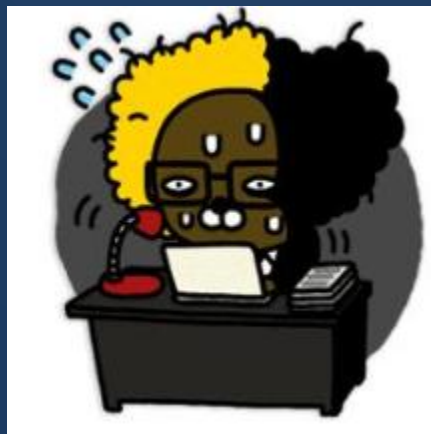


San Francisco Crime Classification

(캐글뽀개기 7월)

Machine Learning & Visualization



PT: 이상열

2015. 07. 18

Contents

Machine Learning & Visualization



Part1

소개

Part2

시각화

Part3

기계학습

Part 1

소개 - Kaggle

사이트 안내

캐글뽀개기 페이스북

(<https://www.facebook.com/#!/groups/kagglebreak/>)

캐글뽀개기 스터디자료

(<http://kagglebreak.github.io/>)

캐글뽀개기 Github 주소

(<https://github.com/KaggleBreak/problems>)

캐글

(<https://www.kaggle.com/>)

Kaggle is a platform for [predictive modelling](#) and [analytics](#) competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. This [crowdsourcing](#) approach relies on the fact that there are countless strategies that can be applied to any predictive modelling task and it is impossible to know at the outset which technique or analyst will be most effective.

Part 1 소개 - 주제



San Francisco Crime Classification

Predict the category of crimes that occurred in the city by the bay



알카트라즈 섬

알카트라즈 섬은 연방 주정부의 형무소로 쓰였던 곳으로 한번 들어가면 절대 나올 수 없다고 해서 '악마의 섬'이라는 별칭이 붙은 곳이다. (샌프란시스코에 있는 섬)

Part 1 소개 - 공공데이터

SF OpenData (<https://data.sfgov.org/>)

SF OpenData

About

Data

Developers

Showcase

Help

SFPD Incidents - Current Year (2015)

Based on SFPD Incidents - from 1 January 2003

Incidents derived from SFPD Crime Incident Reporting system Updated daily, showing 22

Manage

More Views

Filter

Visualize

Export

Discuss

Incidents

Find in this Dataset

Conditional Formatting

Sort & Roll-Up

Filter

Filter this dataset based on contents.

You are in simplified mode. Go advanced now

Date is between

and

IncidentNum is

Category is

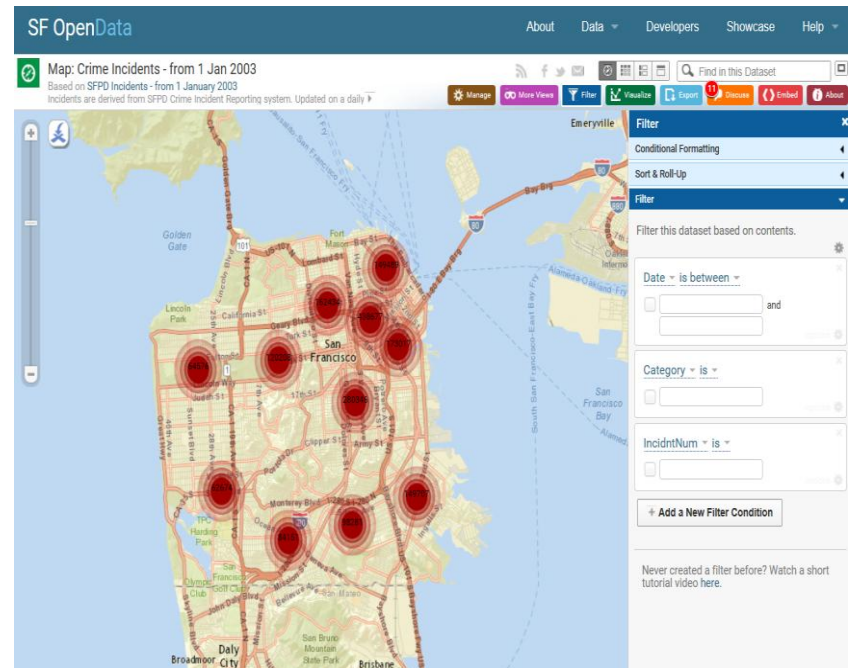
With the following base filters

Date is after

01/01/2015

Never created a filter before? Watch a short tutorial video here.

IncidentNum	Category	Descript	DayOfWeek	Date	Time	PdDistrict
150575042	VANDALISM	MALICIOUS MISCHIEF, VANDALISM	Wednesday	07/01/2015	23:35	INGLESIDE
150575008	WARRANTS	WARRANT ARREST	Wednesday	07/01/2015	23:31	PARK
150575008	OTHER OFFENSES	TRAFFIC COLLISION, HIT & RUN, PRO	Wednesday	07/01/2015	23:31	PARK
150575008	OTHER OFFENSES	TRAFFIC VIOLATION	Wednesday	07/01/2015	23:31	PARK
150575020	OTHER OFFENSES	VIOLATION OF RESTRAINING ORDER	Wednesday	07/01/2015	23:30	BAYVIEW
150575412	VEHICLE THEFT	STOLEN TRUCK	Wednesday	07/01/2015	23:30	NORTHERN
150574953	WARRANTS	ENROUTE TO PAROLE OFFICER	Wednesday	07/01/2015	23:16	BAYVIEW
156158838	VANDALISM	MALICIOUS MISCHIEF, VANDALISM	Wednesday	07/01/2015	23:15	NORTHERN
150575387	NON-CRIMINAL	DEATH REPORT, CAUSE UNKNOWN	Wednesday	07/01/2015	23:00	NORTHERN
150575735	VEHICLE THEFT	STOLEN AUTOMOBILE	Wednesday	07/01/2015	23:00	TARAVAL
150575149	VEHICLE THEFT	STOLEN MOTORCYCLE	Wednesday	07/01/2015	23:00	MISSION
150574903	OTHER OFFENSES	DRIVERS LICENSE, SUSPENDED OR R	Wednesday	07/01/2015	22:55	NORTHERN
150574925	NON-CRIMINAL	AIDED CASE, MENTAL DISTURBED	Wednesday	07/01/2015	22:50	TENDERLOIN
150574975	ASSAULT	BATTERY WITH SERIOUS INJURIES	Wednesday	07/01/2015	22:45	BAYVIEW
150574975	WARRANTS	ENROUTE TO DEPARTMENT OF CORF	Wednesday	07/01/2015	22:45	BAYVIEW
150574975	ROBBERY	ATTEMPTED ROBBERY WITH BODILY	Wednesday	07/01/2015	22:45	BAYVIEW
150576379	VANDALISM	MALICIOUS MISCHIEF, VANDALISM C	Wednesday	07/01/2015	22:45	SOUTHERN
150574878	OTHER OFFENSES	DRIVERS LICENSE, SUSPENDED OR R	Wednesday	07/01/2015	22:40	PARK
150574862	ASSAULT	AGGRAVATED ASSAULT WITH A DEA	Wednesday	07/01/2015	22:35	INGLESIDE
150575666	LARCENY/THEFT	PETTY THEFT OF PROPERTY	Wednesday	07/01/2015	22:30	PARK
150576153	VEHICLE THEFT	STOLEN TRUCK	Wednesday	07/01/2015	22:30	TARAVAL
150574812	SUSPICIOUS OCC	SUSPICIOUS OCCURRENCE	Wednesday	07/01/2015	22:27	SOUTHERN



