Dis 3: HW1 review

Ziwen
Adapted from slides made by Pu Sun

Announcements

- About Discussion

  - Odd-numbered discussions will be presented during the 3-4pm TA-led discussion sections. For example, D3 will be presented this week by Ziwen on Tuesday from 3-4 and by Claudio on Thursday 3-4.
  - Even-numbered discussions will be presented during the 12-1:30 lecture time.

- HW1 Due Date: Jan 17 at 11:30 pm.
- About Final Project Teams

Ziwen will be releasing this assignment by this Thursday, and **you will have until next Thursday to finalize your groups**

Groups will be self-selected and limited to 8-9 students

libraries

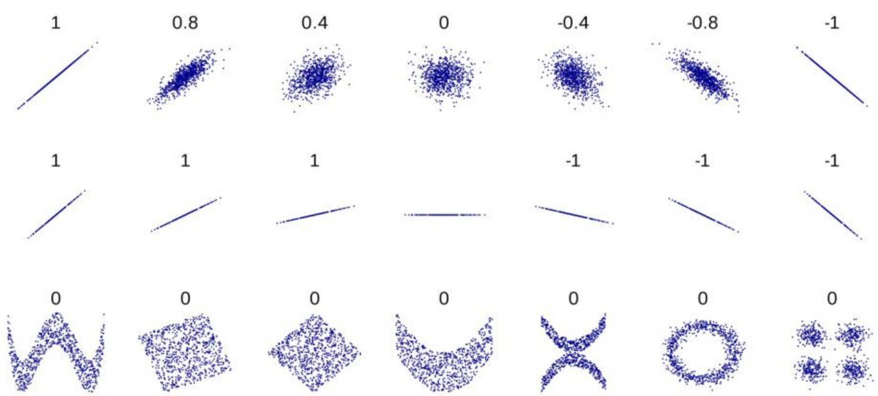
- Pandas
  - A very famous and useful data analyze and visualize tool
  - We will use it for loading data and some calculations
- Scikit-learn
  - Simple and efficient tools for predictive data analysis
- NumPy
  - The fundamental package for scientific computing with Python

