

# 암 환자 RNA정보를 활용한 암 분류 모델 개발

발표자 : 지용기


# 암 분류 모델 개발을 한 이유와 진행 과정

- 이유 : 딥러닝 알고리즘을 공부하기 위해서
- 진행과정 :
  - 1) 공개된 암환자 유전정보 수집( TCGA ) 및 저장( 빅데이터, HBASE )
  - 2) 유전정보(TXT)을 학습에 적합한 형태로 변환
  - 3) 로직스틱회귀를 활용한 모형 개발 <= 오늘 발표는 여기까지
  - 4) MULTILAYER PERCEPTRON을 활용한 모형 개발
  - 5) DBN을 활용한 모형 개발
  - 6) 여러 가지 알고리즘중에서 최적의 성능을 발휘하는 알고리즘은 ??
- 진행된 내용
  - [HTTPS://GITHUB.COM/BIOSPIN/BIGBIO](https://github.com/BIOSPIN/BIGBIO)



# 공개된 암 환자 유전체 데이터

- [HTTPS://CANCERGENOME.NIH.GOV/](https://cancergenome.nih.gov/) <= 수집한 데이터
- [HTTP://ICGC.ORG/](http://icgc.org/)

**THE CANCER GENOME ATLAS**  
National Cancer Institute  
National Human Genome Research Institute

HomeToolsAbout the DataPublication Guidelines

Home

### TCGA Data Portal

Last updated on June 30th, 2016

The TCGA Data Portal is no longer operational and all TCGA data now resides at the Genomic Data Commons. We have provided a link to the following resource which should help in continuing to access and interpret TCGA data.

- [Genomic Data Commons Home](#)

Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	05/31/16
Adrenocortical carcinoma [ACC]	80	80	05/31/16
Bladder Urothelial Carcinoma [BLCA]	412	412	05/27/16
Brain Lower Grade Glioma [LGG]	516	516	05/02/16
Breast invasive carcinoma [BRCA]	1100	1097	05/31/16
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	307	05/26/16
Cholangiocarcinoma [CHOL]	36	36	05/31/16
Colon adenocarcinoma [COAD]	461	461	05/27/16
Esophageal carcinoma [ESCA]	185	185	05/31/16

# 암 환자 유전체 데이터

- 105개의 STUDY에서 수집한 33가지의 암 데이터
- 데이터 종류
  - 1) CLINICAL DATA ( 55,274개 )
  - 2) IMAGES
  - 3) MICROSATELLITE INSTABILITY (MSI)
  - 4) DNA SEQUENCING
  - 5) MIRNA SEQUENCING
  - 6) PROTEIN EXPRESSION ( 21,871 개)
  - 7) MRNA SEQUENCING
  - 8) TOTAL RNA SEQUENCING
  - 9) ARRAY-BASED EXPRESSION ( 67,994개 )
  - 10) DNA METHYLATION ( 29,939 개)
  - 11) COPY NUMBER ( 23,636 개)

## NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

### TCGA BY THE NUMBERS

TCGA produced over

**2.5**  
PETABYTES  
of data

TCGA data describes

**33**  
DIFFERENT  
TUMOR TYPES

...including

**10**  
RARE  
CANCERS

To put this into perspective, 1 petabyte of data is equal to

**212,000**  
DVDs

...based on paired tumor and normal tissue sets collected from

**11,000**  
PATIENTS

...using

**7**  
DIFFERENT  
DATA TYPES

### TCGA RESULTS & FINDINGS



MOLECULAR  
BASIS OF  
CANCER

Improved our understanding of the genomic underpinnings of cancer



TUMOR  
SUBTYPES

Revolutionized how cancer is classified



THERAPEUTIC  
TARGETS

Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.\*

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

### THE TEAM

**20**  
COLLABORATING  
INSTITUTIONS  
across the United States  
and Canada

### WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.



\*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.



