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

공개된 암 환자 유전체 데이터

- HTTPS://CANCERGENOME.NIH.GOV/ <= 수집한 데이터
- HTTP://ICGC.ORG/

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

Home | Tools | About the Data | Publication 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

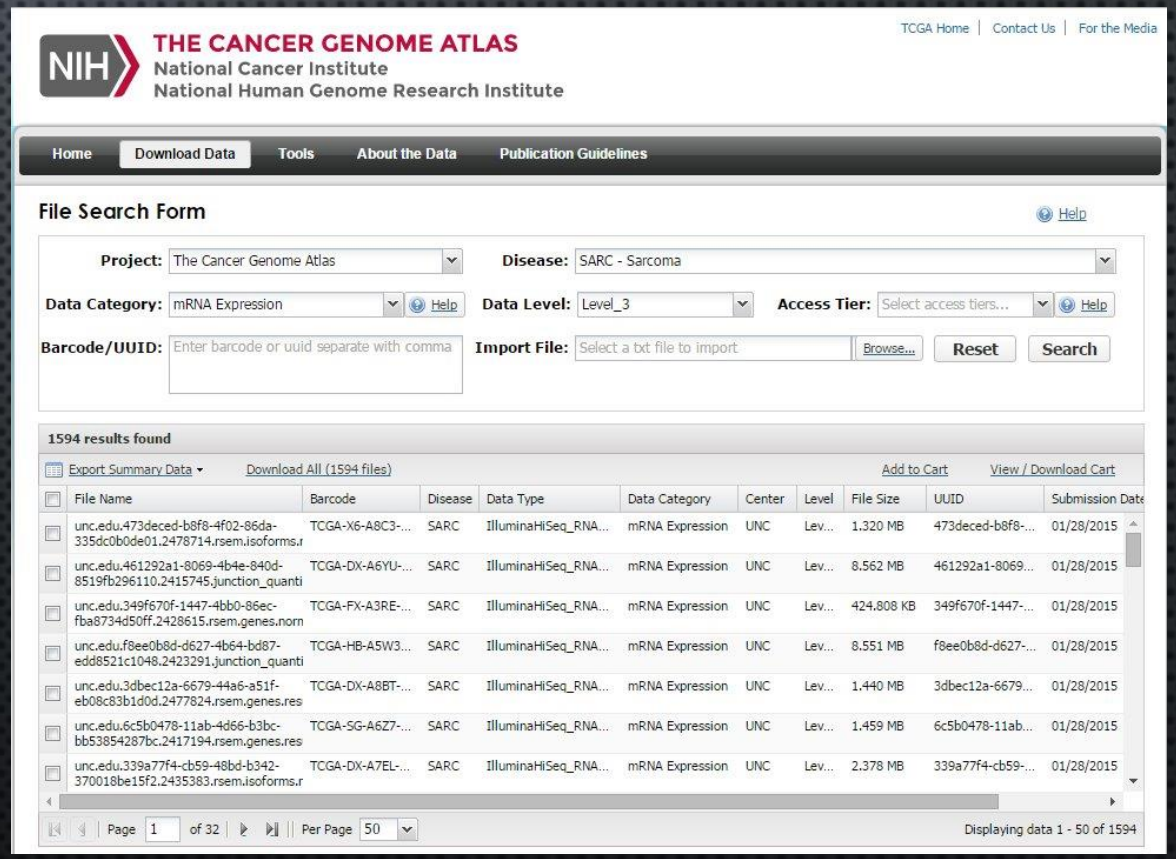
Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	05/31/16
Adrenocortical carcinoma [ACC]	80	80	05/31/16
Bladder Urothelial Carcinoma [BLCA]	412	412	05/27/16
Brain Lower Grade Glioma [LGG]	516	516	05/02/16
Breast invasive 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)

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



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_make_feature.ipynb
- Training용, Validation용, Test용 데이터셋으로 압축하여 파일로 만듬.
- 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
  - ……

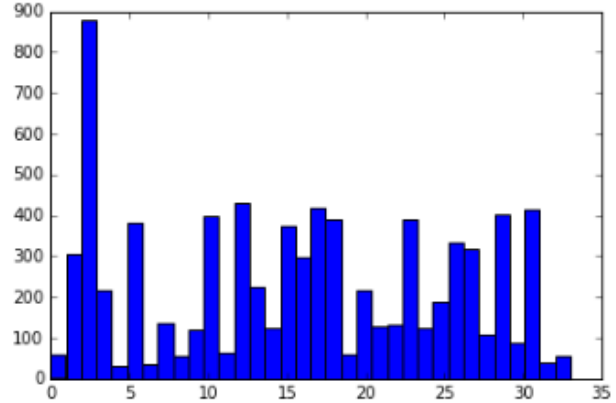
EDA를 통한 데이터 문제점 파악

```
# 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
```