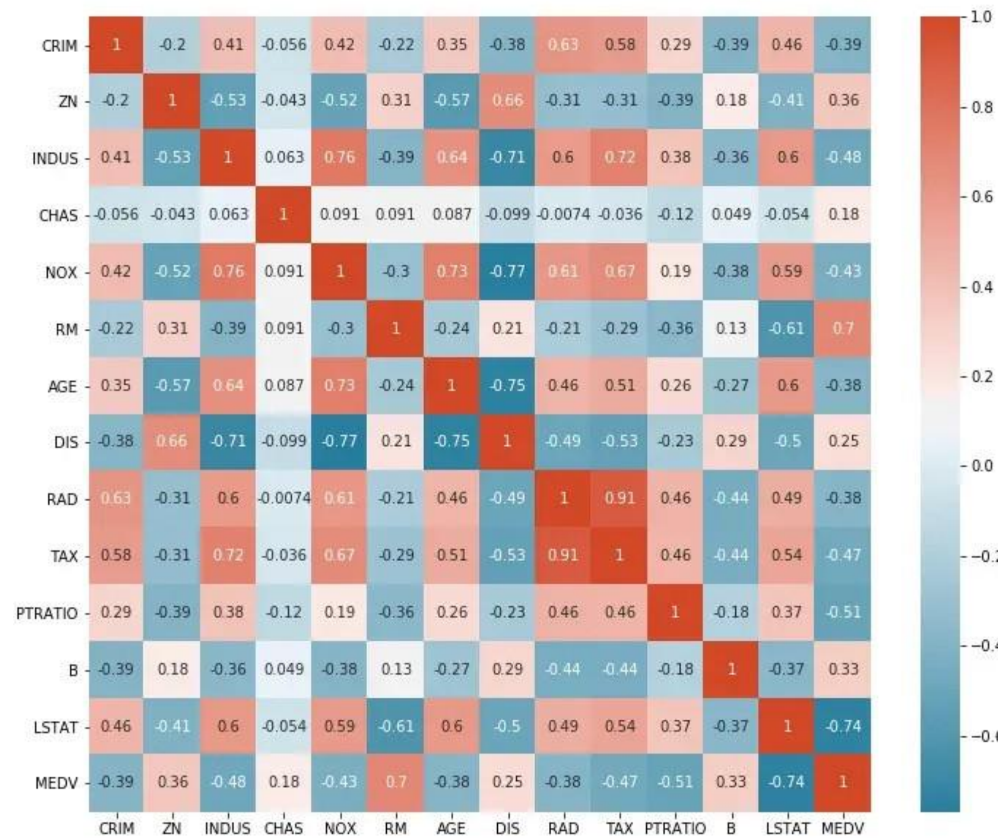
Visual tools

- Motplotlib
- Seaborn
- Pandas
- Plotly

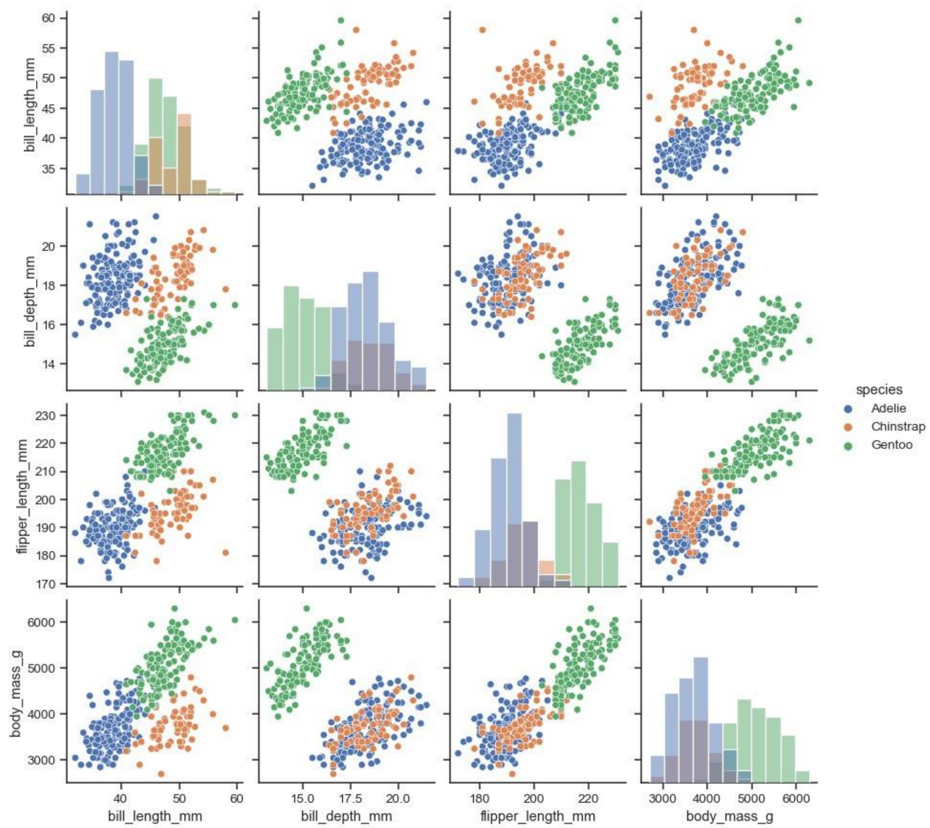
Pearson correlation

- It only calculating linear correlation.
- It will give a value [-1, 1].
- Value close to 0 means lower correlations





Pair plot (a.k.a. scatter plot matrix)