# 암 환자 유전체 데이터 수집

- 모든 암종류의 mRNA 3LEVEL 데이터 양 : 약 350GB
- mRNA 3LEVEL 데이터 구조
  - 7990682D-6A23-47C7-8C8B-32C87061BA10.TAR
  - ACC/
  - BLCA/
  - BRCA/
  - CESC/
  - CHOL/
  - FILE\_ANNOTATIONS.TXT
  - FILE\_MANIFEST.TXT

TCGA Home | Contact Us | For the Media

**NIH THE CANCER GENOME ATLAS**  
National Cancer Institute  
National Human Genome Research Institute

Home Download Data Tools About the Data Publication Guidelines

**File Search Form** [Help](#)

**Project:** The Cancer Genome Atlas **Disease:** SARC - Sarcoma

**Data Category:** mRNA Expression [Help](#) **Data Level:** Level\_3 **Access Tier:** Select access tiers... [Help](#)

**Barcode/UUID:** Enter barcode or uuid separate with comma **Import File:** Select a txt file to import [Browse...](#) **Reset** **Search**

1594 results found

[Export Summary Data](#) [Download All \(1594 files\)](#) [Add to Cart](#) [View / Download Cart](#)

<input type="checkbox"/>	File Name	Barcode	Disease	Data Type	Data Category	Center	Level	File Size	UUID	Submission Date
<input type="checkbox"/>	unc.edu.473deced-b8f8-4f02-86da-335dc0b0de01.2478714.rsem.isoforms.r	TCGA-XG-A8C3-...	SARC	IIluminaHiSeq_RNA...	mRNA Expression	UNC	Lev...	1.320 MB	473deced-b8f8-...	01/28/2015
<input type="checkbox"/>	unc.edu.461292a1-8069-4b4e-840d-8519fb296110.2415745.junction_quant	TCGA-DX-A6YU-...	SARC	IIluminaHiSeq_RNA...	mRNA Expression	UNC	Lev...	8.562 MB	461292a1-8069-...	01/28/2015
<input type="checkbox"/>	unc.edu.349f670f-1447-4bb0-86ec-fba8734d50ff.2428615.rsem.genes.norm	TCGA-FX-A3RE-...	SARC	IIluminaHiSeq_RNA...	mRNA Expression	UNC	Lev...	424.808 KB	349f670f-1447-...	01/28/2015
<input type="checkbox"/>	unc.edu.f8ee0b8d-d627-4b64-bd87-edd8521c1048.2423291.junction_quant	TCGA-HB-A5W3-...	SARC	IIluminaHiSeq_RNA...	mRNA Expression	UNC	Lev...	8.551 MB	f8ee0b8d-d627-...	01/28/2015
<input type="checkbox"/>	unc.edu.3dbec12a-6679-44a6-a51f-eb08c831d0d.2477824.rsem.genes.res	TCGA-DX-A8BT-...	SARC	IIluminaHiSeq_RNA...	mRNA Expression	UNC	Lev...	1.440 MB	3dbec12a-6679-...	01/28/2015
<input type="checkbox"/>	unc.edu.6c5b0478-11ab-4d66-b3bc-bb53854287bc.2417194.rsem.genes.res	TCGA-SG-A6Z7-...	SARC	IIluminaHiSeq_RNA...	mRNA Expression	UNC	Lev...	1.459 MB	6c5b0478-11ab-...	01/28/2015
<input type="checkbox"/>	unc.edu.339a77f4-cb59-48bd-b342-370018be15f2.2435383.rsem.isoforms.r	TCGA-DX-A7EL-...	SARC	IIluminaHiSeq_RNA...	mRNA Expression	UNC	Lev...	2.378 MB	339a77f4-cb59-...	01/28/2015

