Expectation Maximization Algorithm

Overview

- Distribusi normal
- Probabilitas dan likelihood

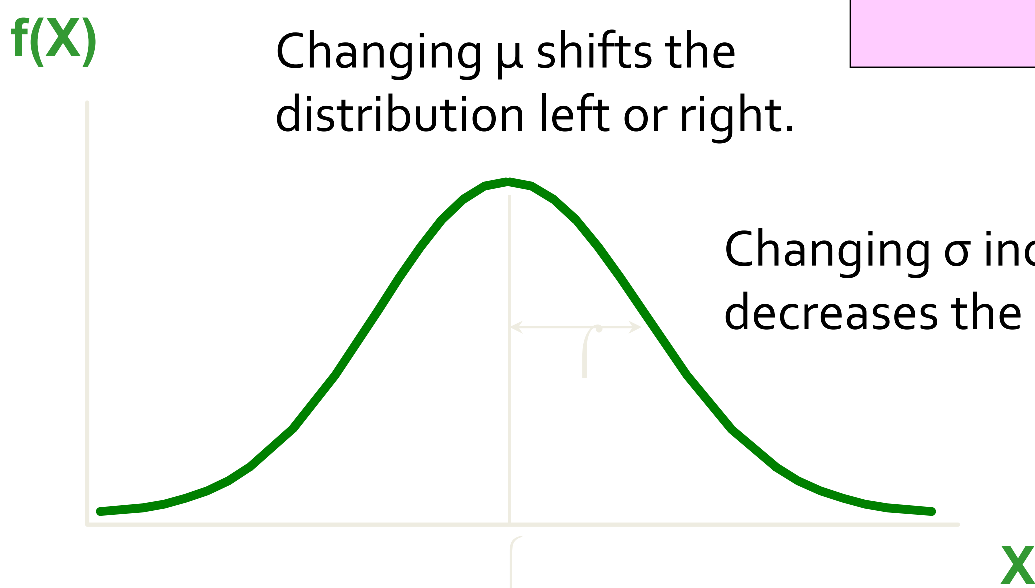
Normal distribution

The Normal Distribution

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

Changing μ shifts the distribution left or right.

Changing σ increases or decreases the spread.



The Normal Distribution: as mathematical function (pdf)

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

This is a bell shaped curve with different centers and spreads depending on μ and σ

<u>Note constants:</u>
π=3.14159
e=2.71828

The Normal PDF

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

It's a probability function, so no matter what the values of μ and σ, must integrate to 1!

Normal distribution is defined by its mean and standard dev.

