應用機器學習

Brian Chan 陳醒凡

課程目標

1. 了解基本的數據分析
2. 了解基本的機器學習(Machine Learning)方法
3. 掌握Python的基本操作和一些有用的package
4. 處理及從網上下載數據
5. 在Python上應用機器學習

今天課堂概要

1. Optimization algorithm in ML

- Optimization framework
- Application: Least square method

2. Statistical foundation of machine learning methods

- Probability distribution
- Estimation framework
- Basics of statistics
- Bayesian probability

1. Optimization algorithm in ML

- Optimization framework

- Application: Least square method

OPTIMIZATION ALGORITHM (BASIC)

$$\min_{\theta} J(\theta)$$
$$\text{such that } \theta \in \Theta$$

1. Objective function, $J(\theta)$
2. Control variables, θ
3. Constraints, $\theta \in \Theta$

Optimization is a very useful issue in ML.

Many ML methods reply optimization algorithms to solve the parameters in the model.

The example includes regression methods and the class of NN models.

EXAMPLES

Constrainted optimization

$$\min_x \; x^2 + 2x + 1$$

$$\text{s.t. } x \in [1,2]$$

$$x \in \mathbb{R}$$

Portfolio optimization

$$\min_x \; x^T \Sigma x$$

$$\text{s.t. } x^T \mathbb{1} = 1$$

$$x \geq 0$$

$$x \in \mathbb{R}^n, \Sigma \in S_+^n$$

Argmax (or argmin) are the points, or elements, of the domain of some function at which the function values are maximized (or minimized).

GRADIENT DESCENT

Gradient descent is one of the most basic but commonly used optimization algorithms. Many advanced optimization algorithm is developed based on gradient descent.

Under some conditions, the optimization can be solved by gradient descent.

$$\theta \leftarrow \theta - \eta \cdot \frac{\partial J(\theta)}{\partial \theta}$$

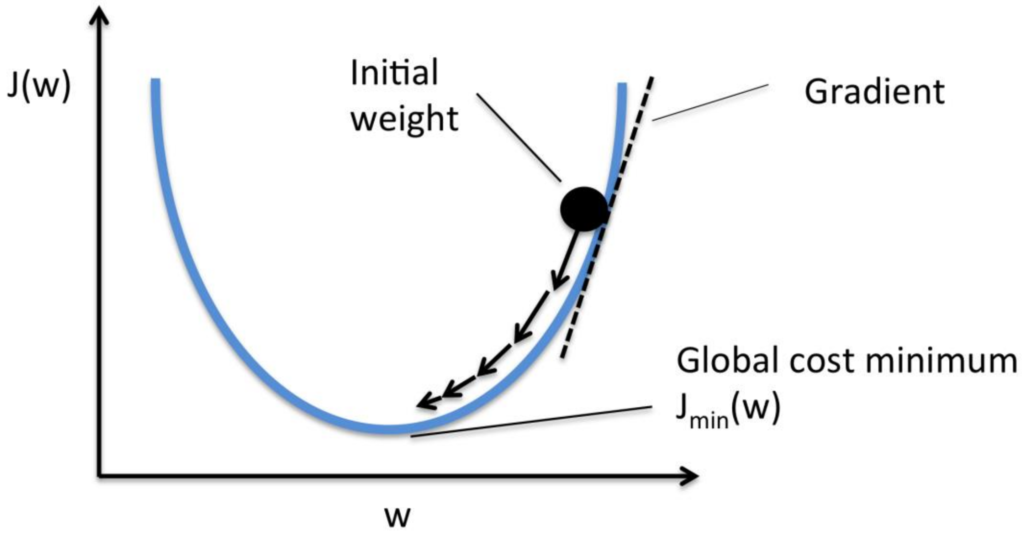
https://en.wikipedia.org/wiki/Gradient_descent

GRADIENT DESCENT

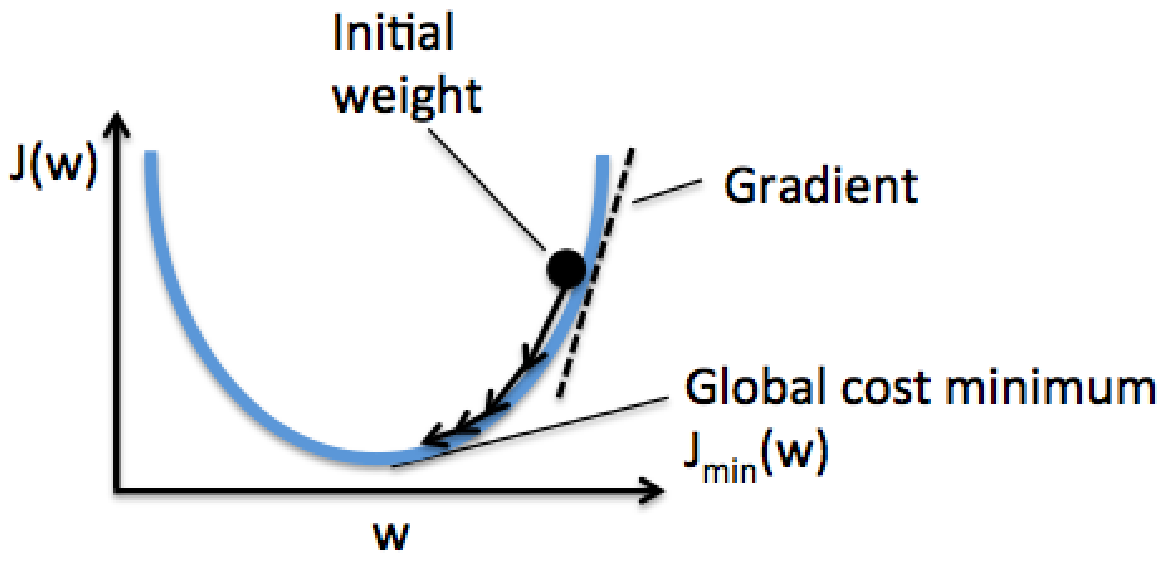
At initial step $i = 1$, the control variable is set to be $\theta^1 = \bar{\theta}$.

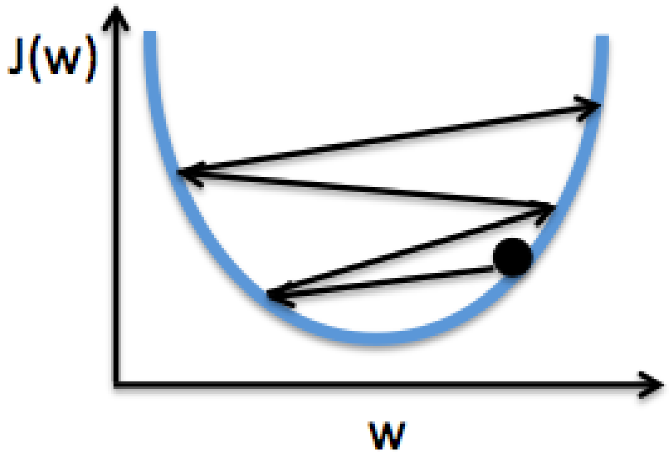
At each iteration, θ^i is updated with $\theta^i \leftarrow \theta^i - \eta \cdot \frac{\partial J(\theta)}{\partial \theta}|_{\theta=\theta^{i-1}}$.

