CS 4980/6980: Introduction to Data Science © Spring 2018

Lecture 4: Linear Regression

Instructor: Daniel L. Pimentel-Alarcón Scribed by: Tuan Phan

This is preliminary work and has not been reviewed by instructor. If you have comments about typos, errors, notation inconsistencies, etc., please email Tuan Phan at tphan36@student.gsu.edu.

4.1 Homework Review

4.2 Introduction

Linear regression is a linear approach for modeling the relationship between two variables by fitting a linear equation to observed data.
Goal: Understand how a variable of interest $\mathbf{Y}$ depends on a sequence of variable $\mathbf{X}$.

Example 4.1. Suppose that we need to find the glusose level:

$$glucose\ level = \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 gender + \beta_5 race + \beta_6 diet + \beta_7 cholesterol$$

The glucose level should be a linear combination of all of:

Glucose level	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

y: variable of interest (glucose level).

Goal: Find $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_D \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_D \end{bmatrix} \epsilon\ \mathbb{R}^D$$

β: coefficient weight parameters
$\mathbf{X}$: feature vector

The β will tell the dependence of glucose level on all of the features(weight, heigh, etc...). Then:
$\mathbf{Y} = \beta_0 x_+ \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_D x_D$

or it can be rewrited as:
$\mathbf{Y} = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}$

Question:
How do we find $\boldsymbol{\beta}$?
When we are able to find vector β, we are able to predict what everybody's gludose level are gonna be based on those features (weight, heigh, etc...).

Solution:
To find $\boldsymbol{\beta}$, we can use "trainning data" (in CSC)/ "samples data" (in Math Statistic), that has all the data of the feature vectors that we care about (weight, height, age, ect) and after we have some samples or some information of some people such as glucose level (variable of interest), we can understand how exactly that variable of interest depend on those features.
We observe Y_1, Y_2, Y_3, ..., Y_N and $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, ..., $\mathbf{x}_N$

Assuming that: $Y_i = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}_i$, for i = 1, 2, ..., N
We need to find:

$$\begin{cases} Y_1 &= \beta^T X_1 \\ Y_2 &= \beta^T X_2 \\ \vdots & \\ Y_N &= \beta^T X_N \end{cases} \qquad \overset{or}{\mathbf{Y} = \boldsymbol{\beta}^T \mathbf{X}}$$

The equation $\mathbf{Y} = \boldsymbol{\beta}^T\mathbf{X}$ can be expanding as:

$$\begin{bmatrix} Y_1, & Y_2, & Y_3, & \cdots, & Y_N \end{bmatrix} = \begin{bmatrix} \beta_1, & \beta_2, & \beta_3, & \cdots, & \beta_N \end{bmatrix}^T \begin{bmatrix} \mathbf{x}_1, & \mathbf{x}_2, & \mathbf{x}_3, & \cdots, & \mathbf{x}_N \end{bmatrix}$$

$$with\ \mathbf{Y} \epsilon\ \mathbb{R}^{1\times N}\ , \mathbf{x}\ \epsilon\ \mathbb{R}^{D\times N}$$

Denotation: $\mathbf{x}$ is a matrix with dimension D × N (same for $\mathbf{Y}$)
Now, we can find β such that $\mathbf{Y} = \boldsymbol{\beta}^T\mathbf{X}$ by using:

$$\widehat{\beta} := \underset{\beta\ \epsilon\ \mathbb{R}^D}{argmin}\ ||(Y - \beta^T\mathbf{X})^T\ ||_2^2 \leftarrow \sum_{i=1}^N (Y_i - \beta^T x)^2$$

Recall:
$||\mathbf{x}||_2$ is the norm type 2, also call the size of the vector or the size of norm of $\mathbf{x}$.

$$||\mathbf{x}||_2^2 := \sum_{i=1}^N x_i^2\ , \qquad with\ \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

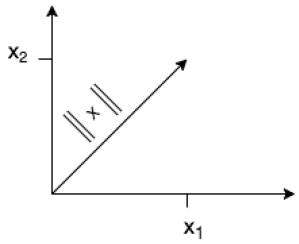

Figure 4.1: Norm of x.

Also, we have: $\mathbf{Y} = Y^T$

- $\mathbf{Y}$ itself is a matrix
- Y^T transpose is a vector

The "arg min " means that $\widehat{\beta}$ is going to be the argument that minimizes those functions over all the possible values of $\mathbb{R}^D$. We don't want to find the smallest β, but we want to find the β that makes those quantities as small as possible. Even if β is huge, we don't need to care about the size of β, but we need to care that β makes those norm small. Because those norms are basically our errors, the overall error of all of our sample.

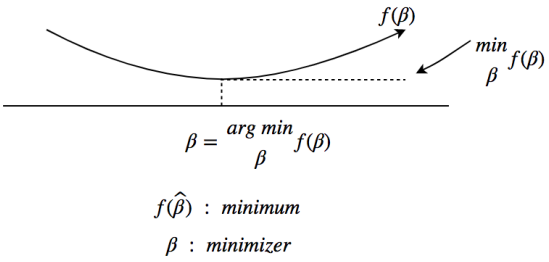

Figure 4.2: Minimum of β.

