Day 3: Overfitting and Generalization

Summer STEM: Machine Learning

Department of Electrical Engineering
NYU Tandon School of Engineering
Brooklyn, New York

June 25, 2020

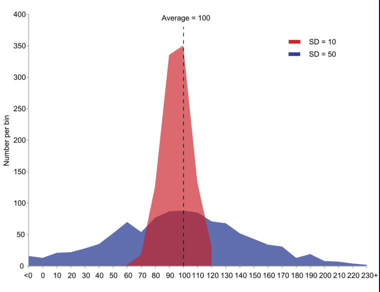
Outline

Basic Concepts

- **Mean** (average value): $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- **Variance** describes the spread of the data with respect to the mean.
- **Covariance** describes the relationship between two variables.

Variance

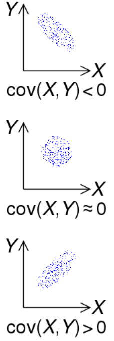
- Variance: $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$



https://en.wikipedia.org/wiki/Variance

Covariance

- Covariance: $\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$



https://en.wikipedia.org/wiki/Covariance

NYU TANDON SCHOOL OF ENGINEERING

Mean, Variance, and Covariance, Correlation Coefficient

- Given feature-target data (x_i, y_i), $i = 1, 2, ..., N$
- Mean:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- Variance:

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \quad \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

- **Covariance**:

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Least Square Solution: Using Statistics

- Solution:

$$f(x) = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

$$w_1 = \frac{\sigma_{xy}}{\sigma_x^2}, \quad w_0 = \bar{y} - w_1\bar{x}$$

- Prediction:

$$f(x) = w_0 + w_1 x$$

Least Square Solution

- Model:

$$f(x) = w_0 + w_1 x$$

- Loss:

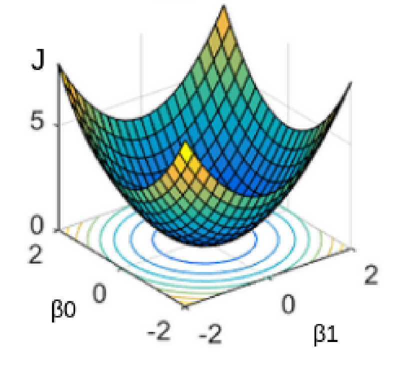
$$J(w_0, w_1) = \frac{1}{N} \sum_{i=1}^{N} \|y_i - f(x_i)\|^2$$

- Optimization: find w_0, w_1 such that $J(w_0, w_1)$ is the least possible value (hence the name "least square").

NYU TANDON SCHOOL OF ENGINEERING

Loss Landscape

Plot the loss against the parameters:



Linear Regression

- Linear models: For scalar-valued feature x, this is $f(x) = w_1 x + w_0$
- One of the simplest machine learning model, yet very powerful.
- Two ways to get the solution, we will show them later.

Least Square Solution: Using Pseudo-Inverse

- For N data points (x_i, y_i) we have,

$$
\begin{aligned}
y_1 &\approx w_0 + w_1 x_1 \\
y_2 &\approx w_0 + w_1 x_2 \\
&\vdots \\
y_N &\approx w_0 + w_1 x_N .
\end{aligned}
$$

Linear Regression

- In matrix form we have,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \approx \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

- We can write it as $Y \approx X\mathbf{w}$. We call X the design matrix.
- Exercise: verify $\|Y - X\mathbf{w}\|^2 = \sum_{i=1}^{N} \|y_i - (w_0 + w_1 x_i)\|^2$

Linear Least Square

- $\min_{\mathbf{w}} \frac{1}{N} \|Y - X\mathbf{w}\|^2$
- Using the psuedo-inverse (only square matrices have an inverse),

$$
\begin{aligned}
Y &= X\mathbf{w} \\
X^T Y &= X^T X \mathbf{w} \\
(X^T X)^{-1} X^T Y &= (X^T X)^{-1} X^T X \mathbf{w} \\
(X^T X)^{-1} X^T Y &= \mathbf{w}.
\end{aligned}
$$

Linear Regression

- What if we have multivariate data with $\mathbf{x}$ being a vector?
- Ex: $\mathbf{x}_i = [x_{i1}, x_{i2}]^T$

$$
\begin{aligned}
y_1 &\approx w_0 + w_1 x_{11} + w_2 x_{12} = \hat{y}_1 \\
y_2 &\approx w_0 + w_1 x_{21} + w_2 x_{22} = \hat{y}_2 \\
&\vdots \\
y_N &\approx w_0 + w_1 x_{N1} + w_2 x_{N2} = \hat{y}_N
\end{aligned}
$$

- The model can be written as $\hat{y}_i = \mathbf{w}^T \phi(\mathbf{x}_i)$, here both $\mathbf{w} = [w_0, w_1, w_2]^T$ and $\mathbf{x}$ are vectors. $\phi(\mathbf{x})$ is a feature transformation that transforms the original feature to $\phi(\mathbf{x}_i) = [1, x_{i1}, x_{i2}]^T$.

Multilinear Regression

- In matrix-vector form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_{n2} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

- Solution remains the same $(X^T X)^{-1} X^T Y = \mathbf{w}$
- Exercise: open `demo_multilinear.ipynb`