Repeat this until $|\theta^i - \theta^{i-1}| < \delta$.



GRADIENT DESCENT





SOME MORE OPTIMIZATION ALGORITHM IN ML

Stochastic gradient descent

Momentum

Adagrad

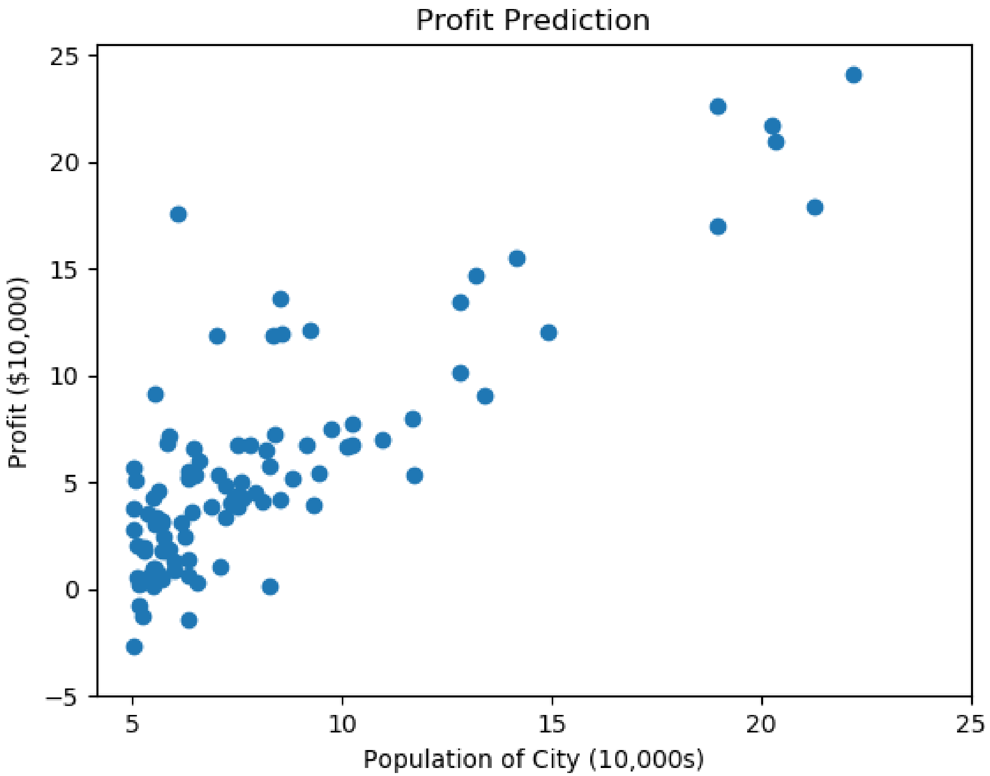
RMSProp

Adam

and some more …

http://ruder.io/optimizing-gradient-descent/index.html#adam

EXAMPLE: LEAST SQUARE METHOD



Profit Prediction
Profit ($10,000)
Population of City (10,000s)
25
20
15
10
5
0
−5
5
10
15
20
25

LEAST SQUARE METHOD

Let y_i be observed data for $i = 1, \ldots, N$.

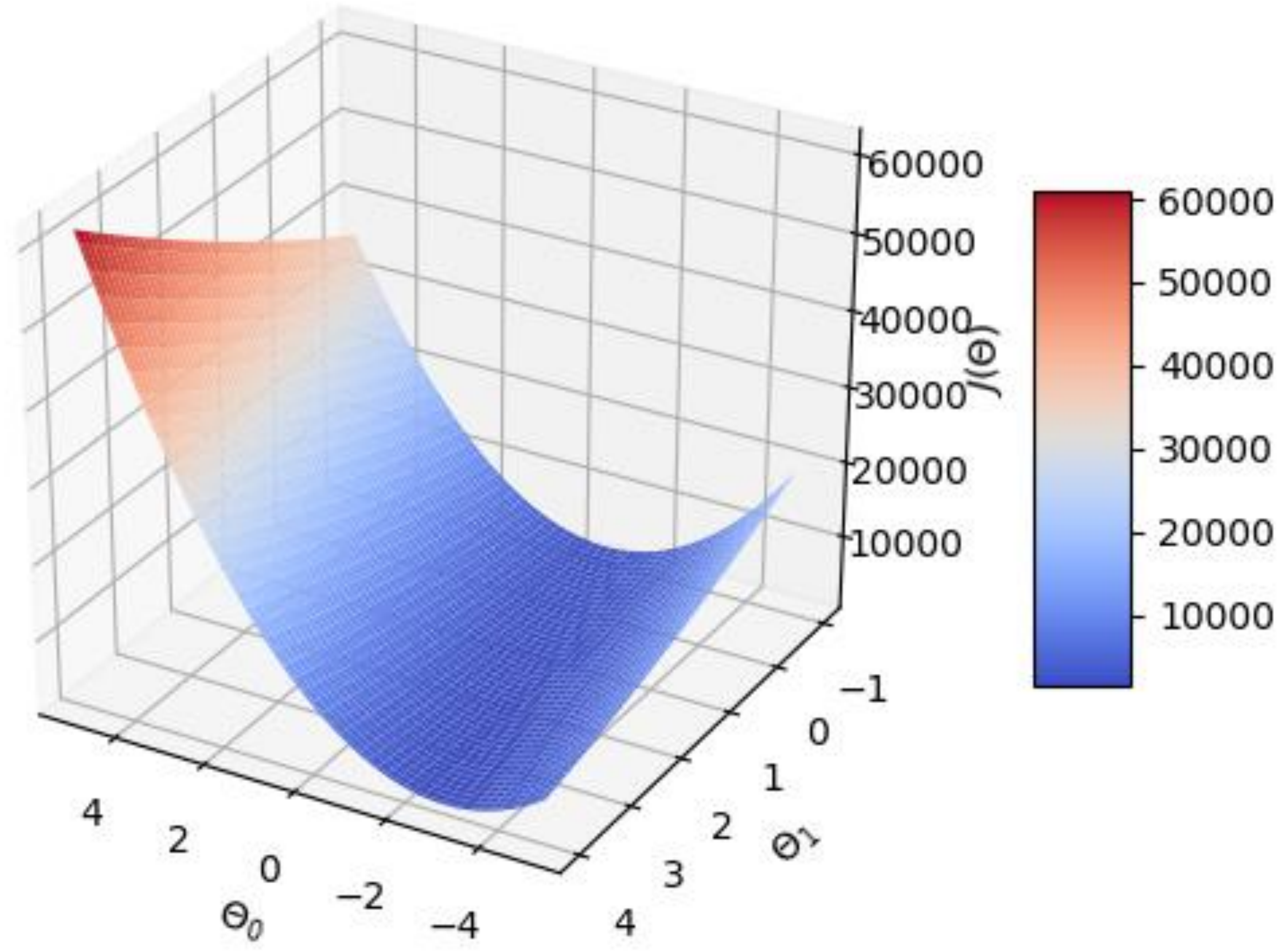
$$J(\theta_0, \theta_1) = \frac{1}{N}\sum_{i=1}^{N}[y_i - (\theta_0 + \theta_1 x_i)]^2$$