$\widehat{\beta}$ is called minimizer or the argument that minimize the function. There can be more than one value for β.

Continue with:

$$\widehat{\beta} := \underset{\beta \,\epsilon\, \mathbb{R}^D}{arg\ min}\ ||(\mathbf{Y} - \beta^T \mathbf{X})^T\ ||_2^2$$

Remember that, all variables we discussed and the feature vector such as $\mathbf{x}_i$ has D dimension. In D-dimension, we can model glucose as height, weight, age, etc. Suppose we have $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ in one-dimension, then we can plot it in the cholesterol level graph to predict the cholesterol level of each person.

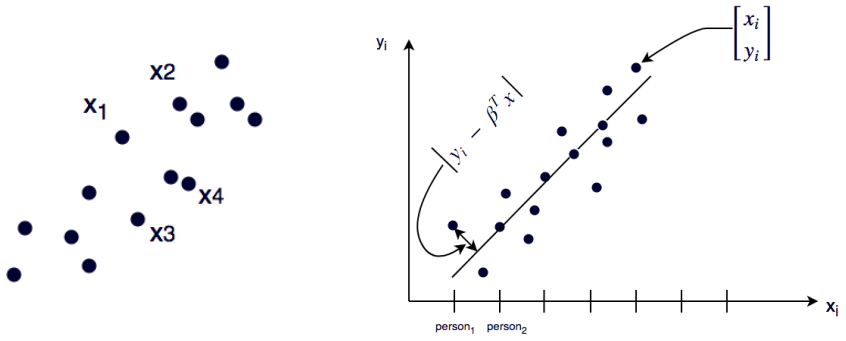

Figure 4.3: Cholesterol Level

Now, we are looking to find some β that explains $\mathbf{Y}$ as a function of $\mathbf{X}$ and this is similiar to the function: y = mx + b

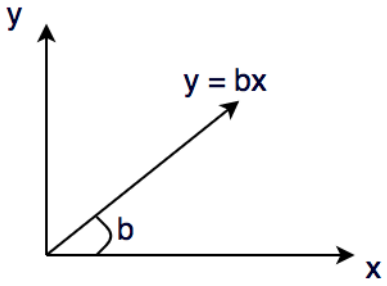

Figure 4.4: y = bx

We need to find a line that explains all of these vectors as accurately as possible. For example, in the graph above, as soon as we find b, we would find the slope of the line. So as soon as we find the β in the linear equation, we will also find that high dimensional line.

Note: For any line that are drawed, there always some errors; for instance, there are points that are not exactly lie on the line (Figure 4.3). The distance from those points to the line, which uses to explain data, is measure as $|y_i - \beta^T x|$, are error. Adding all of those errors, we need to figure out the best line that achieves those minimum errors. Therefore, we have:

$$\widehat{\beta} := \underset{\beta \,\epsilon\, \mathbb{R}^D}{arg\ min} \; ||(\mathbf{Y} - \beta^T \mathbf{X})^T \; ||_2^2 \; = \; \underset{\beta \,\epsilon\, \mathbb{R}^D}{arg\ min} \; (\mathbf{Y} - \beta^T \mathbf{X}) \; (\mathbf{Y} - \beta^T \mathbf{X})^T$$

For $(\mathbf{Y}-\beta^T\mathbf{X})\,(\mathbf{Y}-\beta^T\mathbf{X})^T$, we can look at this in an other way for for better understanding. Assume that: $(\mathbf{Y}-\beta^T\mathbf{X}) = \mathbf{Z}^T$ and $(\mathbf{Y}-\beta^T\mathbf{X})^T = \mathbf{Z}$ then we have:

$$\underbrace{(\mathbf{Y}-\beta^T\mathbf{X})}_{\mathbf{Z}^T}\underbrace{(\mathbf{Y}-\beta^T\mathbf{X})^T}_{\mathbf{Z}} = \mathbf{Z}^T\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_1, & \mathbf{Z}_2, & \mathbf{Z}_3, & \cdots, & \mathbf{Z}_N\end{bmatrix}\begin{bmatrix}\mathbf{Z}_1,\\ \mathbf{Z}_2,\\ \mathbf{Z}_3,\\ \vdots\\ \mathbf{Z}_N\end{bmatrix}$$

$$with:\ \mathbf{Z} = \begin{bmatrix}Z_1,\\ Z_2,\\ Z_3,\\ \vdots\\ Z_N\end{bmatrix} := \begin{bmatrix}Y_1-\beta^T X_1,\\ Y_2-\beta^T X_2,\\ Y_3-\beta^T X_3,\\ \vdots\\ Y_N-\beta^T X_N\end{bmatrix}$$

Next lecture: Solve for β.