$E(X)=\mu =$ $$\int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

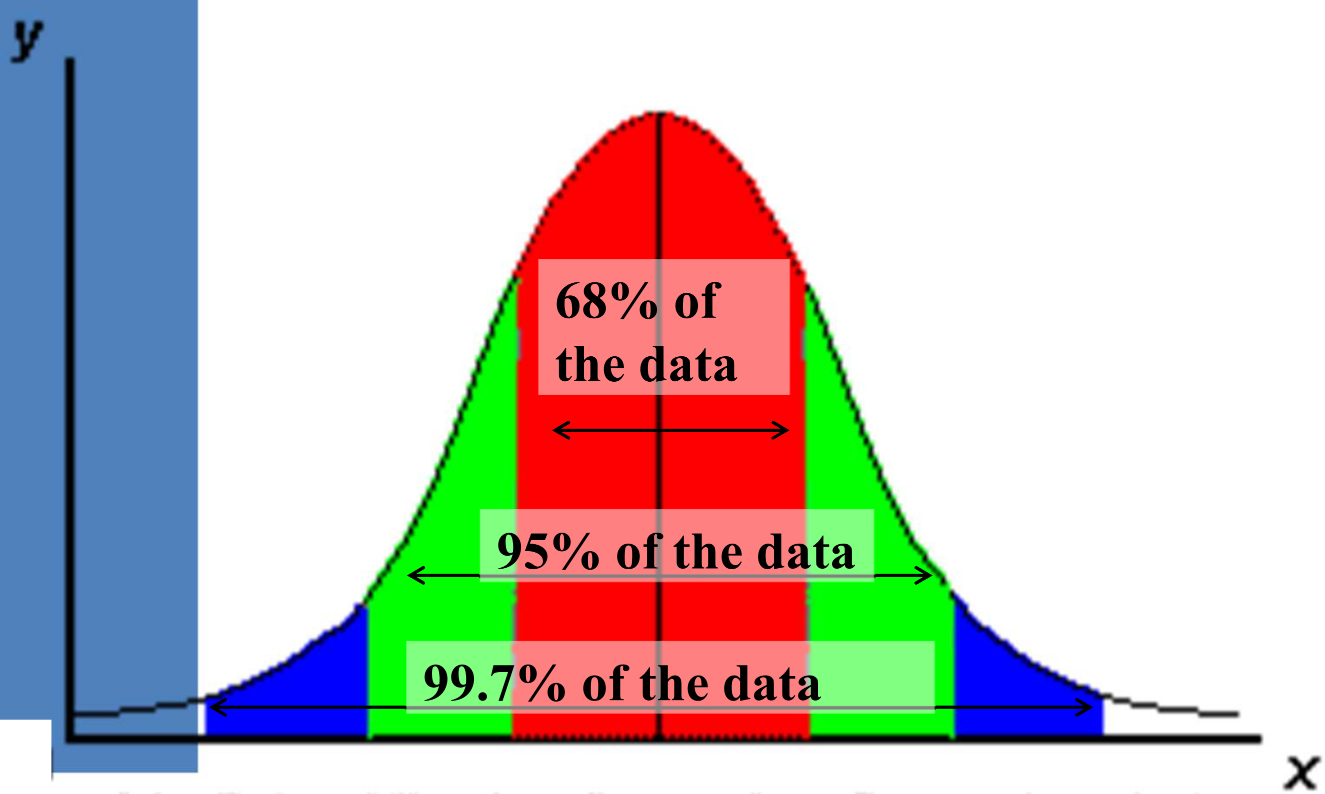
$Var(X)=\sigma^2 =$ $$\int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx) - \mu^2$$

Standard Deviation(X)=σ

**The beauty of the normal curve:

No matter what μ and σ are, the area between $\mu-\sigma$ and $\mu+\sigma$ is about 68%; the area between $\mu-2\sigma$ and $\mu+2\sigma$ is about 95%; and the area between $\mu-3\sigma$ and $\mu+3\sigma$ is about 99.7%. Almost all values fall within 3 standard deviations.

68-95-99.7 Rule



68-95-99.7 Rule in Math terms…

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .68$$

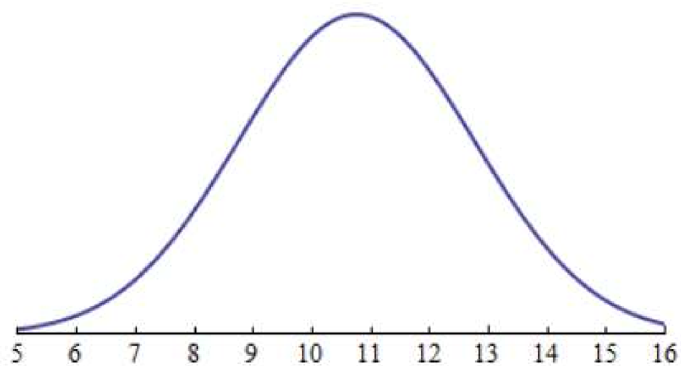
$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .997$$