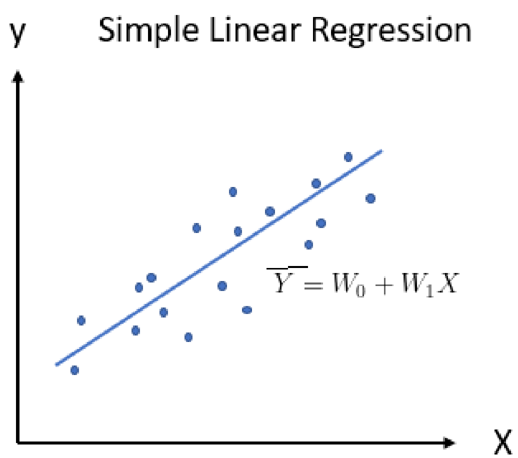


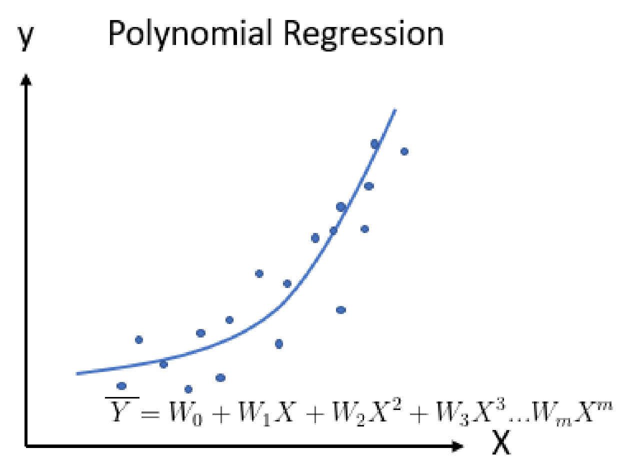
Linear and Polynomial regression

$$y_i \;=\; \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i \; (i = 1, 2, \dots, n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}.$$





