Foundation of Analytics: Lecture 1

Dihui Lai

dlai@wustl.edu

August 18, 2019

CONTENT

- Data Analytics in Science, Finance, Insurance, Health Care etc.
- Mathematics of Data Analytics & Artificial Intelligence
- Computational Tools, Library, Packages, Softwares
- Review of Linear Algerbra

Kepler's Law of Planetary Motion

Planet	Distance to Sun (AU)	Period(days)
Mercury	0.389	87.77
Venus	0.724	224.70
Earth	1	365.25
Mars	1.524	686.95
Jupiter	5.2	4332.62
Saturn	9.510	10759.2

What is the mathematical model for $Period = f(DistanceSun)$?

Realtor Housing Price

house price	logitude	lattitude	age	oceanProx	size	...
452600.0	-122.23	37.88	41	NEAR BAY	85768	...
358500.0	-122.22	37.86	21	NEAR BAY	40803	...
352100.0	-122.24	37.85	52	NEAR BAY	63085	...
...	...	...				

What is the mathematical model for House Price $= f(location, size, ...)$?

Predict Heart Disease

heart disease	age	chest pain type	fbs	thalach	gender	...
Yes	63	0	1	150	F	...
No	45	1	0	170	F	...
No	70	0	0	168	M	...
No	30	3	0	190	F	...
Yes	55	2	0	148	M	...
No	26	1	1	155	M	...
...	...	...				

What is the model for heart disease $= f(age, gender, fbs, ...)$?

Mathematic Model for Structured Data

Given a dataset

$$\begin{bmatrix} y^1 \\ y^2 \\ y^3 \\ y^i \\ \vdots \\ y^n \end{bmatrix} \& \begin{bmatrix} x_1^1 & x_2^1 & x_3^1 & \cdots & x_m^1 \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_m^2 \\ x_1^3 & x_2^3 & x_3^3 & \cdots & x_m^3 \\ x_1^i & x_2^i & x_3^i & \cdots & x_m^i \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_1^n & x_2^n & x_3^n & \cdots & x_m^n \end{bmatrix}$$

What is the model for $y = f(x_1, x_2, x_3, ...x_m)$?

What is f?

Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_m x_m$$

Variation

$$ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_m x_m$$

What is f?

When y is observables of a random process and the same $x_1, x_2...x_m$ will leads to different y i.e.

$$y \sim P(x_1, x_2...x_m)$$

For example: Logistic Regression; Poisson Regression; Generalized Linear Model

What is f?

When close math formula does not provide good enough approximation for the problem?

$$f \sim \mathit{Neuralnetwork}; \mathit{Tree}; \mathit{RandomForest}$$

How to find f?

Training Algorithms:

- Maximum Likelihood Estimation; Entropy Maximization
- Gradient Descent; Stochastic Gradient Descent;
- Greedy Seach

Mathematics of Data Analytics

- Linear Algerbra (to Handle High Dimension Data Space)
- Statistics (to Handle Randomness in Data)
- Calculus (to Find the Optimal Solution/Model/Function)

Linear Algebra Review

- Matrix Addition, Multiplication,
- Inverse $XX^{-1} = X^{-1}X = I$,
- Transpose M^T
- Linear Combination $a\vec{x} + b\vec{y} + c\vec{z}$
- Dot Product $\vec{x} \cdot \vec{y} = x_1y_1 + x_2y_2 + x_3y_3$
- Geometric Ineterpretation of Linear Algebra: Linear Independent, Linearly Dependent

Python

- Python; 'pip' installtion tools
- Packages: numpy; sklearn etc.
- IDE: jupyter-notebook, pyCharm etc.
- Virtual Environment