Probabilitas dan likelihood

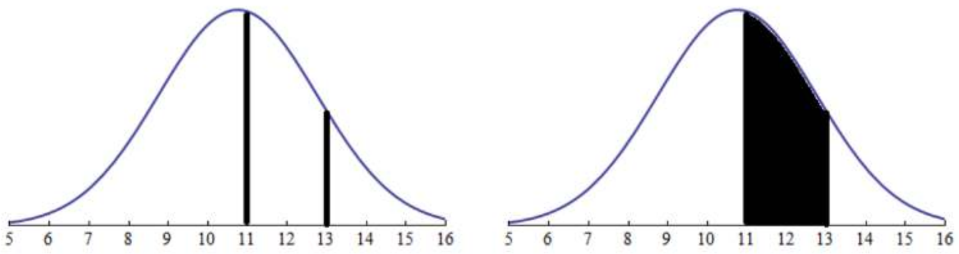
Probabilitas

- **Probability** is the measure of the likelihood that an event will occur.
- The basic idea is out of all given occurrences, what is the certainty that a specific event will occur?
- Let us say we have a normal distribution graph of the average marks of students in a surprise test. (this concept will apply to all continuous distributions)



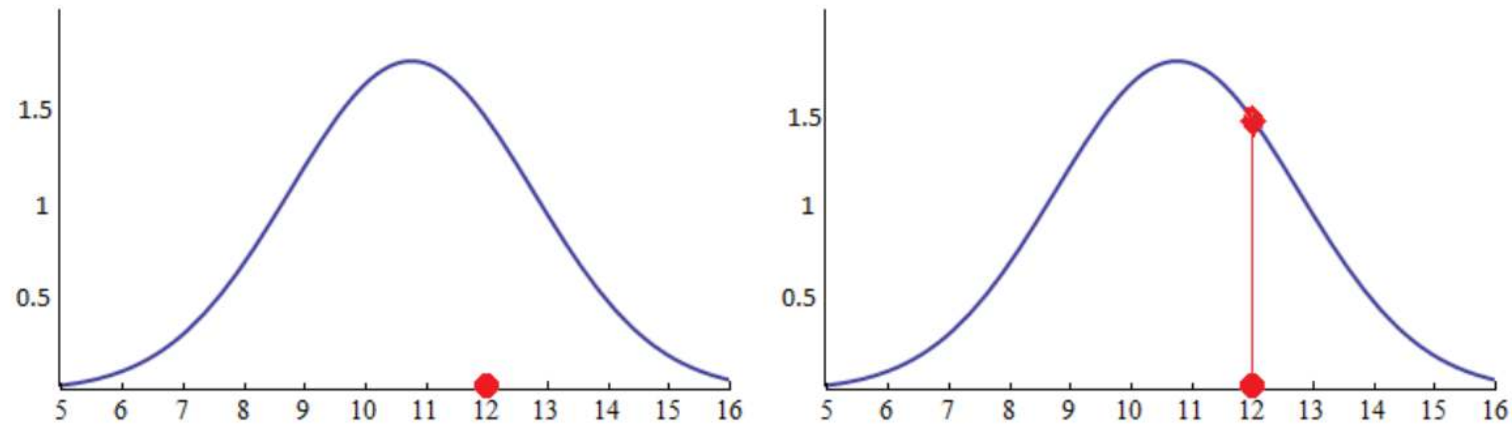
- Now, the probability that a randomly selected student will have marks between 11–13 marks is the area under the curve between those 2 points.
- mathematically,

P(marks between 11 and 13 marks | mean=11 and std = 3) = 0.31



Likelihood

- likelihood function (often simply a likelihood) is a function of parameters within the parameter space that describes the probability of obtaining the observed data.
- **L(mean=11 and std = 3 | student scored 12 marks) = 1.48**



- Probabilities are the areas under fixed distribution

P(data|distribution)

- Likelihoods are the y-axis values for fixed data points with distributions that can be moved.

L(distribution| data)

- Finally, **Probability quantifies anticipation (of outcome), likelihood quantifies trust (in the model).**

Bayesian Theori

Bayes' Theorem

- Bayes' Theorem shows the relationship between a conditional probability and its inverse.

