearization

sometime, we need to pre-process the data to use the method at hand. consider

$$f(x) = \beta_2 \exp(\beta_1 x)$$

The error to be minimized is

$$E_2(\beta_1, \beta_2) = \sum_{k=1}^{n} (\beta_2 \exp(\beta_1 x_k) - y_k)^2$$

Applying the minimization condition, we get

$$\frac{\partial E_2}{\partial \beta_1} = 0 : \quad \sum_{k=1}^{n} 2(\beta_2 \exp(\beta_1 x_k) - y_k)\beta_2 x_k \exp(\beta_1 x_k) = 0$$

$$\frac{\partial E_2}{\partial \beta_2} = 0 : \quad \sum_{k=1}^{n} 2(\beta_2 \exp(\beta_1 x_k) - y_k)\exp(\beta_1 x_k) = 0$$

The 2x2 system we are solving is

$$\beta_2^2 \sum_{k=1}^{n} x_k \exp(2\beta_1 x_k) - \cancel{\beta_2} \sum_{k=1}^{n} x_k y_k \exp(\beta_1 x_k) = 0$$

$$\beta_2 \sum_{k=1}^{n} \exp(2\beta_1 x_k) - \sum_{k=1}^{n} y_k \exp(\beta_1 x_k) = 0$$

This system is nonlinear and might not have a solution, but in this particular case, we can linearize the system

Let $Y = \ln(y)$, $X = x$, $\beta_3 = \ln \beta_2$

Taking the ln of both side of the original equation, we get

$$\ln(y) = \beta_1 x + \ln \beta_2 \Rightarrow Y = \beta_1 X + \beta_3$$