<https://data.sfgov.org/Public-Safety/Map-Crime-Incidents-from-1-Jan-2003/gxxq-x39z>

Part 1

소개 – 데이터 필드

test.csv (884262 obs, 86,7MB)

sampleSubmission.csv

train.csv (878049 obs, 121MB)

The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set.

Data fields

- **Dates** – timestamp of the crime incident
- **Category** – category of the crime incident (only in train.csv). **This is the target variable you are going to predict.**
- **Descript** – detailed description of the crime incident (only in train.csv)
- **DayOfWeek** – the day of the week
- **PdDistrict** – name of the Police Department District
- **Resolution** – how the crime incident was resolved (only in train.csv)
- **Address** – the approximate street address of the crime incident
- **X** – Longitude
- **Y** – Latitude

Part 1 소개 - 평가 방법

$$\text{logloss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}),$$

<https://www.kaggle.com/c/sf-crime/details/evaluation>

Submissions are evaluated using the multi-class logarithmic loss

$$-\frac{1}{3} * (0 * \log(1/3) + 1 * \log(1/3) + 0 * \log(1/3)) = 0.3662$$

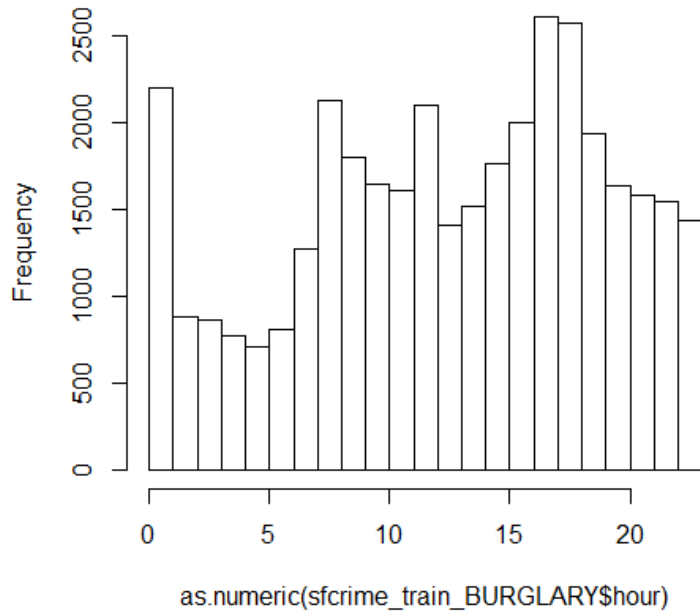
Part 1

데이터 로딩 및 정제

각자 데이터 로딩 및 정제하는 시간을 갖겠습니다. (30분 정도)

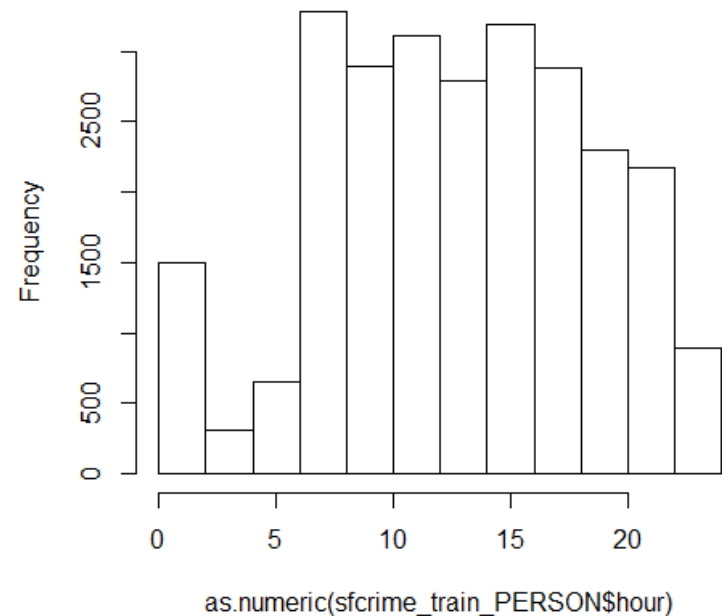
Part 2 시각화 - 시간별 범죄수

Histogram of as.numeric(sfcrime_train_BURGLARY\$hour)