Page 1 of 32 Per Page 50 Displaying data 1 - 50 of 1594

# MRNA DATA의 구성

- JUNCTION\_QUANTIFICATION.TXT
- RSEM.GENES.RESULTS
- RSEM.GENES.NORMALIZED\_RESULTS
- RSEM.ISOFORMS.NORMALIZED\_RESULTS
- BT.EXON\_QUANTIFICATION.TXT
- XXXXX.RSEM.GENES.NORMALIZED\_RESULTS 구조
  - GENE\_ID          NORMALIZED\_COUNT
  - ? | 729884      0.0000
  - ? | 8225 924.9305
  - A1BG | 1 63.0213
  - A1CF | 29974    0.0000
  - A2BP1 | 54715   0.9268
  - A2LD1 | 87769   116.8211

<https://www.ncbi.nlm.nih.gov/gene/?term=A1BG>



# MRNA DATA를 빅데이터 시스템 저장과 변환

- [HTTPS://GITHUB.COM/BIOSPIN/DEEPBIO/BLOB/MASTER/EXERCISE01/MRNA\\_UPLOAD\\_SCRIPT.IPYNB](https://github.com/biospin/deepbio/blob/master/exercise01/mrna_upload_script.ipynb)
- [HTTPS://GITHUB.COM/BIOSPIN/DEEPBIO/BLOB/MASTER/EXERCISE01/MRNA\\_MAKE\\_FEATURE.IPYNB](https://github.com/biospin/deepbio/blob/master/exercise01/mrna_make_feature.ipynb)
- TRAINING용, VALIDATION용, TEST용 데이터셋으로 압축하여 파일로 만듦.
- [HTTPS://DRIVE.GOOGLE.COM/OPEN?ID=0B6BSLTLVNAGfN2DIZ0P1OTFTYzG](https://drive.google.com/open?id=0B6BSLTLVNAGfN2DIZ0P1OTFTYzG)
  - MRNA\_20160125-200855\_TYPE1\_00.PKL.GZ
  - MRNA\_20160125-200855\_TYPE1\_01.PKL.GZ
  - MRNA\_20160125-200855\_TYPE1\_02.PKL.GZ
  - MRNA\_20160125-200855\_TYPE1\_03.PKL.GZ
  - .....

# EDA를 통한 데이터 문제점 파악

```
# 각각의 데이터셋에서 feature와 label을 구분
```

```
train_x, train_y = TrainSet
```

```
validation_x, validation_y = ValidationSet
```

```
test_x, test_y = TestSet
```

```
# 각각의 size 출력
```

```
print train_x.shape ; print train_y.shape
```

```
print validation_x.shape ; print validation_y.shape
```

```
print test_x.shape ; test_y.shape
```

```
(7945, 20502)
```

```
(7945,)
```

```
(1679, 20502)
```

```
(1679,)
```

```
(1679, 20502)
```

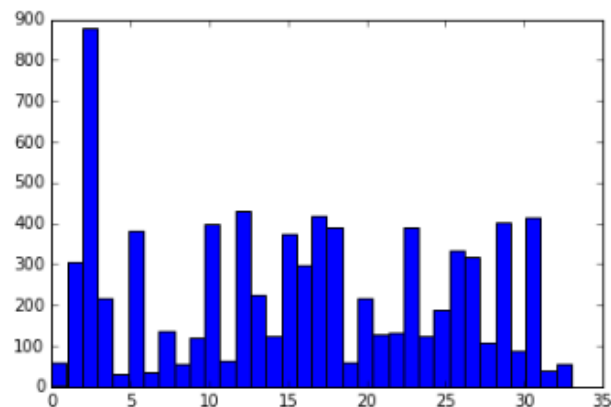
```
(1679,)
```

```
print "type of cancer:", np.unique(train_y)
```

```
plt.hist(train_y, bins=34)
```

```
plt.show()
```

```
type of cancer: [ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24  
25 26 27 28 29 30 31 32 33]
```