- i.e. it allows us to make an inference from
- the probability of a hypothesis given the evidence to
- the probability of that evidence given the hypothesis
- and vice versa

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- P(A) – the PRIOR PROBABILITY – represents your knowledge about A before you have gathered data.
- e.g. if 0.01 of a population has schizophrenia then the probability that a person drawn at random would have schizophrenia is 0.01

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- P(B|A) – the CONDITIONAL PROBABILITY – the probability of B, given A.
- e.g. you are trying to roll a total of 8 on two dice. What is the probability that you achieve this, given that the first die rolled a 6?

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- So the theorem says:
- The probability of A given B is equal to the probability of B given A, times the prior probability of A, divided by the prior probability of B.

A Simple Example

P(late|car) = 0.5

P(late|bus) = 0.20

P(late|train) = 0.01

Mode of transport:	Probability he is late:
Car	50%
Bus	20%
Train	1%

P(car|late) = ????

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- Suppose that Bob is late one day.
- His boss wishes to estimate the probability that he traveled to work that day by car.
- He does not know which mode of transportation Bob usually uses, so he gives a prior probability of 1 in 3 to each of the three possibilities.

P(car) = 0.33

P(bus) = 0.33

P(train) = 0.33

P(late) = P(late|car)*P(car)+P(late|bus)*P(bus)+P(late|train) *P(train)

A Simple Example

- $P(A|B) = P(B|A)\ P(A)\ /\ P(B)$
- $P(car|late) = P(late|car) \times P(car)\ /\ P(late)$

- $P(late|car) = 0.5$ (he will be late half the time he drives)
- $P(car) = 0.33$ (this is the boss' assumption)
- $P(late) = 0.5 \times 0.33 + 0.2 \times 0.33 + 0.01 \times 0.33$

(all the probabilities that he will be late added together)

- $P(car|late) = 0.5 \times 0.33\ /\ 0.5 \times 0.33 + 0.2 \times 0.33 + 0.01 \times 0.33$

  $= 0.165\ /\ 0.71 \times 0.33$

  $= 0.7042$

EM algorithm

EM algorithm

- The EM algorithm is an iterative optimization method that finds the maximum likelihood estimate (MLE) of parameters in problems where hidden/missing/latent variables are present

K-Means → EM

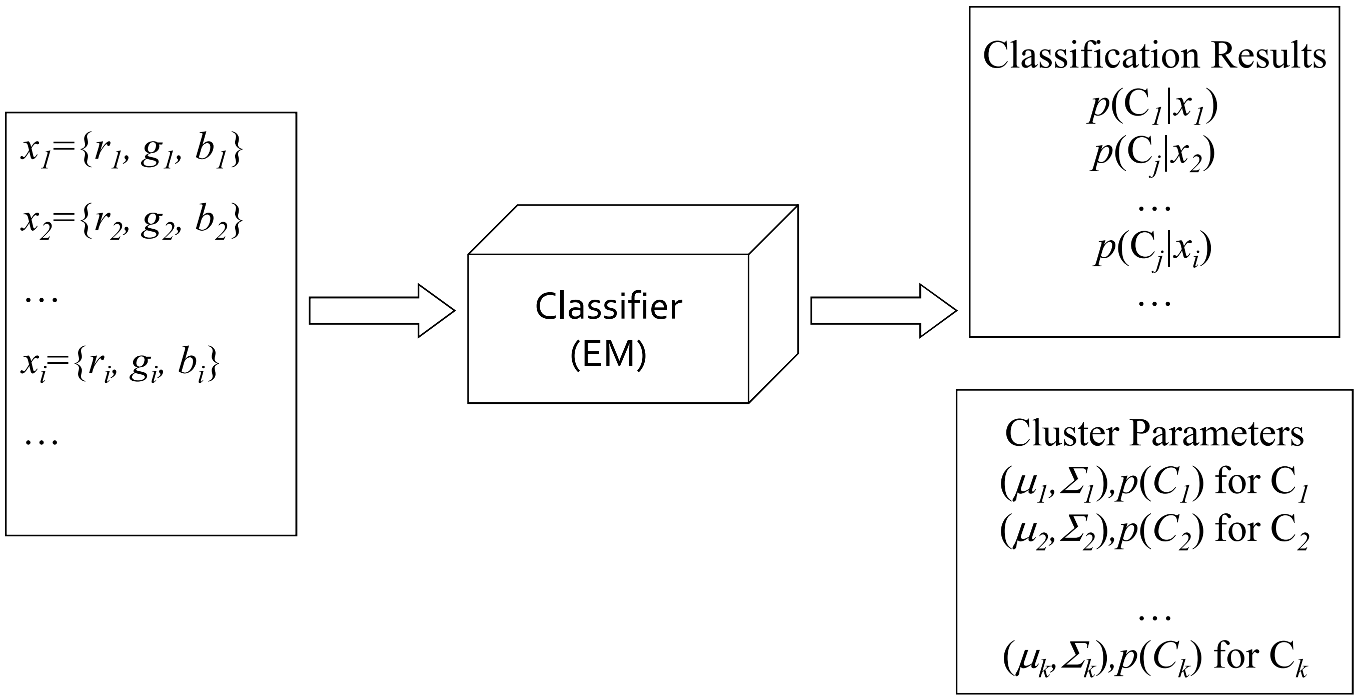
- Boot Step:
  - Initialize K clusters: $C_1, \ldots, C_K$