주거 침입

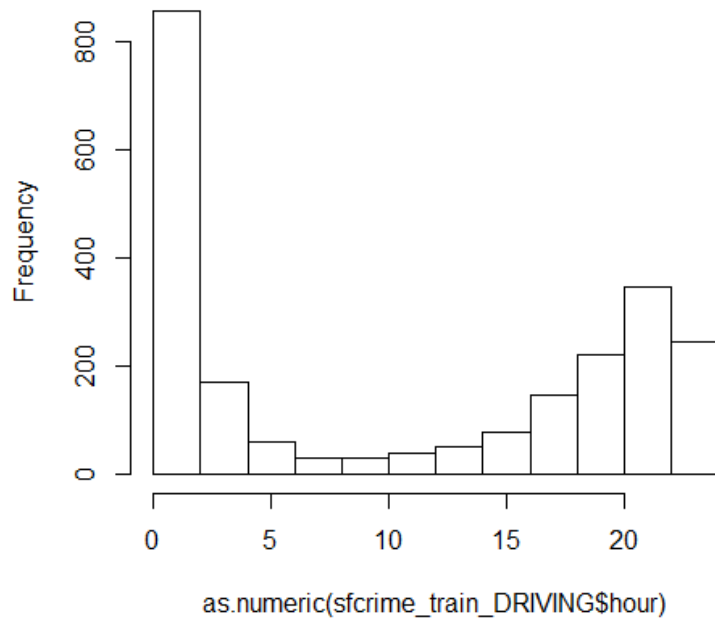
Histogram of as.numeric(sfcrime_train_PERSON\$hour)



실종 사건

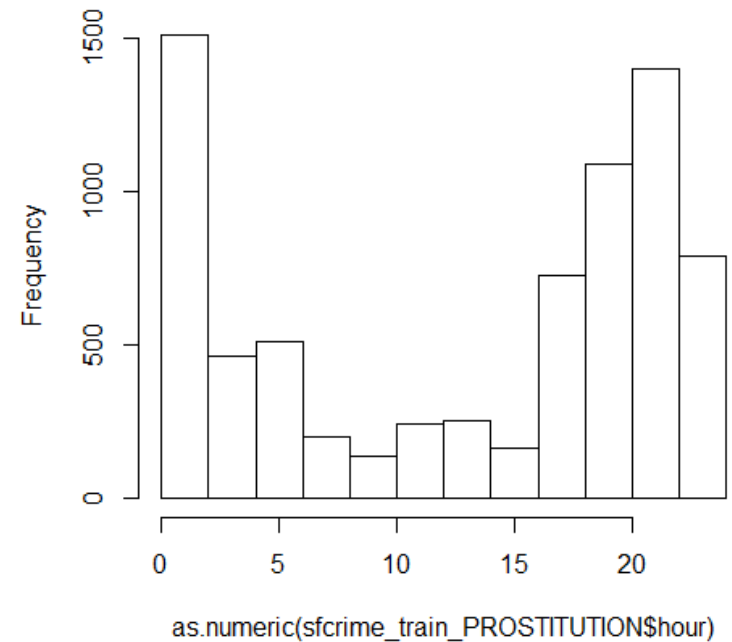
Part 2 시각화 - 시간별 범죄수

Histogram of as.numeric(sfcrime_train_DRIVING\$hour)



음주 운전

Histogram of as.numeric(sfcrime_train_PROSTITUTION\$hour)



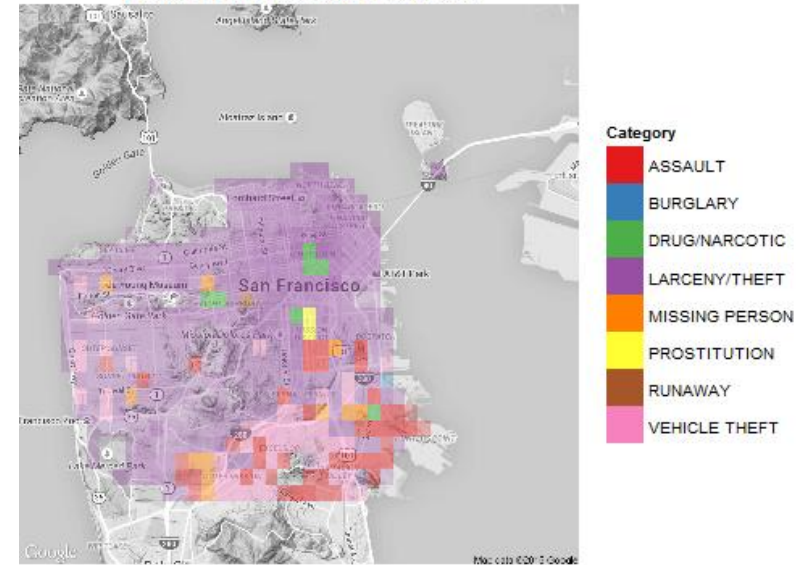
성매매

Part 2 시각화 - 지구별 범죄

Top Crimes in San Francisco

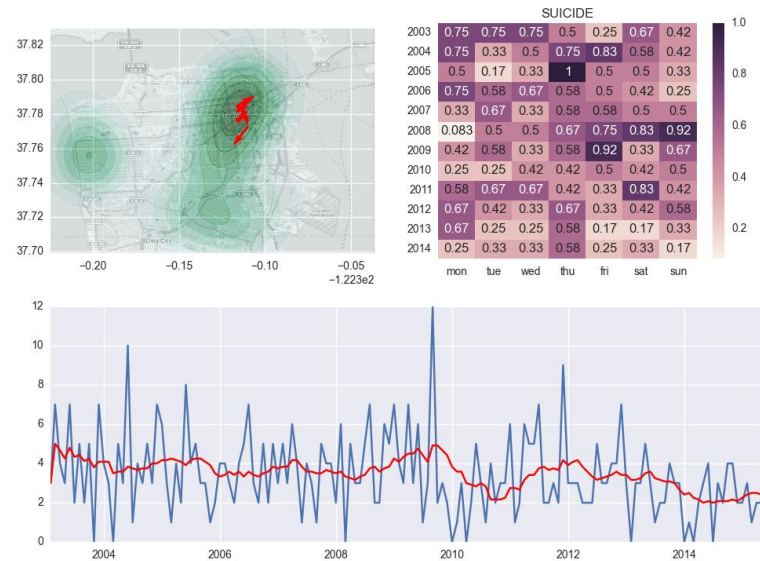
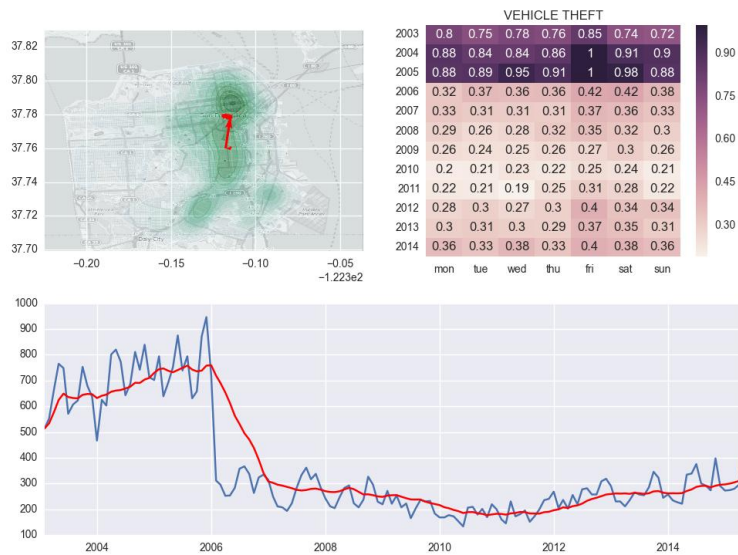