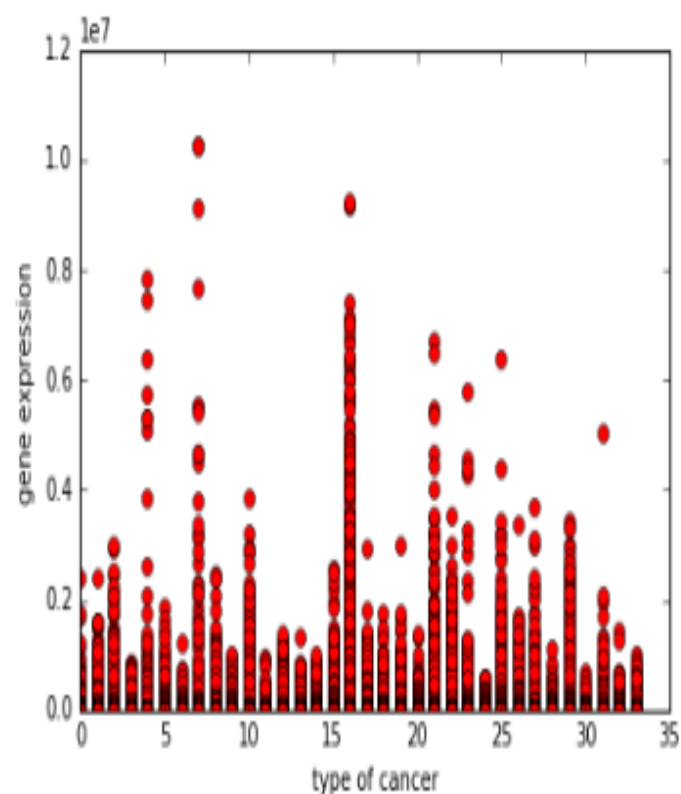


```
plt.plot(train_y, train_x, 'ro')
```

```
plt.xlabel('type of cancer')
```

```
plt.ylabel('gene expression')
```

```
plt.show()
```





# EDA를 통한 데이터 문제점 파악

```
# 각각의 데이터셋에서 feature와 label을 구분
train_x, train_y = TrainSet
validation_x, validation_y = ValidationSet
test_x, test_y = TestSet
```

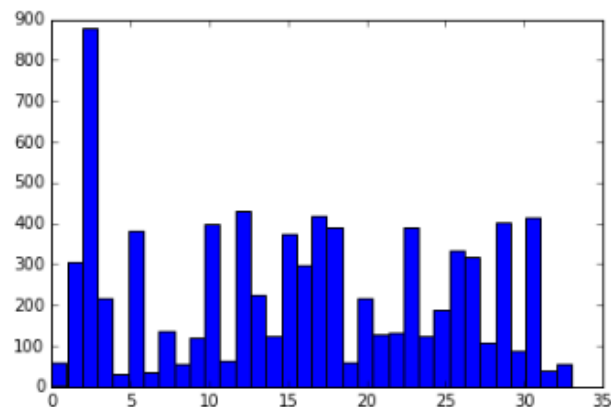
```
# 각각의 size 출력
```

```
print train_x.shape ; print train_y.shape
print validation_x.shape ; print validation_y.shape
print test_x.shape ; test_y.shape
```

```
(7945, 20502)
(7945,)
(1679, 20502)
(1679,)
(1679, 20502)
(1679,)
```

```
print "type of cancer:", np.unique(train_y)
plt.hist(train_y, bins=34)
plt.show()
```