LEAST SQUARE METHOD

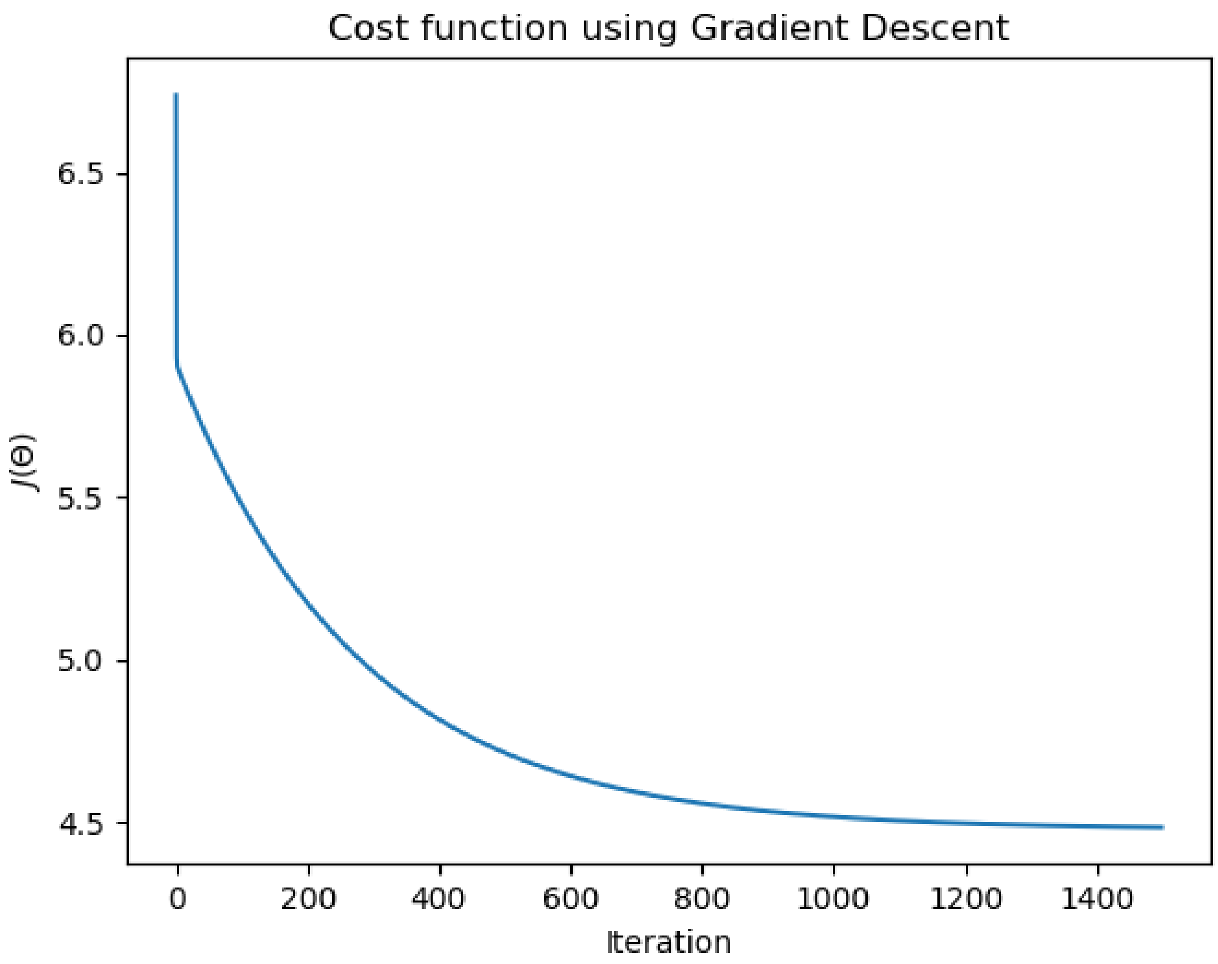
Let y be observed data such that $y = (y_1, y_2, \dots, y_N)^T$ and $Z = (1, X)^T$.

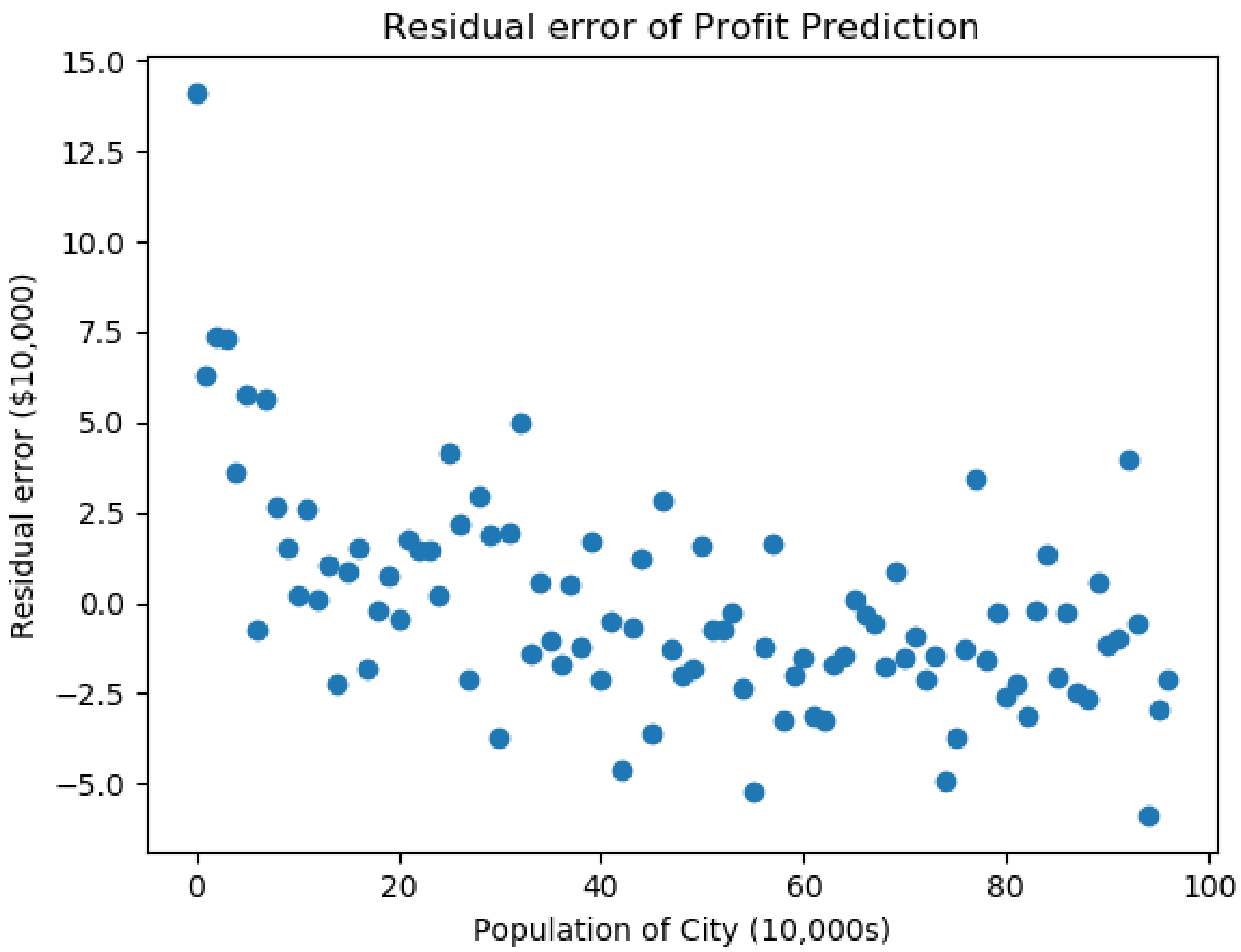
$$J(\theta) = \frac{1}{N} ||y - Z \cdot \theta||^2,$$

