1 Introduction

Given an example $x = \{x^1, x^2, ..., x^d\}$, where x^d is the d-th feature, the linear model is defined as

$$f(x) = w^1x^1 + w^2x^2 + ... + w^dx^d + b. \tag{1}$$

Its' vectorized formulation can be written:

$$f(x) = \mathbf{w}^T X + b, \tag{2}$$

where $\mathbf{w} = \{w^1; w^2; ...; w^d\}$.

2 Linear Regression

Problem Def. Given the dataset $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, where $x_n = \{x_n^1, x_n^2, ..., x_n^d\}$ and $y \in \mathbb{R}$, the linear regression is to learn a model to predict the value y.
For convenience, we consider d=1 such that $x_n \in \mathbb{R}$. With Eq. 1, the linear regression can be constructed as:

$$f(x_n) = wx_n + b \ \text{ s.t. } \ f(x_n) = y_n. \tag{3}$$

How to calculate w and b? The problem lies in the measure between the predict and the true value. The mean squared error (MSE) is commonly used in the regression task:

$$\begin{aligned} Error &= \sum_i^n (f(x_n) - y_n)^2, \\ Error &= \sum_i^n (wx_n + b - y_n)^2. \end{aligned} \tag{4}$$

Least squares method. [1]

$$\begin{aligned} \frac{\partial E}{\partial w} &= 2(w\sum_i^n x_i^2 - \sum_i^n (y_i - b)x_i), \\ \frac{\partial E}{\partial b} &= 2(nb - \sum_i^n (y_i - wx_i)). \end{aligned} \tag{5}$$

The closed-form:

$$\begin{aligned} w &= \frac{\sum_i^n y_i(x_i - \bar{x})}{\sum_i^n x_i^2 - 1/m(\sum_i^n x_i)^2}, \\ b &= 1/m\sum_i^n (y_i - wx_i). \end{aligned} \tag{6}$$

Generalizing to the multivariate linear regression, in which $x_n = \{x_n^d, ..., x_n^d, 1\}$, the error can be written as:

$$\begin{aligned} Error &= |\mathbf{w}X - \mathrm{Y}|^2 \\ &= (\mathbf{w}^T X - \mathrm{Y})^T(\mathbf{w}X - \mathrm{Y}) \\ &= (X^T\mathbf{w}^T - Y^T)(\mathbf{w}X - Y) \\ &= X^T\mathbf{w}^T\mathbf{w}X - X^T\mathbf{w}^TY - Y^T\mathbf{w}X + Y^TY. \end{aligned} \tag{7}$$

The derivative:

$$\frac{\partial E}{\partial w} = 2X^T(X\mathbf{w} - Y). \tag{8}$$

The closed-form:

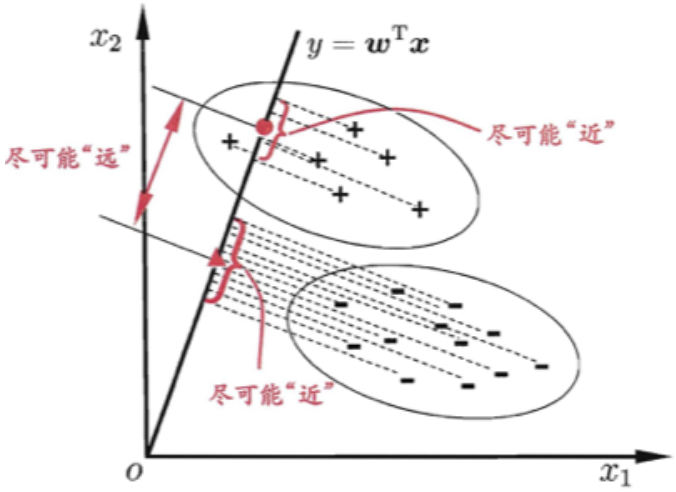
$$\mathbf{w} = (X^TX)^{-1}X^TY, \tag{9}$$

where (X^TX) has to be a full-rand matrix.

[1] https://colab.research.google.com/drive/1h_1gtTb86X4KKrmBbgw13mMOT-3j62jf#scrollTo=jYkD1fYw85w7

3 Linear Model for Classification

Problem Def. Given the dataset $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, where $x_n = \{x_n^1, x_n^2, ..., x_n^d\}$ and $y = \{0, ..., C\}$, the classification model is to predict the class $y_n = C_i$.



We transform the linear regression problem to the classification as the figure shown. This is also called Linear Discriminant Analysis (LDA), which assumes that each class come from an individual Gaussian distribution. The objective can be formulated as minimizing the cross-distance among intra-class projection while maximizing the inter-classes projection:

$$Error = -\frac{||\mu_0 - \mu_1||^2}{\Sigma_0 + \Sigma_1}, \tag{10}$$

where μ denotes the mean and Σ is the variance.
Formally, the variance is defined as within-class scatter matrix:

$$S_w = \Sigma_0 + \Sigma_1 = \sum_0 (x - \mu_0)(x - \mu_0)^T + \sum_1 (x - \mu_1)(x - \mu_1)^T \tag{11}$$

The difference between class is between-class scatter matrix:

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T. \tag{12}$$

After mapping to the line, the objective used for optimization is:

$$J = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}. \tag{13}$$

The derivative:

$$\frac{\partial J}{\partial w} = \frac{2S_b\mathbf{w} \times \mathbf{w}^T S_w \mathbf{w} - 2S_w\mathbf{w} \times \mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} = 0 \tag{14}$$

The closed-form:

$$\begin{aligned} &S_b\mathbf{w} \times \mathbf{w}^T S_w \mathbf{w} = S_w\mathbf{w} \times \mathbf{w}^T S_b \mathbf{w} \\ &S_b\mathbf{w} = S_w\mathbf{w}(\frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}) \\ &(S_w^{-1} S_b)\mathbf{w} = \lambda\mathbf{w} \end{aligned} \tag{15}$$

If $(S_w^{-1} S_b)$ is non-singular matrix, $\mathbf{w}$ is the Eigenvector of $(S_w^{-1} S_b)$.

4 Recommended Reading

[1] https://en.m.wikipedia.org/wiki/Linear_discriminant_analysis.
[2] https://en.m.wikipedia.org/wiki/Logistic_regression
[3] https://stats.stackexchange.com/questions/95247/logistic-regression-vs-lda-as-two-class-classifiers
[4] https://en.m.wikipedia.org/wiki/Generalized_linear_model