Prevalent Crimes in San Francisco



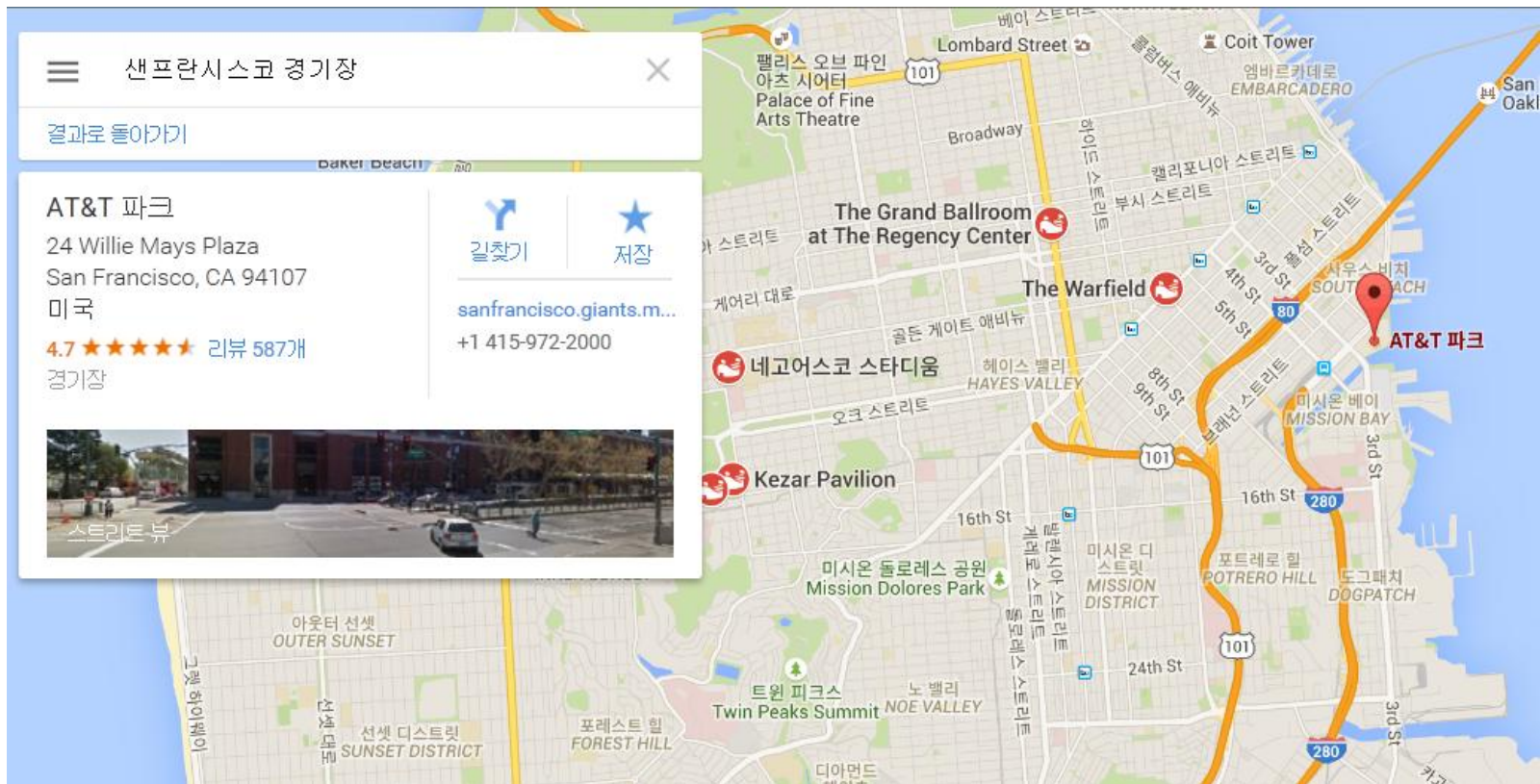
Top 10 범죄

Part 2 시각화 - 연도별 범죄수



차량 도난 및 자살 맵 밀도분포, 히트맵, 시계열 도표

Part 2 시각화 – 메이저리그 AT&A 파크



Part 2

시각화 - 월드시리즈 우승(2012, 2014)



2012년 월드시리즈 우승



2014년 월드시리즈 우승

Part 2

시각화 - 월드시리즈 우승(2012, 2014)

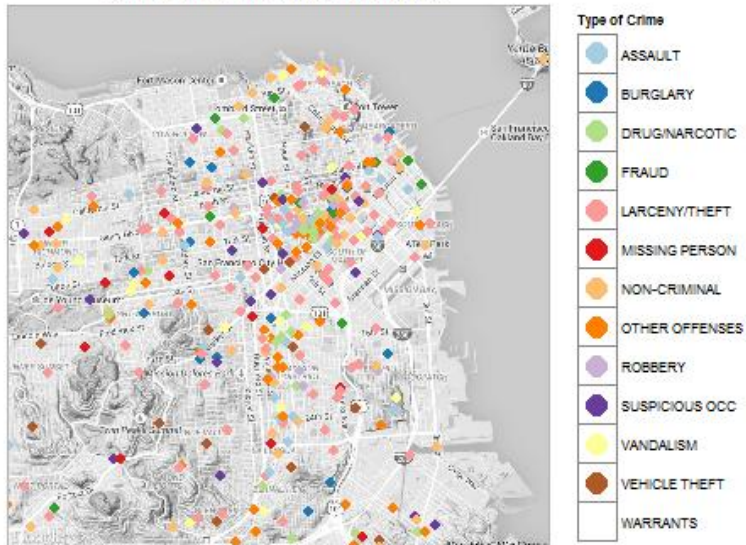
Baseball fans cause chaos in San Francisco after Giants win the World Series over Kansas City Royals