where $\theta = (\theta_1, \theta_2)^T$.



60000
50000
40000
30000
20000
10000
J(Θ)
60000
50000
40000
30000
20000
10000
4
2
0
−2
−4
Θ0
−1
0
1
2
3
4
Θ1


Cost function using Gradient Descent
6.5
6.0
5.5
5.0
4.5
J(Θ)
0
200
400
600
800
1000
1200
1400
Iteration


Residual error of Profit Prediction
Residual error ($10,000)
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
0
20
40
60
80
100
Population of City (10,000s)

2. Statistical foundation of machine learning methods

- Probability distribution
- Estimation framework
- Basics of statistics
- Bayesian probability

PROBABILITY (DISCRETE)

- Event
- Probability measure

Discrete probability distribution:

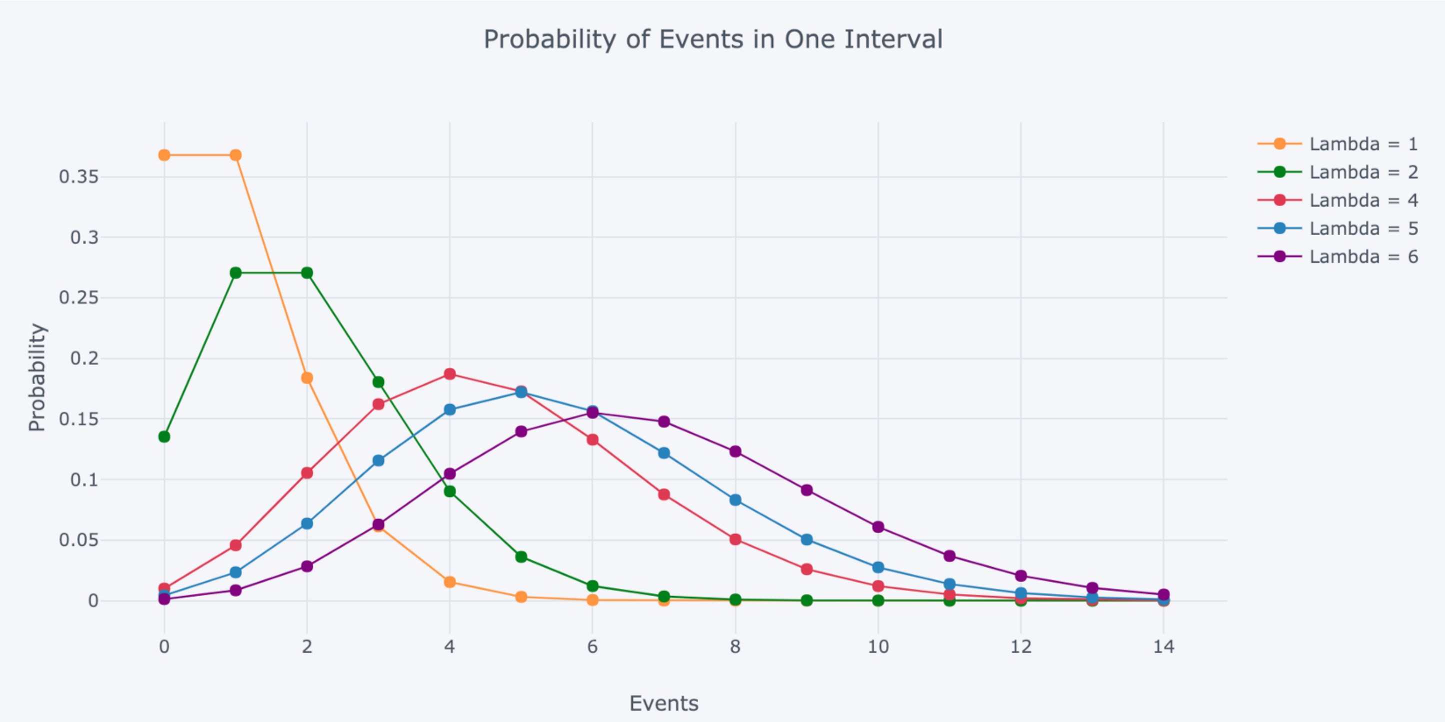
P({Head}) = P({Tail}) = 0.5

P({1}) = P({2}) = P({3}) = P({4}) = P({5}) = P({6}) =1/6

P({1},{2}) = P({1})+P({2}) = 1/6+1/6=1/3

NUMBER OF METEORS IN 1 HOUR

Poisson distribution



https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459

CREDIT MIGRATION

	Moody's	S&P	Fitch	Meaning
Investment Grade	Aaa	AAA	AAA	Prime
	Aa1	AA+	AA+	High Grade
	Aa2	AA	AA	
	Aa3	AA-	AA-	
	A1	A+	A+	Upper Medium Grade
	A2	A	A	
	A3	A-	A-	
	Baa1	BBB+	BBB+	Lower Medium Grade
	Baa2	BBB	BBB	
	Baa3	BBB-	BBB-	
Junk	Ba1	BB+	BB+	Non Investment Grade Speculative
	Ba2	BB	BB	
	Ba3	BB-	BB-	
	B1	B+	B+	Highly Speculative
	B2	B	B	
	B3	B-	B-	
	Caa1	CCC+	CCC+	Substantial Risks
	Caa2	CCC	CCC	Extremely Speculative
	Caa3	CCC-	CCC-	In Default w/ Little Prospect for Recovery
	Ca	CC	CC+	
		C	CC	In Default
			CC-	
	D	D	DDD	

