Lecture Note - 01: Introduction to Data Science: Toy Problem, Linear Algerbra, Randomness

Dihui Lai

August 29, 2021

Contents

1 Data Science - Toy Problem

1.1 Toy Problem

Suppose you learned from 3 of your friends who went on shopping recently, who bought pants and socks. The number and costs are shown as below:

	Pants	Socks	Cost
John	1	1	23
Lisa	1	2	26
David	1	1	24

According to John and Lisa, the prices of a pant and a sock can be calcuated as $P = 20$ and $S = 3$, respectively. However, David should have paid 23 dollar given the inferred prices. Why did David pay 24 dollar instead of 23? It could be due to price variation. How can we get a good estimation of the prices?

1.2 Price Estimation

To get a good estimation of the prices of socks and pants, we can use the following error function

$$\epsilon = (P + S - 23)^2 + (P + 2S - 26)^2 + (P + S - 24)^2$$

Ideally, we would like to have our estimated socks (S) and pants (P) price as close to the real cost, i.e. minimize ϵ. Use basic calculus knowledge, we know the optimal P and S should satisfy the following equations.

$$\begin{cases} \frac{\partial \epsilon}{\partial P} = 0 \\ \frac{\partial \epsilon}{\partial S} = 0 \end{cases} \tag{1}$$

$$\begin{cases} \frac{\partial \epsilon}{\partial P} = 2(P + S - 23) + 2(P + 2S - 26) + 2(P + S - 24) = 0 \\ \frac{\partial \epsilon}{\partial S} = 2(P + S - 23) + 4(P + 2S - 26) + 2(P + S - 24) = 0 \end{cases} \Longrightarrow \begin{cases} 9P + 12S - 219 = 0 \\ 8P + 12S - 198 = 0 \end{cases} \tag{2}$$

$$\begin{cases} P = 21 \\ S = 2.5 \end{cases} \tag{3}$$

2 Structured Data and Linear Model

2.1 Tabular Data and Matrix

In general, if we want to consider a model of m types of goods and collect data from n people. The toy model can be generalized to a problem that needs to estimate m variables on n data points

$$\begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{bmatrix} \sim \begin{bmatrix} x_1^1 & x_2^1 & x_1^3 & \dots & x_m^1 \\ x_1^2 & x_2^2 & x_1^2 & \dots & x_m^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^n & x_2^n & x_1^n & \dots & x_m^n \end{bmatrix}$$

Using vector notation, we have

$$Y = \vec{y} = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{bmatrix}$$

,

$$X = [\vec{x}_1, \vec{x}_2, \dots, \vec{x}_m] = \begin{bmatrix} x_1^1 & x_2^1 & x_1^3 & \dots & x_m^1 \\ x_1^2 & x_2^2 & x_1^2 & \dots & x_m^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^n & x_2^n & x_1^n & \dots & x_m^n \end{bmatrix}$$

Where $\vec{x}_j$ is a column vector/matrix of size $n \times 1$

$$\vec{x}_j = \begin{bmatrix} x_j^1 \\ x_j^2 \\ \vdots \\ x_j^n \end{bmatrix}, j = 1, 2, 3, ..., m$$

The vectors $\vec{x}_i$ are called covariates, or predictors. $\vec{y}$ is normally called target variable

2.2 Linear Model

If we assume $\vec{y}$ is linearly dependent on $\vec{x}$s, we have a linear model

$$\hat{\vec{y}} = \beta_1 \vec{x}_2 + \beta_2 \vec{x}_2 + \cdots + \beta_m \vec{x}_m$$

An optimal model should estimate $\hat{\vec{y}}$ as close as $\vec{y}$. We can define an error function, for example, using sum square error

$$\epsilon = (\hat{\vec{y}} - \vec{y}) \cdot (\hat{\vec{y}} - \vec{y}) = \sum_{i=1}^{n} (\hat{y}^i - y^i)^2$$

To have the error term to be minimized, we want to have

$$\frac{\partial \epsilon}{\partial \beta_j} = 0, \text{ j=1, 2, 3,..., m}$$

This problem is called least square estimation. How can we solve the problem? We are going to solve it in our lecture-3

2.3 Other Models

In general, $\vec{y}$ could be any function of $\vec{x}$ i.e. $\vec{y} = f(\vec{x})$.

- Kepler's Law: $T^2 \sim r^3$. Note Kepler's law becomes linear if we do a log transformation on both side i.e. $2 \log T = 3 \log r$
- House price: $P \sim f(size, location)$

3 Geometric Interpretation, Visualization and Randomness

3.1 Pants, Socks and Cost from 100 Friends

Suppose now you collect data from 100 friends. Everyone of them has bought a few pants and some socks from the same store. In a perfect world, everyone remember the number of pants/socks they bought and the corresponding cost. You end up having the complete data set like this:

	Socks	Pants	Cost
Person1	3	6	129
Person2	8	6	144
Person3	8	5	124
Person4	8	3	84
Person5	3	6	129
...	...	...	...
Person100	...	...	...