Read more: <http://www.dailymail.co.uk/news/article-2813586/Bumgarner-Giants-beat-KC-3-2-Series-Game-7.html>

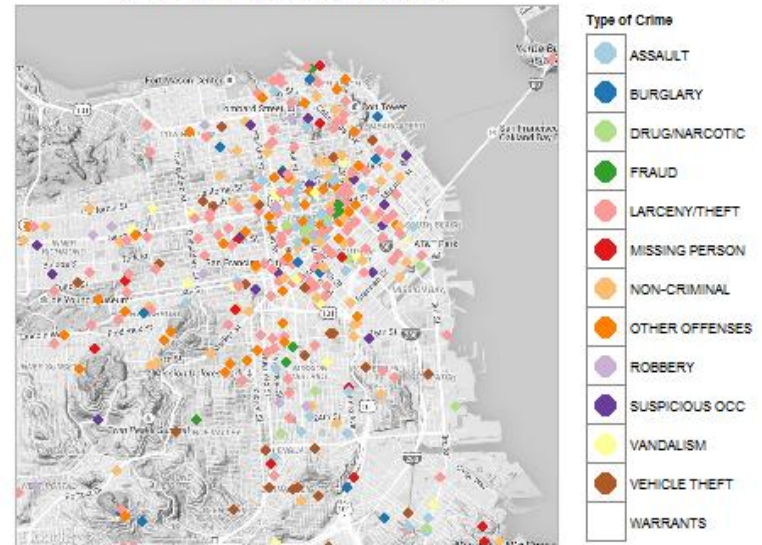


Part 2 시각화 – 2012? 2014?

2012 World Series in San Francisco



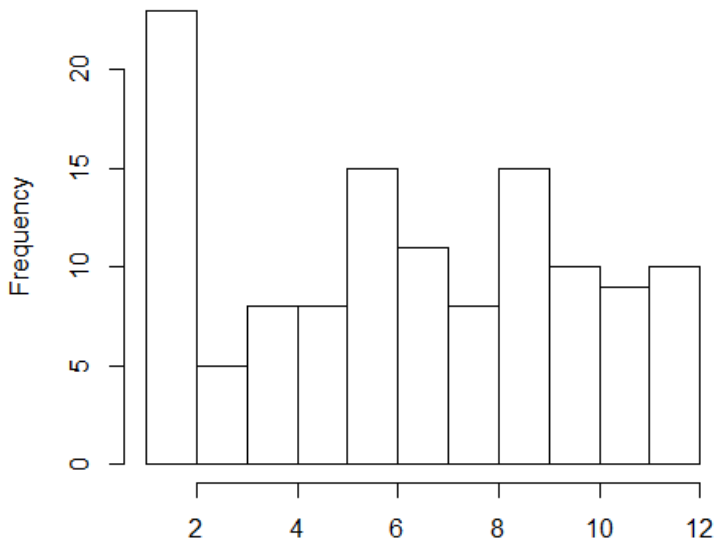
2014 World Series in San Francisco



2012년과 2014년 월드시리즈가 있었던 10월 달의 Top 범죄 지도

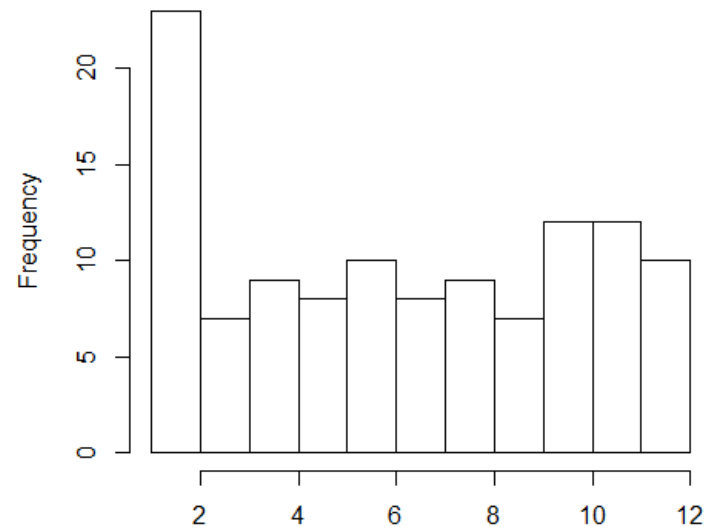
Part 2 시각화 – 2012? 2014?

```
crime_train[sfcrime_train$Category == "ARSON" & sfcr
```



```
sfcrime_train[sfcrime_train$Category == "ARSON" & sfcrime_train$year
```

```
crime_train[sfcrime_train$Category == "ARSON" & sfcr
```



```
sfcrime_train[sfcrime_train$Category == "ARSON" & sfcrime_train$year
```