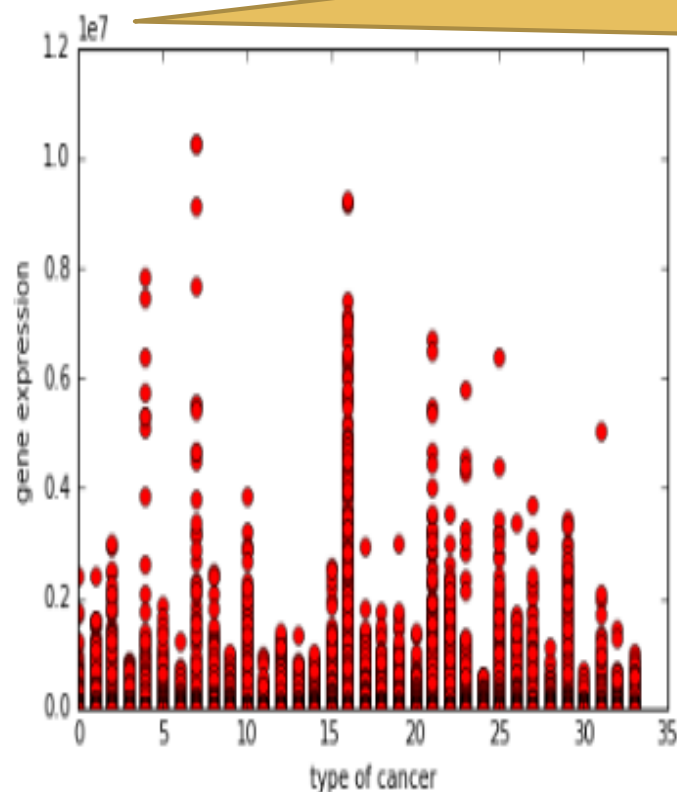
```
type of cancer: [ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
 25 26 27 28 29 30 31 32 33]
```



7945 x  
20502 행렬  
=>  
glm()로는  
처리 불가

암종류별  
샘플수의  
차이가 많이  
발생

```
plt.plot(train_y, train_x, 'ro')
plt.xlabel('type of cancer')
plt.ylabel('gene expression')
plt.show()
```



스케일의  
변차가 너무  
크고, 극단적인  
이상치가 있음.

# 7945 X 20502 행렬에서 B(베타)값 구하기

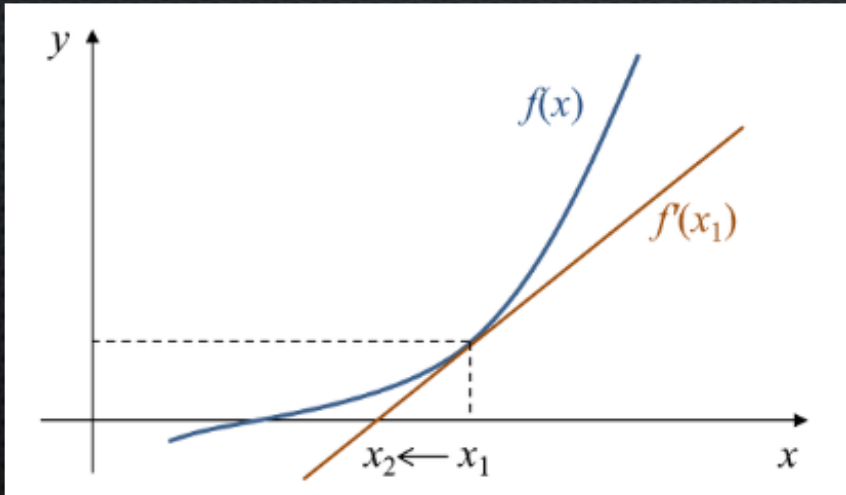
- 뉴턴-랩슨법(NEWTON-RAPHSON METHOD)

- 아래와 같이  $x$ 에 대한 7차 방정식의 해는??