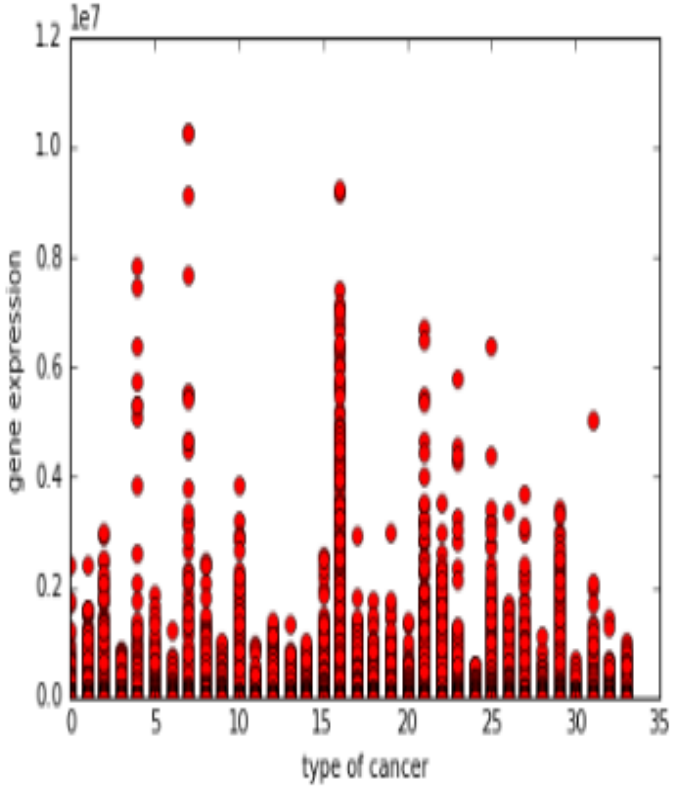
```
(7945, 20502)
(7945,)
(1679, 20502)
(1679,)
(1679, 20502)
(1679,)
```

```
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

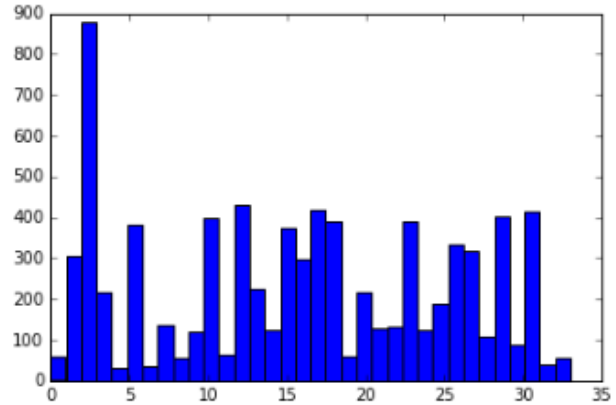
```
# 각각의 데이터셋에서 feature와 label을 구분
train_x, train_y = TrainSet
validation_x, validation_y = ValidationSet
test_x, test_y= TestSet