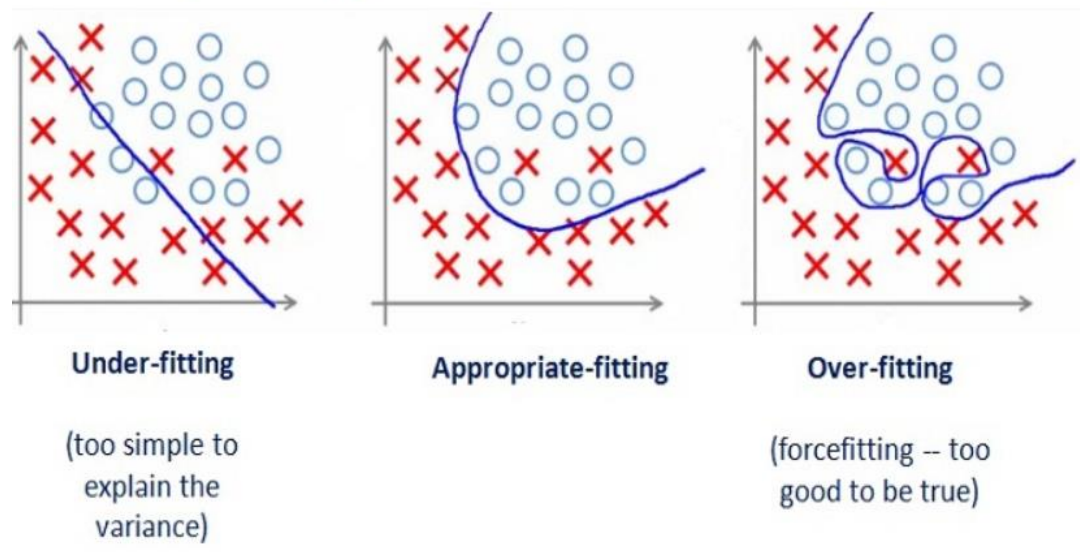
SSE and MSE and variance

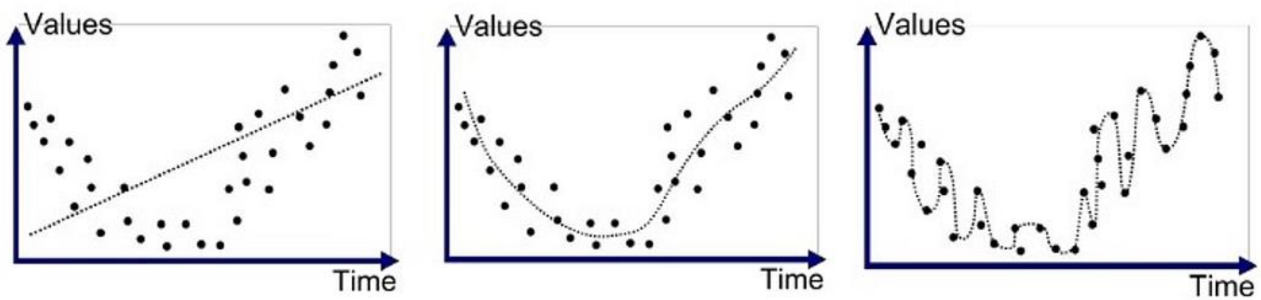
SSE: sum of squared error
MSE: mean of squared error

$$\underbrace{\sum_{i=1}^{\overset{\text{number of samples}}{\boxed{n}}}\left(\overset{\text{real value}}{\boxed{Y_i}} - \overset{\text{predicted value}}{\boxed{\hat{Y}_i}}\right)^2}_{\text{sum of the errors of all samples}}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