Table 1.8

One-year transition matrix (%)

Initial rating	Rating at year-end (%)							
	AAA	AA	A	BBB	BB	B	CCC	Default
AAA	90.81	8.33	0.68	0.06	0.12	0	0	0
AA	0.70	90.65	7.79	0.64	0.06	0.14	0.02	0
A	0.09	2.27	91.05	5.52	0.74	0.26	0.01	0.06
BBB	0.02	0.33	5.95	86.93	5.30	1.17	0.12	0.18
BB	0.03	0.14	0.67	7.73	80.53	8.84	1.00	1.06
B	0	0.11	0.24	0.43	6.48	83.46	4.07	5.20
CCC	0.22	0	0.22	1.30	2.38	11.24	64.86	19.79

Source: Standard & Poor's CreditWeek (15 April 96)

Markov Chain

https://www.moodys.com/sites/products/ProductAttachments/DRD/CTM_Methodology.pdf

MOMENTS

Expectation:

$E[X] = \sum \mathrm{P}(x_i) x_i$

Expectation:

$E[X] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$

Variance:

$Var[X] = \sum \mathrm{P}(x_i)(x_i - \mu)^2$

Variance:

$Var[X] = \frac{1}{6} \cdot (1 - 3.5)^2 + \frac{1}{6} \cdot (2 - 3.5)^2 + \frac{1}{6} \cdot (3 - 3.5)^2 + \frac{1}{6} \cdot (4 - 3.5)^2 + \frac{1}{6} \cdot (5 - 3.5)^2 + \frac{1}{6} \cdot (6 - 3.5)^2$

Covariance:

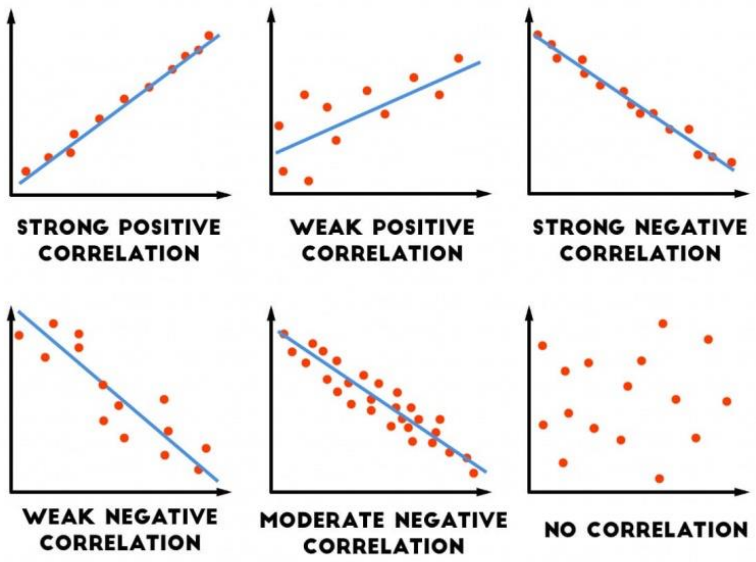
$Cov[X, Y] = E[(X - \mu_x)(Y - \mu_y)]$

Correlation:

$Corr[X, Y] = \frac{Cov[X,Y]}{\sigma_x \sigma_y}$

COVARIANCE & CORRELATION

Dice1\Dice2	0	1
0	1/4	1/4
1	1/4	1/4



Covariance:

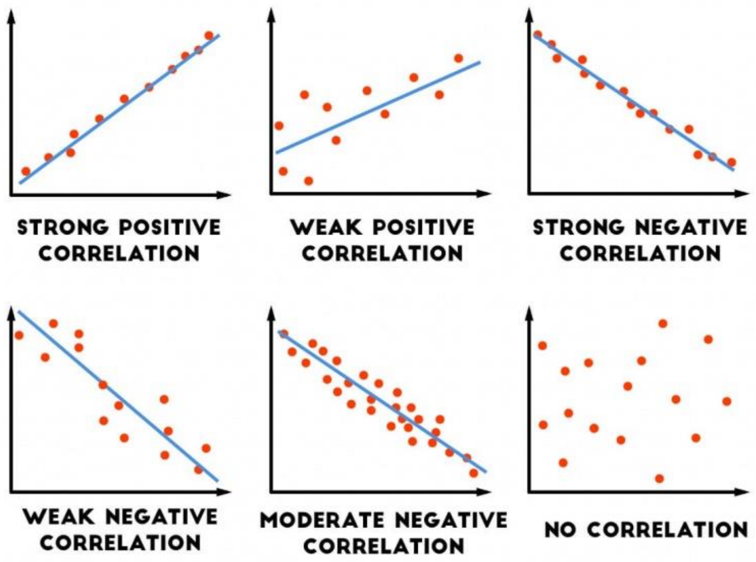
$$Cov[X,Y] = E\left[(X-\mu_x)\left(Y-\mu_y\right)\right] = \frac{1}{4}\cdot(0-0.5)(0-0.5) + \frac{1}{4}\cdot(1-0.5)(0-0.5) + \frac{1}{4}\cdot(0-0.5)(1-0.5) + \frac{1}{4}\cdot(1-0.5)(1-0.5) = 0$$

Correlation:

$$Corr[X,Y] = \frac{Cov[X,Y]}{\sigma_x\sigma_y}$$

COVARIANCE & CORRELATION

Dice1\Dice2	0	1
0	2/4	0
1	0	2/4

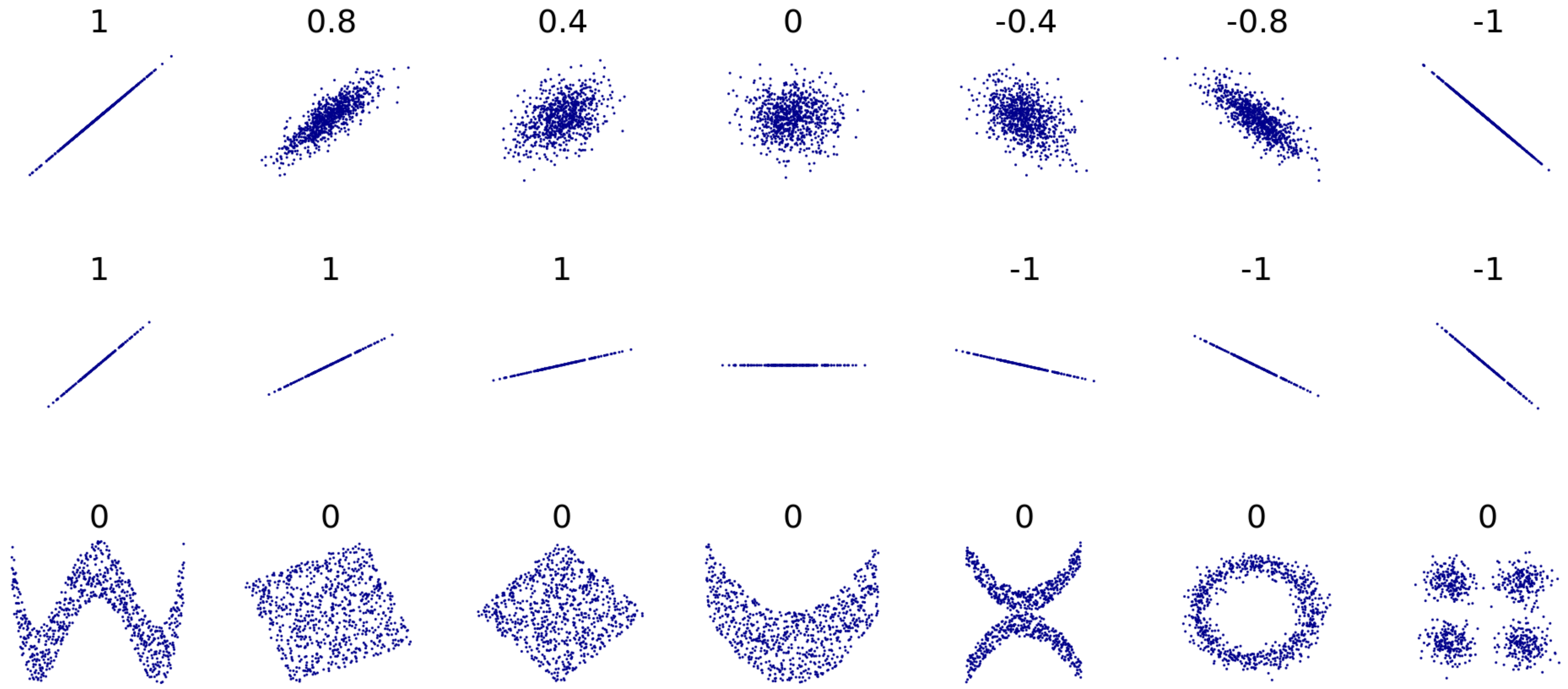


Covariance:

$$Cov[X,Y] = E\left[(X-\mu_x)\left(Y-\mu_y\right)\right] = \frac{2}{4}\cdot(0-0.5)(0-0.5) + \frac{2}{4}\cdot(1-0.5)(1-0.5) = 0.25$$

Correlation:

$$Corr[X,Y] = \frac{Cov[X,Y]}{\sigma_x \sigma_y}$$


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

REMINDER ABOUT CORRELATION

Relationship with regression

Interpretation of correlation:

When correlation = 1

When correlation = 0

Multivariate:

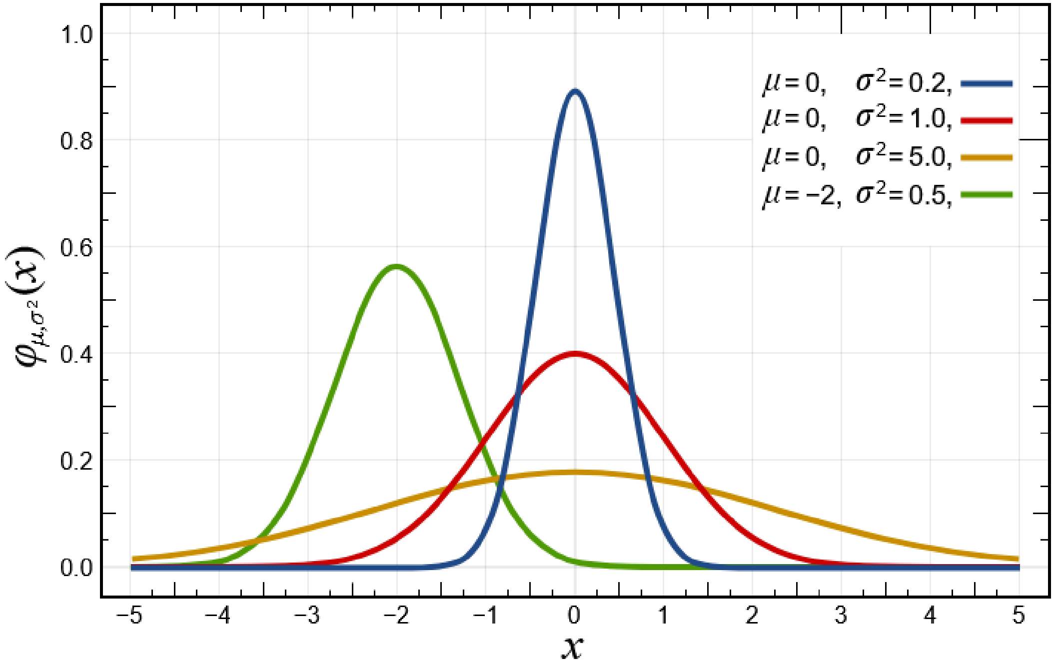
$$\Sigma = E[(X-\mu)(X-\mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{pmatrix},$$

where $X = (x_1, x_2, \dots, x_d)^T$.

PROBABILITY (CONTINUOUS)