2012년, 2014년 월별 방화 사건

Part 2

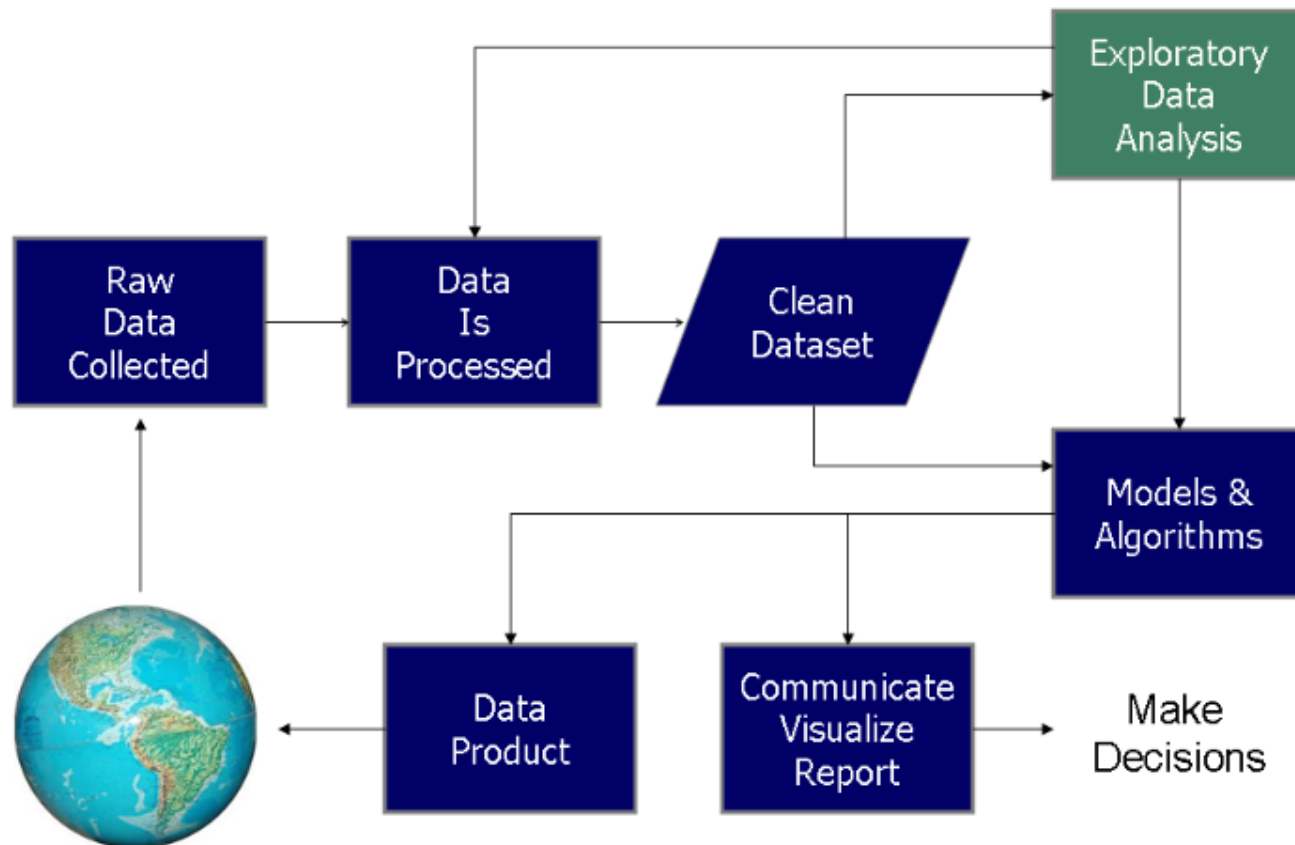
시각화 - 연습

각자 시각화 예제를 해보는 시간을 갖겠습니다. (30분 정도)

Part 3

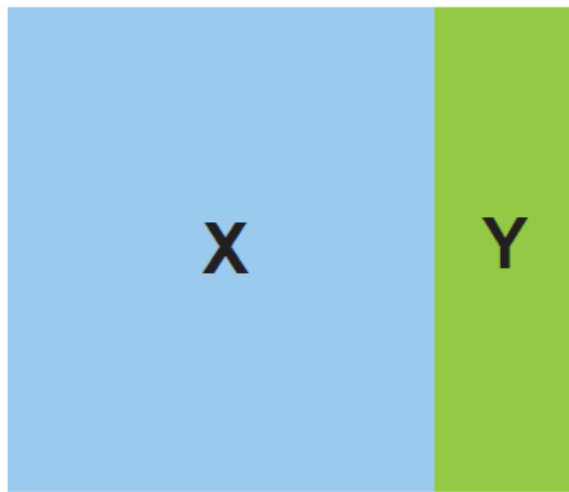
기계학습 - 소개

Data Science Process

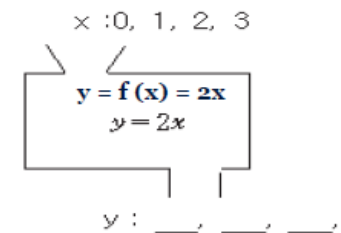
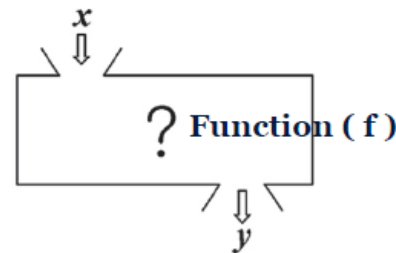


Part 3 기계학습 - 소개

[1] 지도학습(Supervised Learning)



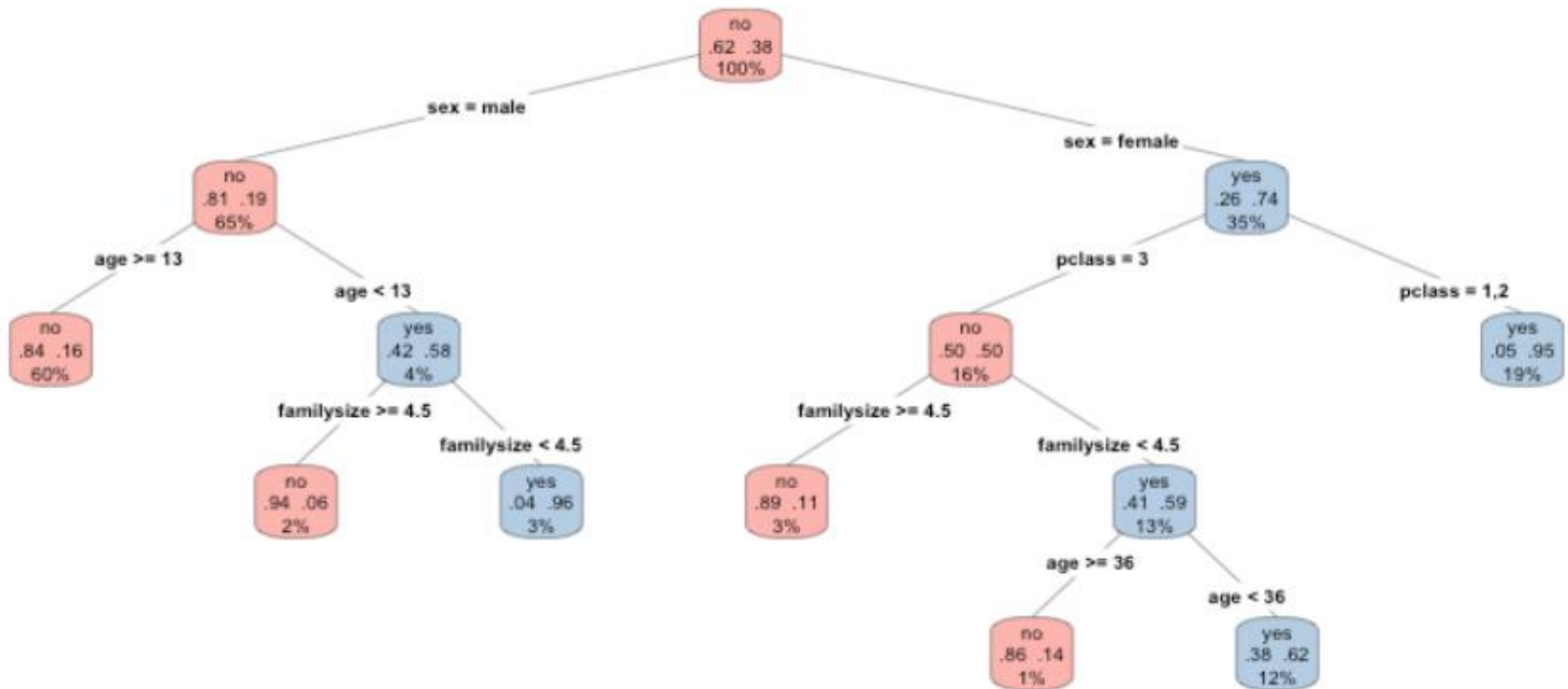
Predictors (예측변수) Response (반응변수)
Independent (독립변수) Dependent (종속변수)
Input (인풋변수) Output (아웃풋변수)