# 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,)
```

7945 x 20502 행렬 => glm()로는 처리 불가

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

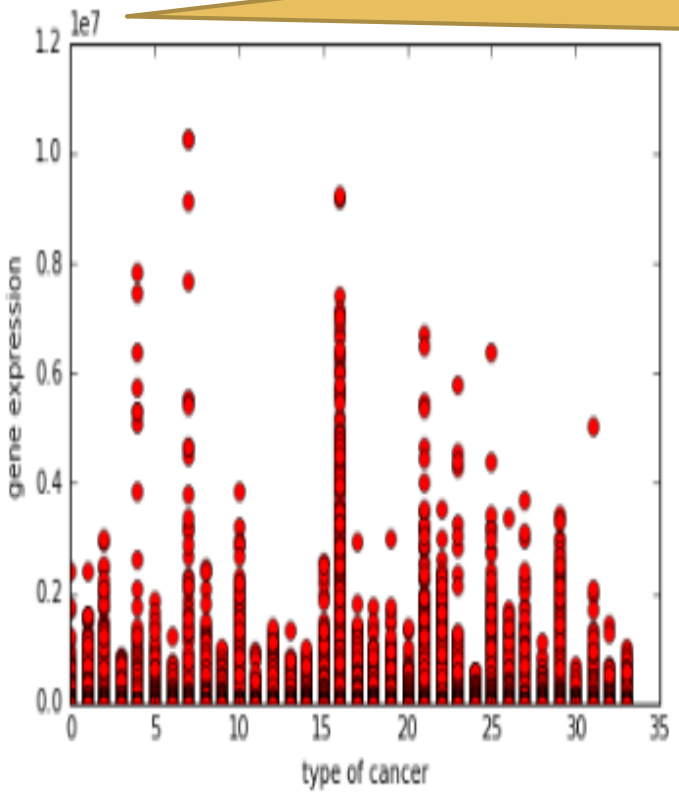
```
type of cancer: [ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
 25 26 27 28 29 30 31 32 33]
```



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.

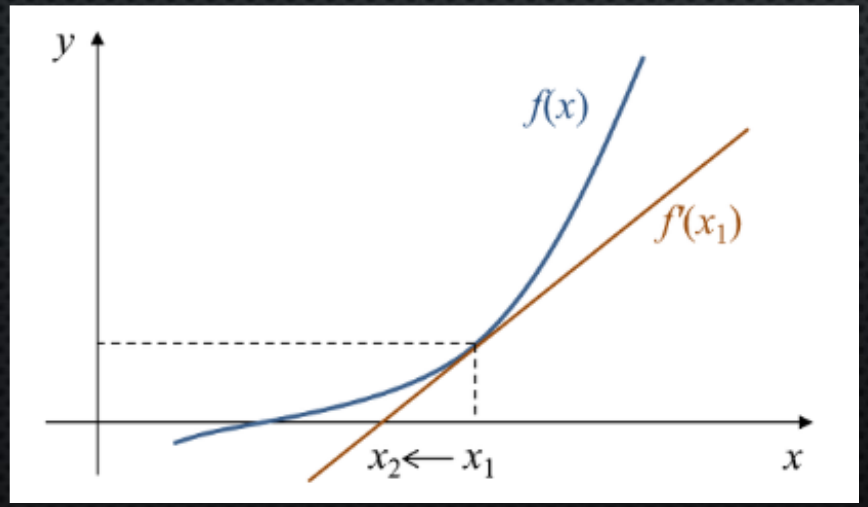


7945 X 20502 행렬에서 B(베타)값 구하기

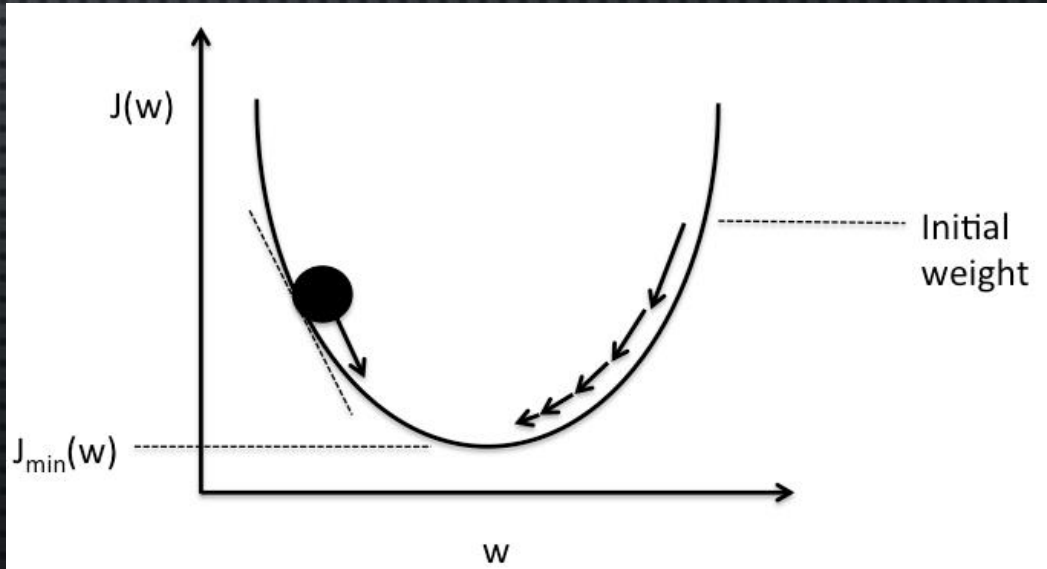