Continuous probability distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

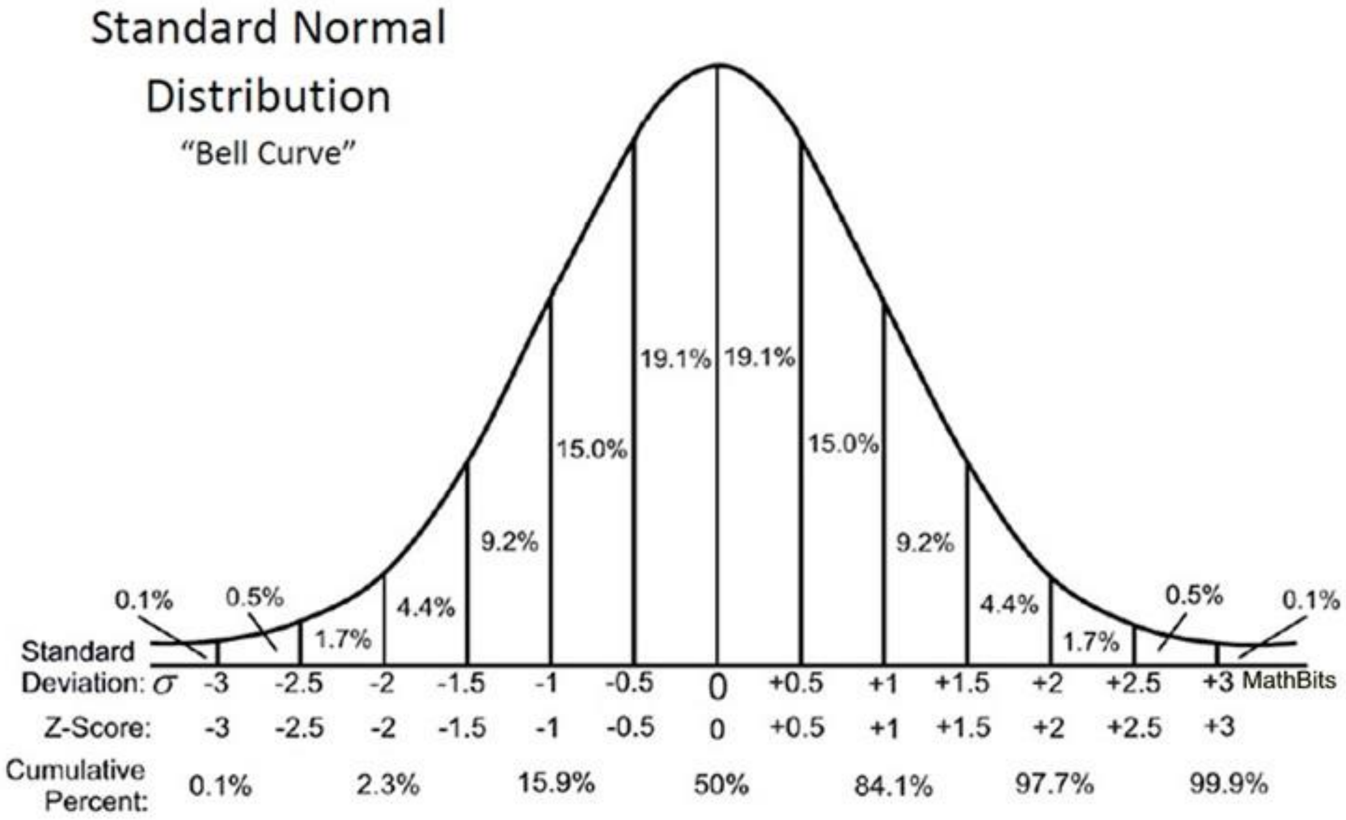


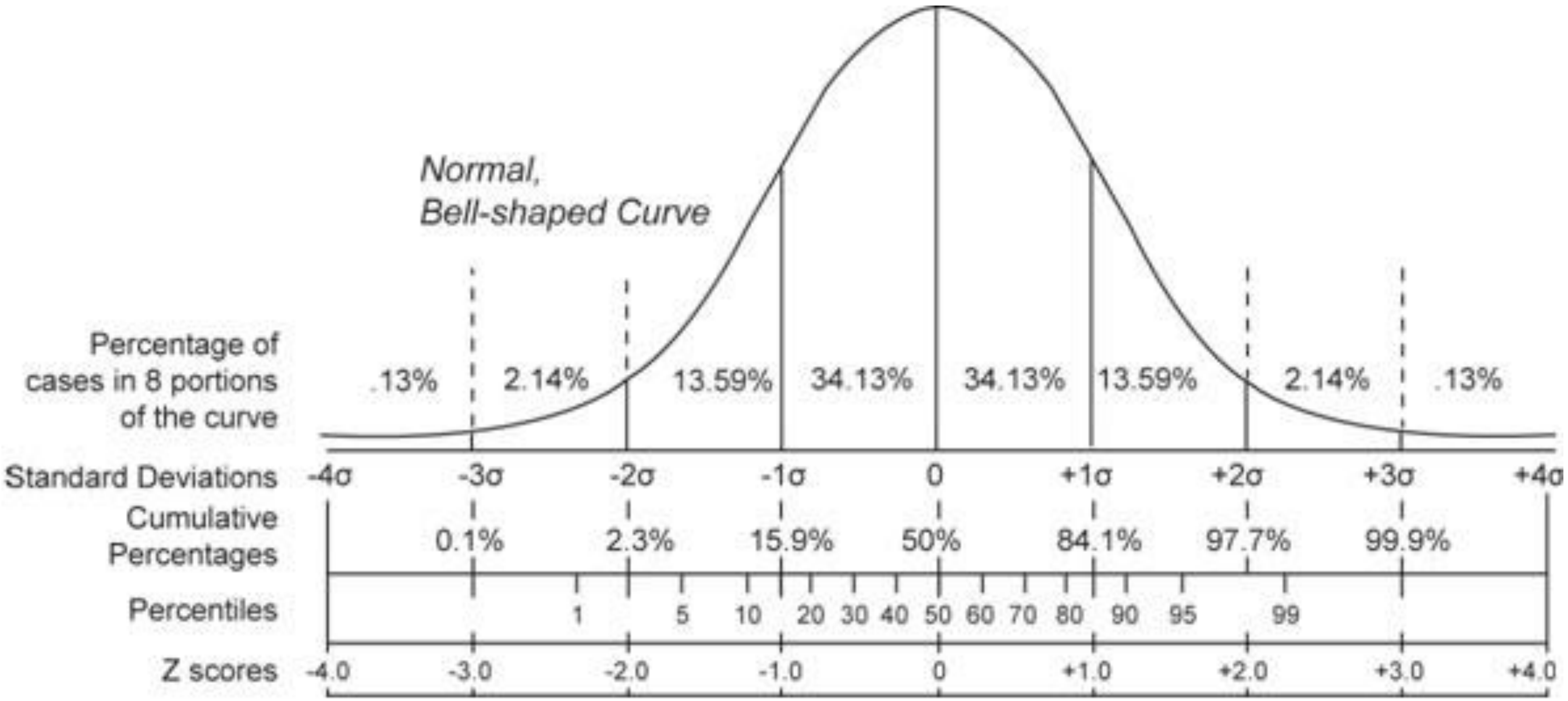
Conditions:

1. $\int f(x)\,dx = 1$

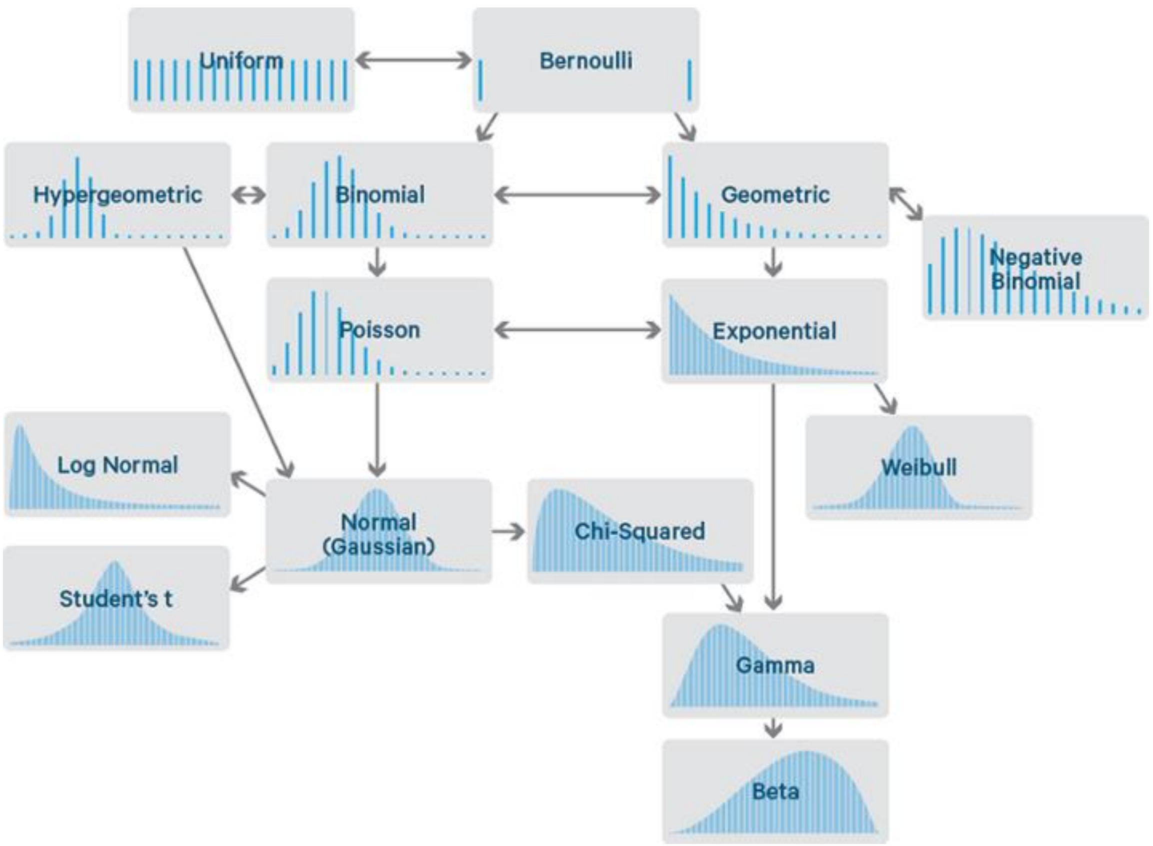
2. $f(x) \geq 0$

GAUSSIAN DISTRIBUTION




Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0

FAMILY OF PROBABILITY DENSITY



https://medium.com/diogo-menezes-borges/introduction-to-statistics-for-data-science-7bf596237ac6

ESTIMATION FRAMEWORK

Parameter θ

A statistical parameter is a quantity characterize probability distribution or statistical model.

Estimator $\hat{\theta}$

An estimator is a quantity as an estimate of a given quantity based on observed data.

Unbiased estimator $\theta = E[\hat{\theta}]$

Consistent estimator $\hat{\theta} \xrightarrow{p} \theta$, as $n \to \infty$

EXAMPLE

Sample Mean $\hat{\mu} = \frac{\sum x_i}{n}$

One can show

$$E[\hat{\mu}] = \mu$$
$$var[\hat{\mu}] = \frac{\sigma_x^2}{\sqrt{n}}$$
$$\hat{\mu} \xrightarrow{p} \mu, \text{ as } n \to \infty$$

One can see that estimator of sample average $\hat{\mu}$ is an unbiased and consistent estimator.

HYPOTHESIS TEST

The t test (also called Student's T Test) compares two averages (means) and tells you if they are different from each other.

The t test also tells you how significant the differences are.

The t score is a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference there is between groups.

A p-value is the probability that the results from your sample data occurred by chance. (Null hypothesis is correct)

KEY STEPS OF TESTING

1. Determine a null and alternative hypothesis (H_0 and H_1)
2. Collect sample data
3. Determine a confidence interval and degrees of freedom
4. Calculate the t-statistics and critical t-value
5. Determine reject the null or not

T-TEST (UNEQUAL SAMPLE SIZES & UNEQUAL VARIANCES)

1. Determine a null and alternative hypothesis (H_0 and H_1)
2. Collect sample data
3. Determine a confidence interval and degrees of freedom

$$df = n_x + n_y - 2$$

4. Calculate the t-statistics and critical t-value

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}} \qquad \text{for } S^2 = \frac{\sum(x-M)^2}{n-1}$$

5. Determine reject the null or not (reject if $|t| > t_\alpha$)

BAYESIAN PROBABILITY

Conditional probability:

$$P(B \text{ and } A) = P(B|A)P(A) \qquad or \qquad P(B) = \frac{P(B \text{ and } A)}{P(A)},$$

where

P(A) and P(B) are the probabilities of observing A and B.

P(A | B) is a conditional probability: the likelihood of event A occurring given that B is true.

P(B|A) is also a conditional probability: the likelihood of event B occurring given that A is true.

Bayesian theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \left(= \frac{P(A \text{ and } B)}{P(B)}\right)$$

https://zh.wikipedia.org/zh-hk/%E6%9D%A1%E4%BB%B6%E6%A6%82%E7%8E%87

EXAMPLE

	Women	Men	
With short hair	25	48	73
With long hair	25	2	27
	50	50	100

P(Woman) = 25/50 = 0.5
P(Man) = 25/50 = 0.5

P(Woman with long hair|Woman) = 25/50 = 0.5
P(Woman with long hair) = P(Woman with long hair|Woman) *P(Woman) = 0.5*0.5 = 0.25

EXAMPLE

$$\mathrm{P(man|long\ hair)} = \frac{\mathrm{P(long\ hair|man)P(man)}}{\mathrm{P(long\ hair)}}$$

$$\mathrm{P(man|long\ hair)} = \frac{\frac{2}{50}*\frac{1}{2}}{\frac{2}{50}*\frac{1}{2}+\frac{25}{50}*\frac{1}{2}}=2/27=0.074$$

Bayesian probability has a wide range of applications in machine learning and statistical analysis. It will be applied in Bayesian classifier in later class.

SUMMARIZE

Optimization -> foundation of regression, SVM, Neural Network and etc…

Probability -> Regression, Hypothesis testing

Bayesian -> Bayesian classifier

SUMMARY

1. Framework of optimization
2. Probability distribution
3. Bayesian probability
4. Least square method

下一課...

迴歸分析及例子:

1. 線性迴歸基礎
2. 線性迴歸的統計特質
3. 非線性迴歸 (Nonlinear regression)
4. 正規化迴歸 (Regularized regression)