훈련데이터로부터 입출력사이의 함수를 만들어내는 기술. 출력 값의 형태에 따라 분류, 회귀문제로 정의될 수 있음.

Part 3 기계학습 - 소개

1-1) 분류(Classification)

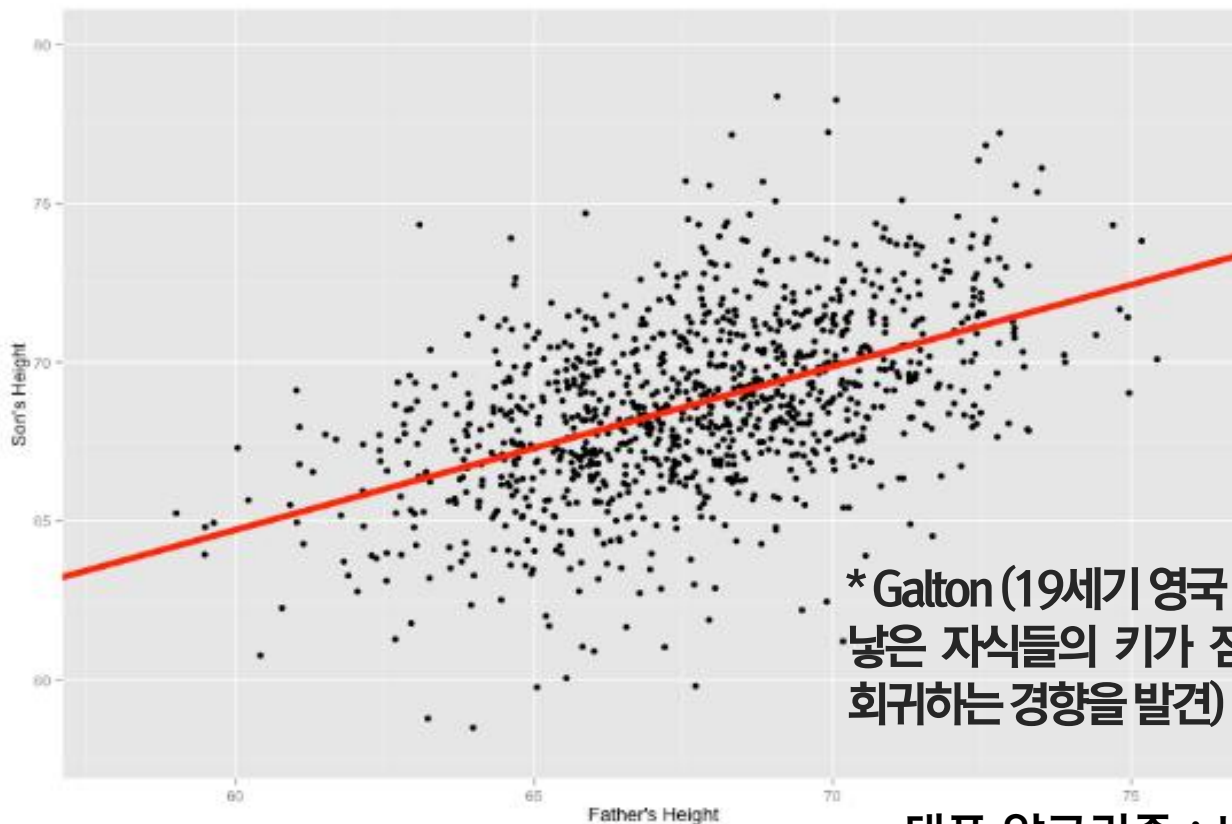


* 타이타닉 예제(데이터를 통한 생존 분류)

대표 알고리즘 : CART, LDA, SVM ...