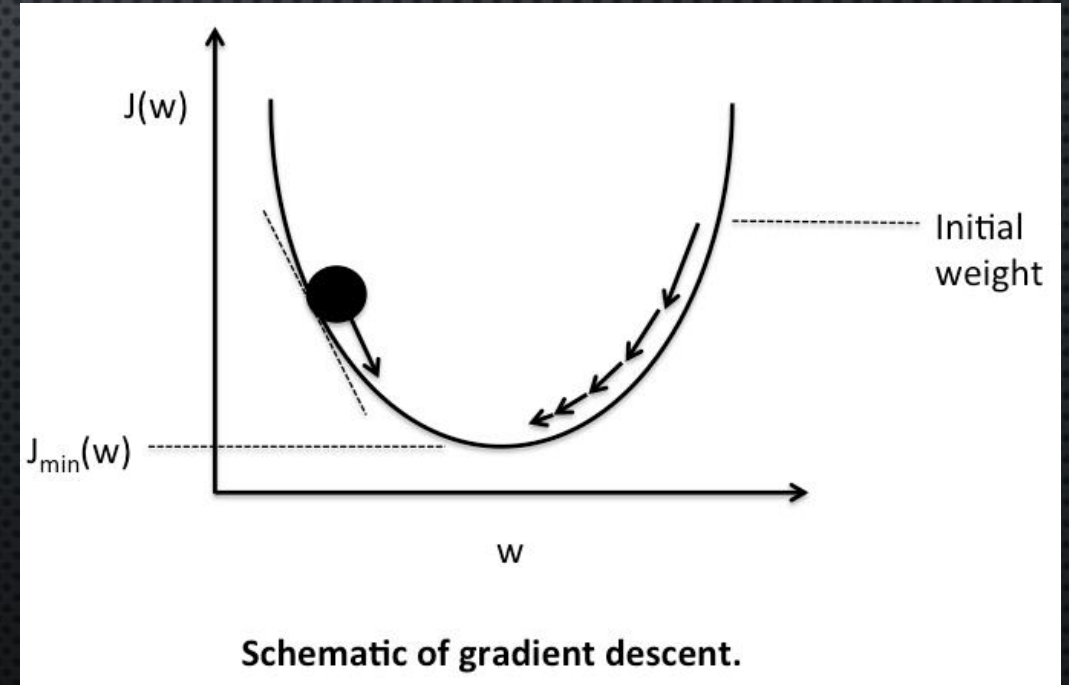
$$f(x)=x^7-2x^6+x^5+7x^2-3x+11=0$$

- 임의의  $x_1$ 에서 접선의 기울기(미분)해서

- 기울기가 양수 일때 접점은 ( 왼쪽 , 오른쪽 ) ??
- 기울기가 음수 일때 접점은 ( 왼쪽 , 오른쪽 ) ??
- 기울기가 클때는 접점이 ( 가까움, 널리 있음 ) ??
- 기울기가 작을때는 접점이 ( 가까움, 널리 있음 ) ??