Overfitting and Underfitting



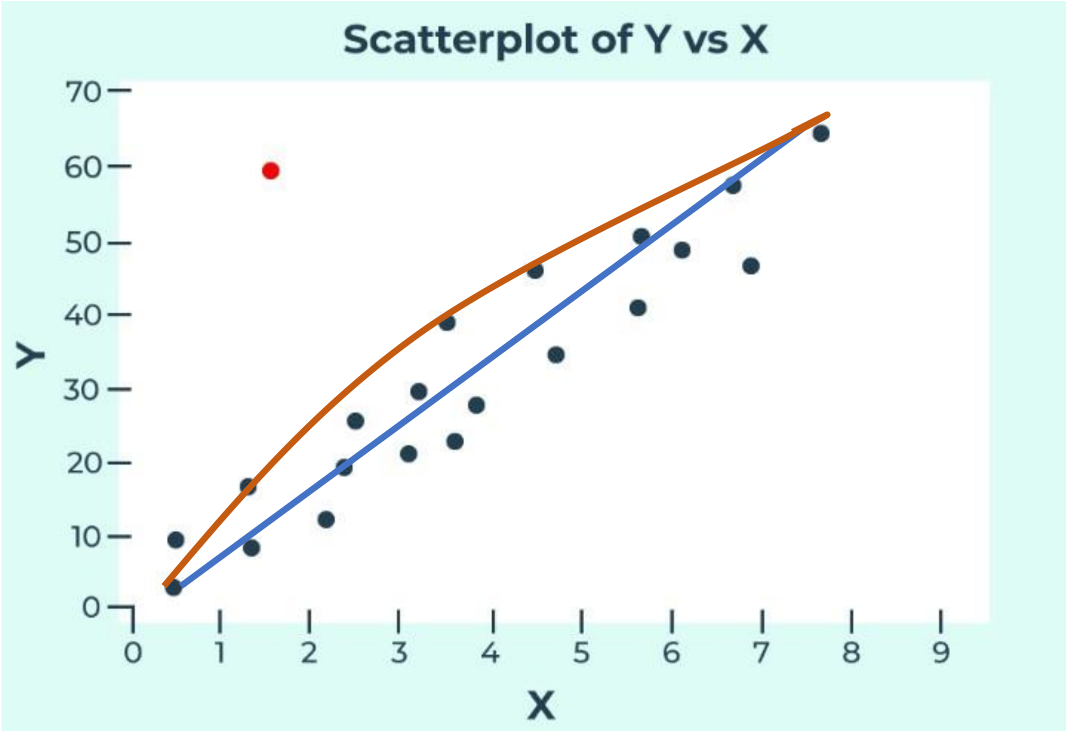


Underfitted Good Fit/Robust Overfitted

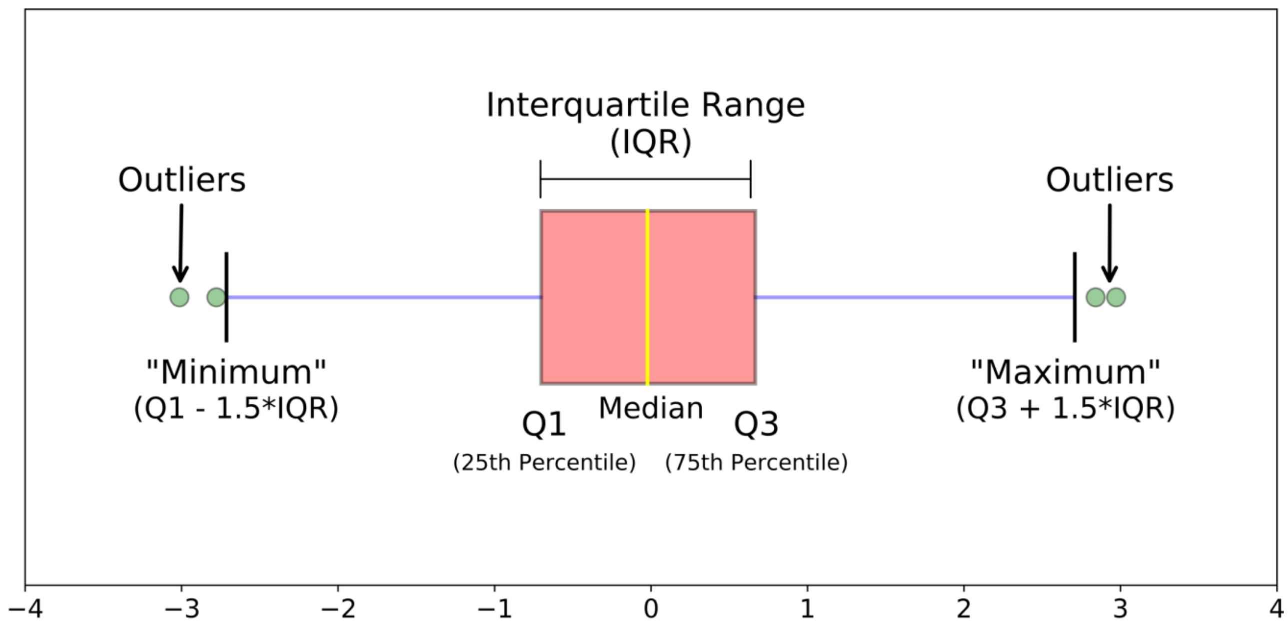
Overfitting and Underfitting

- Ideal case: low bias (in training) and low variance
- Overfitting: low bias (in training) and high variance
- Underfitting: High bias (in training and test datasets)