- 뉴턴-랩슨법(NEWTON-RAPHSON METHOD)
  - 아래와 같이 x에 대한 7차 방정식의 해는??

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

  - 임의의 x1에서 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) ??

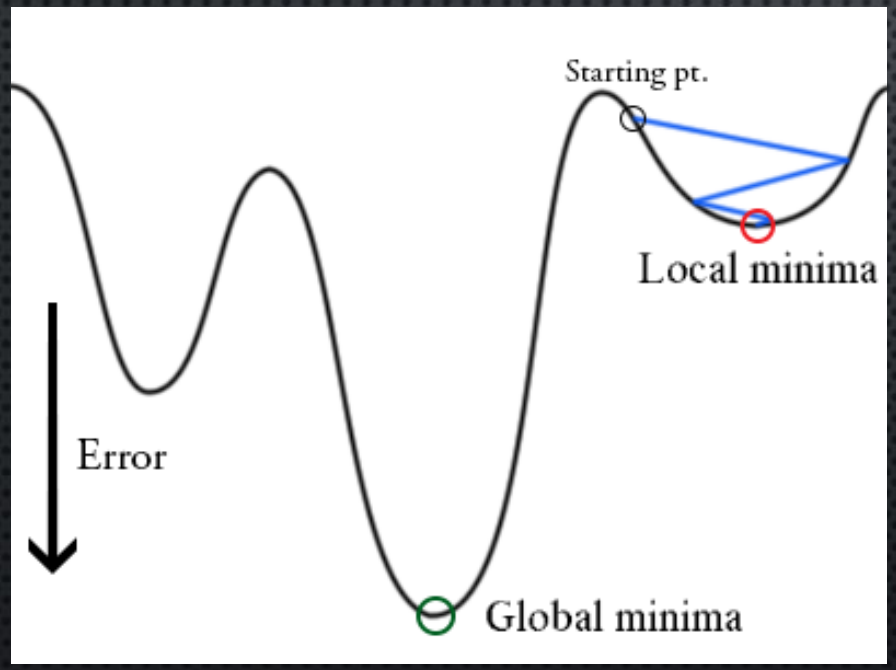


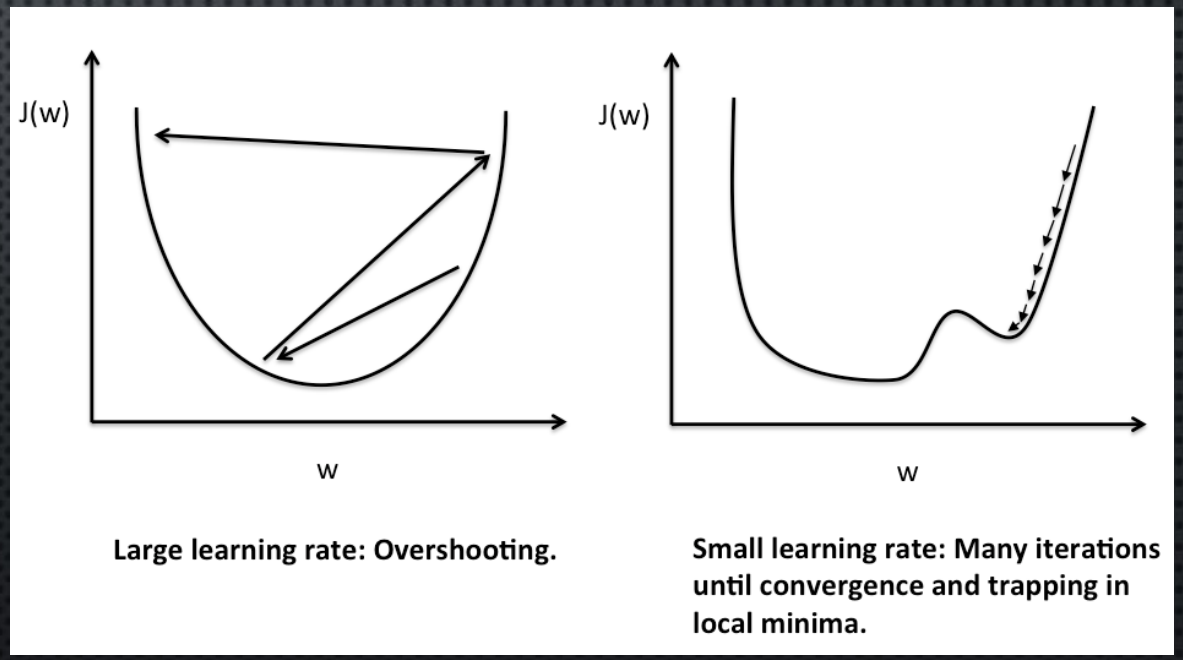
- 확률적 기울기 하강(STOCHASTIC GRADIENT DESCENT)



Schematic of 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

- 참조 : https://www.tensorflow.org/versions/r0.12/tutorials/mnist/beginners/index.html#softmax-regressions