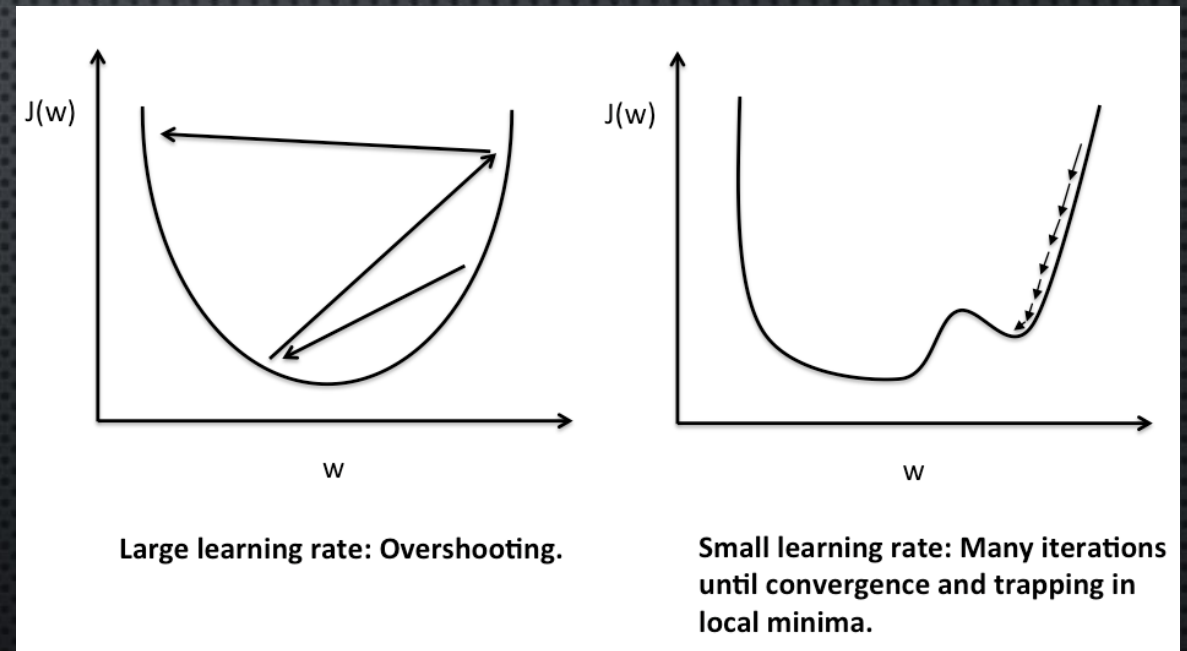
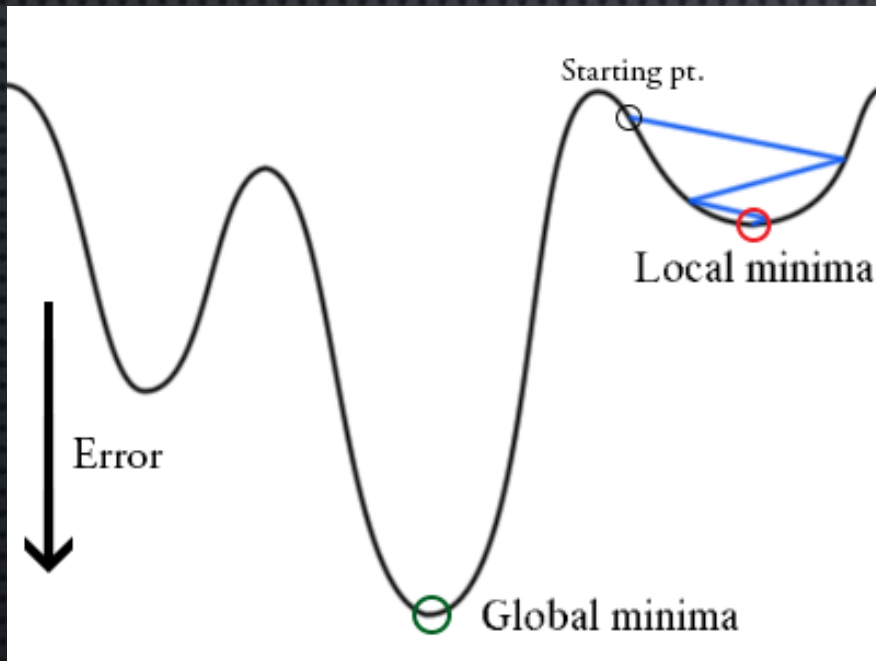


- 확률적 기울기 하강(STOCHASTIC GRADIENT DESCENT)





# 확률적 기울기 하강(STOCHASTIC GRADIENT DESCENT)



스케일의 변차가 너무 크고, 극단적인 이상치 문제

- 정규화(NORMALIZATION )

$$\frac{X - \mu}{\sigma}$$

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$



## 암종류별 샘플수의 차이가 많이 발생

- 샘플수가 적은 암종류는 제거
- 샘플수가 많은 암종류는 일부만 추출
- 실습에는 모든 데이터 사용함

## 실습 코드

- [https://github.com/biospin/bigbio/blob/master/part03/week03\\_160517/tensorflow%ED%99%9C%EC%9A%A9%20%EC%95%94%EC%A2%85%EB%A5%98%20%EC%98%88%EC%B8%A1\\_%EC%B5%9C%EC%A2%85.ipynb](https://github.com/biospin/bigbio/blob/master/part03/week03_160517/tensorflow%ED%99%9C%EC%9A%A9%20%EC%95%94%EC%A2%85%EB%A5%98%20%EC%98%88%EC%B8%A1_%EC%B5%9C%EC%A2%85.ipynb)
- 참조 : <https://www.tensorflow.org/versions/r0.12/tutorials/mnist/beginners/index.html#softmax-regressions>