Outliers



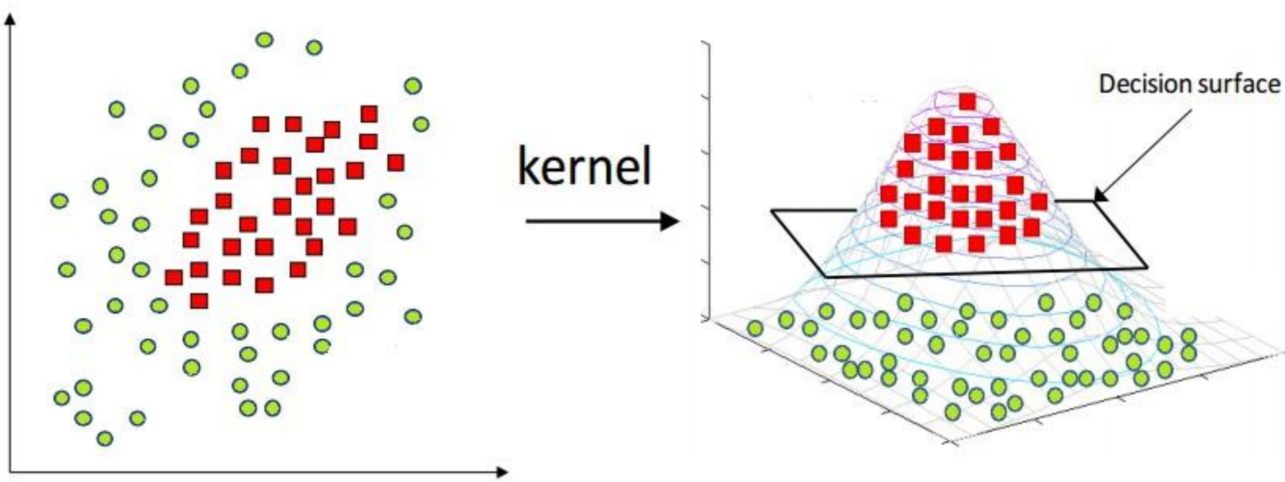
Outliers

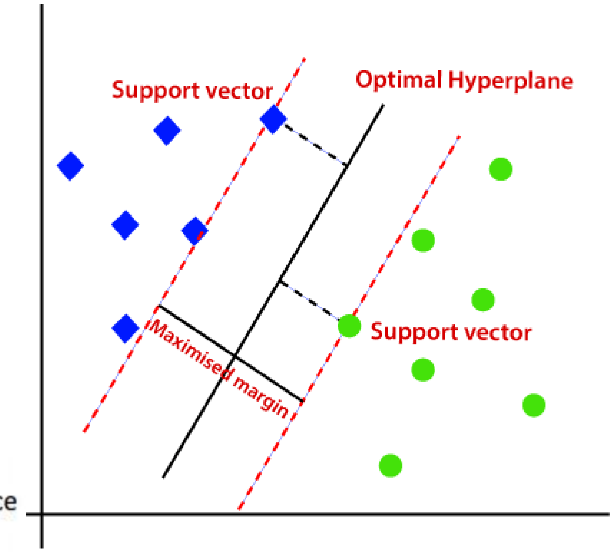


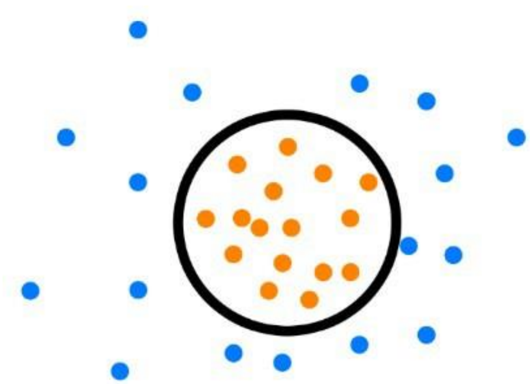
Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
"Maximum"
(Q3 + 1.5*IQR)
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