    $(\mu_{j,} \Sigma_j)$ and $P(C_j)$ for each cluster j.
- Iteration Step:
  - Estimate the cluster of each data

    $p(C_j \mid x_i)$ ➡ Expectation
  - Re-estimate the cluster parameters

$(\mu_j, \Sigma_j), p(C_j)$ For each cluster j ➡ Maximization

EM Classifier



EM Classifier (Cont.)

Input (Known)

$x_1=\{r_1, g_1, b_1\}$

$x_2=\{r_2, g_2, b_2\}$

…

$x_i=\{r_i, g_i, b_i\}$

…

Output (Unknown)

Cluster Parameters

(μ_1, Σ_1), $p(C_1)$ for C_1

(μ_2, Σ_2), $p(C_2)$ for C_2

…

(μ_k, Σ_k), $p(C_k)$ for C_k

Classification Results

$p(C_1|x_1)$

$p(C_j|x_2)$

…

$p(C_j|x_i)$

…

Expectation Step

Input (Known)		Input (Estimation)		Output
$x_1=\{r_1, g_1, b_1\}$ $x_2=\{r_2, g_2, b_2\}$ … $x_i=\{r_i, g_i, b_i\}$ …	+	Cluster Parameters (μ_1, Σ_1), $p(C_1)$ for C_1 (μ_2, Σ_2), $p(C_2)$ for C_2 … (μ_k, Σ_k), $p(C_k)$ for C_k	$\Rightarrow$	Classification Results $p(C_1 \vert x_1)$ $p(C_j \vert x_2)$ … $p(C_j \vert x_i)$ …

$$p(C_j \mid x_i) = \frac{p(x_i \mid C_j) \cdot p(C_j)}{p(x_i)} = \frac{p(x_i \mid C_j) \cdot p(C_j)}{\sum_j p(x_i \mid C_j) \cdot p(C_j)}$$

Maximization Step

Input (Known)

$x_1=\{r_1, g_1, b_1\}$

$x_2=\{r_2, g_2, b_2\}$

$\dots$

$x_i=\{r_i, g_i, b_i\}$

$\dots$

+

Input (Estimation)