By looking at the number closely, you figured out that the price of a pant is 20 dollars and the price of a sock is 3 dollars. And it is consistent across the whole data set.

3.2 Incomplete Data and Visualization

In reality it is very hard to get all information you need to build a pricing model for pants and socks, not everyone is going to give you the complete information about their purchase. Let us

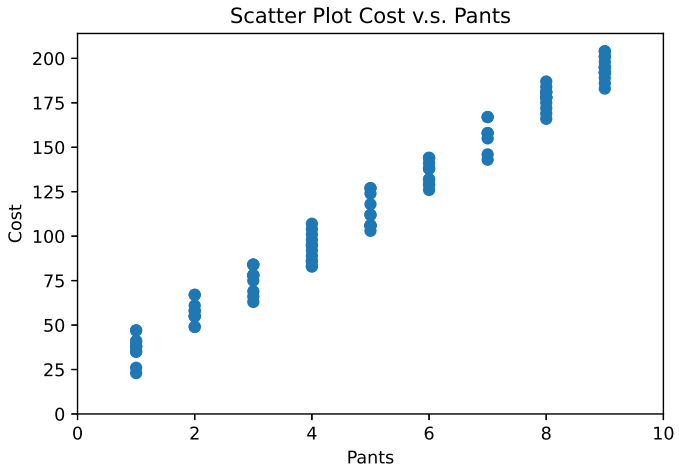

Figure 1: Scatter plot of total cost against the number of pants bought

assume that all your friends only tell you the number of pants that they bought and the total cost of their purchase. So you end up having a data set like this:

	Pants	Cost
Person1	6	129
Person2	6	144
Person3	5	124
Person4	3	84
Person5	6	129
...	...	...
Person100	...	...

If you do not know that your friends actually bought things other than pants, the data looks a bit puzzeling. Because some data points imply that the price of a pant is \$21.5 and some imply \$24. If you can not have a bettter guess, you would assume the price of pants varies in the market and the fluctuation looks a lot like some sourt of noise. You can confirm your guess by making a scatter plot (Figure 1).

On the other hand, if all your friends only tells you the number of socks that they bought and you have a data set.

	Socks	Cost	
Person1	3	129	
Person2	8	144	
Person3	8	124	
Person4	8	84	
Person5	3	129	
...	...	...	...
Person100	...	...	...

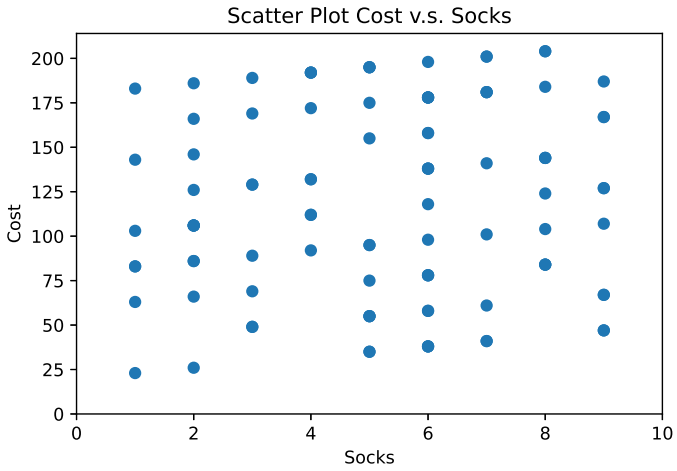

Figure 2: Scatter plot of total cost against the number of socks bought

The scatter plot looks even more puzzeling (Figure 2). The cost looks as if it is not dependent on the number of socks bought at all. In this case you might guess that the cost of buying socks is totally random.

3.3 Randomness caused by High Dimension Projection

In our toy model, we are considering the cost of buying two types of item. Mathematically, the total cost is a perfectly linear function of the number of pants/socks bought i.e. $cost_{total} = price_{pants} \times pants + price_{socks} \times socks$. The price of pants/socks are deterministic and there is no randomness at all. However, the total cost appears as if the there are some random impact that causes the price os pants/socks to change

While dealing with real world data problem, missing information is almost guaranteed. For example, while modeling the house price, it is unlikely that we will know the price that the buyers are willing to pay; while modeling a public company's stock price, it is almost impossible to know all information related to the company. **Therefore, we have to make good assumptions about the information that we do not have, a.k.a. noise**.