One class SVM outliers

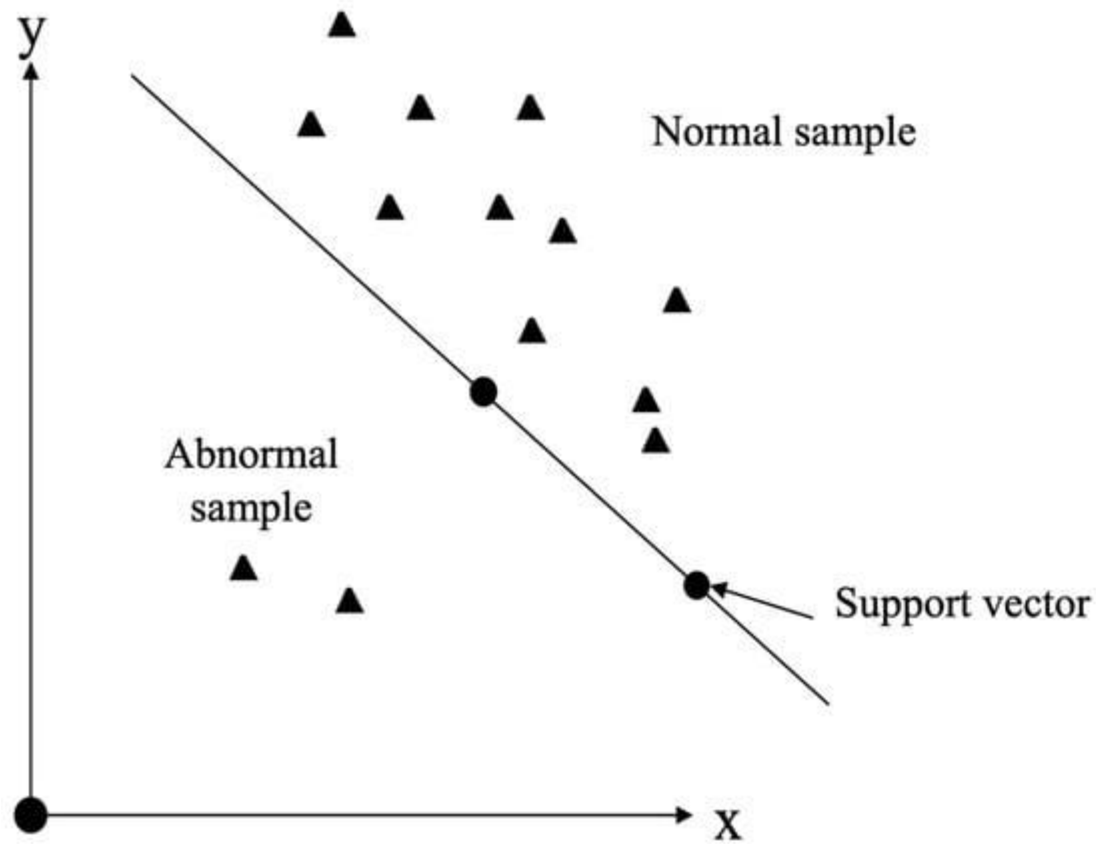
SVM:







OC SVM



OLS Example

$$\boxed{W = (X^{!}X)^{"\#}\, X^{!}\mathrm{y}}$$

$$x = \begin{bmatrix} 5 & 7 & 3 \end{bmatrix}$$

$$y = \begin{bmatrix} 2 \\ 11 \\ 8 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 5 \\ 1 & 7 \\ 1 & 3 \end{bmatrix}$$

$$X^{!} = \begin{bmatrix} 1 & 1 & 1 \\ 5 & 7 & 3 \end{bmatrix}$$

$$X^{!}X = \begin{bmatrix} 3 & 15 \\ 15 & 83 \end{bmatrix}$$

$$(X^{!}X)^{"\#} = \begin{bmatrix} \frac{83}{24} & -\frac{5}{8} \\ -\frac{5}{8} & \frac{1}{8} \end{bmatrix}$$

$$(X^{!}\,X)^{"\#}X^{!} = \begin{bmatrix} \frac{1}{3} & -\frac{11}{12} & \frac{19}{12} \\ 0 & \frac{1}{4} & -\frac{1}{4} \end{bmatrix}$$

$$\boxed{W = \begin{bmatrix} \frac{39}{12} & \frac{3}{4} \end{bmatrix}}$$