Classification Results

$p(C_1|x_1)$

$p(C_j|x_2)$

$\dots$

$p(C_j|x_i)$

$\dots$

$\Rightarrow$

Output

Cluster Parameters

(μ_1, Σ_1), $p(C_1)$ for C_1

(μ_2, Σ_2), $p(C_2)$ for C_2

$\dots$

(μ_k, Σ_k), $p(C_k)$ for C_k

$$\mu_j = \frac{\sum_i p(C_j \mid x_i) \cdot x_i}{\sum_i p(C_j \mid x_i)} \qquad \Sigma_j = \frac{\sum_i p(C_j \mid x_i) \cdot (x_i - \mu_j) \cdot (x_i - \mu_j)^T}{\sum_i p(C_j \mid x_i)} \qquad p(C_j) = \frac{\sum_i p(C_j \mid x_i)}{N}$$

EM Algorithm

- Boot Step:
  - Initialize K clusters: $C_1, \ldots, C_K$

    (μ_j, Σ_j) and $P(C_j)$ for each cluster j.
- Iteration Step:
  - Expectation Step

    $$p(C_j \mid x_i) = \frac{p(x_i \mid C_j) \cdot p(C_j)}{p(x_i)} = \frac{p(x_i \mid C_j) \cdot p(C_j)}{\sum_j p(x_i \mid C_j) \cdot p(C_j)}$$
  - Maximization Step

$$\mu_j = \frac{\sum_i p(C_j \mid x_i) \cdot x_i}{\sum_i p(C_j \mid x_i)} \qquad \Sigma_j = \frac{\sum_i p(C_j \mid x_i) \cdot (x_i - \mu_j) \cdot (x_i - \mu_j)^T}{\sum_i p(C_j \mid x_i)} \qquad p(C_j) = \frac{\sum_i p(C_j \mid x_i)}{N}$$

Data	NIlai
1	2
2	4
3	1
4	5
5	7

Example

- Initialization: K= 2
  - $P(C1)=0.5$
  - $P(C2)=0.5$
  - $\mu\ (C1) = (1,2)$
  - $\mu\ (C2) = (3,4)$
  - $\sum(C1)=(1,1)$
  - $\sum(C2)=(2,2)$

Data	NIlai
1	2
2	4
3	1
4	5
5	7

Example

- P(C1)=0.5
- P(C2)=0.5
- μ (C1) =(2)
- μ (C2) =(5)
- ∑(C1)=(1)
- ∑(C2)=(2)

- Expectation
  - P(C1|data1)= P(data1|C1)*P(C1)/P(data1)
  - P(C2|data1)= P(data1|C1)*P(C1)/P(data1)

  - P(data1|C1)= 0.398

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

Data	Nilai	P(x!C1)	p(x|C2)	P(C1|x)	P(C2|x)		my	sigma		Pc1	Pc2
1	2	0.3989423	0.021024	0.570458	0.046423	1	2	1		0.5	0.5
2	4	0.053991	0.155348	0.077203	0.343024	2	5	2			
3	1	0.2419707	0.003653	0.346	0.008067						
4	5	0.0044318	0.199471	0.006337	0.440452						
5	7	1.487E-06	0.073381	2.13E-06	0.162033						

Data	Nilai	P(x!C1)	p(x|C2)	P(C1|x1)	P(C2|x1)		my	sigma		Pc1	Pc2
1	2	0.548791	0.019931	0.599057	0.036724		1.827428	0.7119		0.2	0.2
2	4	0.0203588	0.209616	0.022224	0.386233		4.809504	1.538041			
3	1	0.3464651	0.002318	0.378199	0.00427						
4	5	0.0004768	0.256341	0.00052	0.472328						
5	7	3.867E-09	0.054513	4.22E-09	0.100444						

Data	Nilai	P(x!C1)	p(x|C2)	P(C1|x1)	P(C2|x1)		my	sigma		Pc1	Pc2
1	2	0.8940706	0.010643	0.606386	0.015714		1.66781	0.386906		0.2	0.2
2	4	0.0009134	0.272406	0.000619	0.402174		4.789857	1.104047			
3	1	0.5794402	0.000541	0.392994	0.000798						
4	5	6.047E-07	0.354191	4.1E-07	0.522919						
5	7	1.138E-16	0.039553	7.72E-17	0.058395						