Lecture Note - 01

Dihui Lai

March 5, 2020

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	2	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.

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

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

,

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

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

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

An optimal model should estimate $\hat{\vec{y}}$ as close as $\vec{y}$. e.g.

$$\epsilon = (\hat{\vec{y}} - \vec{y}) \cdot (\hat{\vec{y}} - \vec{y})$$

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

How can we solve the problem?

2.3 General Models

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

- Kepler's Law: $T^2 \sim r^3$
- House price: $P \sim f(size, location)$

3 Geometric Interpretation and Visualization

Scatter plot and projection operation pick any two columns from matrix $[Y, X]$