Part 3 기계학습 - 소개

1-2) 회귀(Regression)

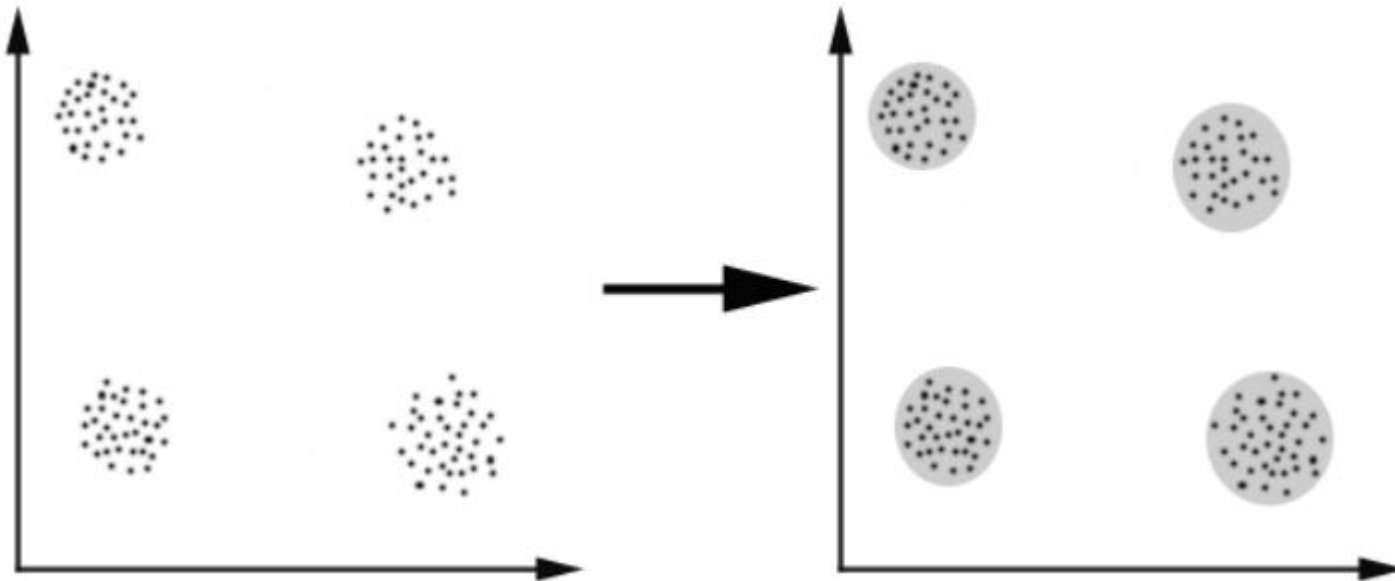


* Galton (19세기 영국 유전학자로 키 큰 부모들이 낳은 자식들의 키가 점점 커지지 않고 평균 키로 회귀하는 경향을 발견)

대표 알고리즘 : Least Squares, ARIMA...

Part 3 기계학습 - 소개

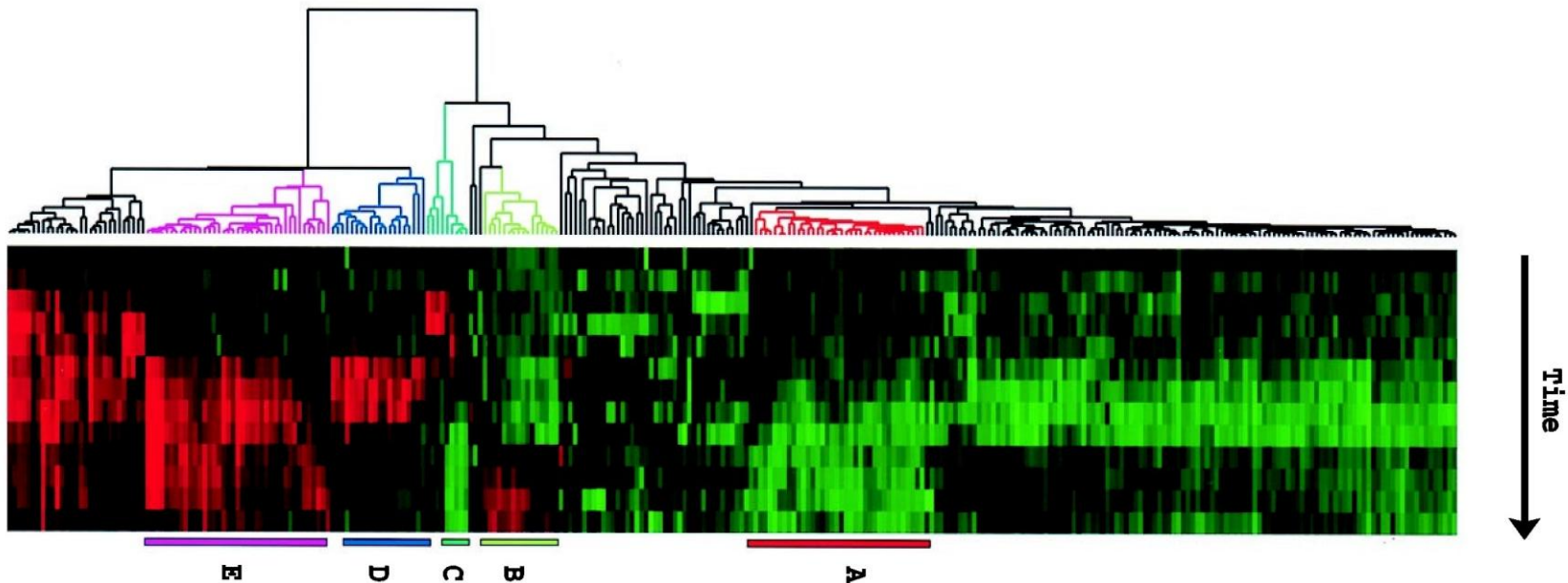
[2] 자율학습(Unsupervised Learning)



입력 값간의 패턴에 근거하여 학습을 진행하는 방법. 지도학습과 달리 입력 값만 있음 (출력 값이 없음)

Part 3 기계학습 - 소개

2-1) 군집(Clustering)



cDNA microarray with elements representing approximately 8,600 distinct human genes. (Samples were taken at time 0, 15min, 1 hr, 2 hr, 3 hr, 4hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr)

대표 알고리즘 : Kmeans, Hierarchical Clustering

Part 3 기계학습 - 소개

2-2) 연관규칙(Association Rule)

MARKET BASKET ANALYSIS



*98% of people who purchased items A and B
also purchased item C*

$$\begin{aligned}\text{support}(A \Rightarrow B) &= P(A \cup B) \\ \text{confidence}(A \Rightarrow B) &= P(B|A) \\ &= \frac{P(A \cup B)}{P(A)} \\ \text{lift}(A \Rightarrow B) &= \frac{\text{confidence}(A \Rightarrow B)}{P(B)} \\ &= \frac{P(A \cup B)}{P(A)P(B)}\end{aligned}$$

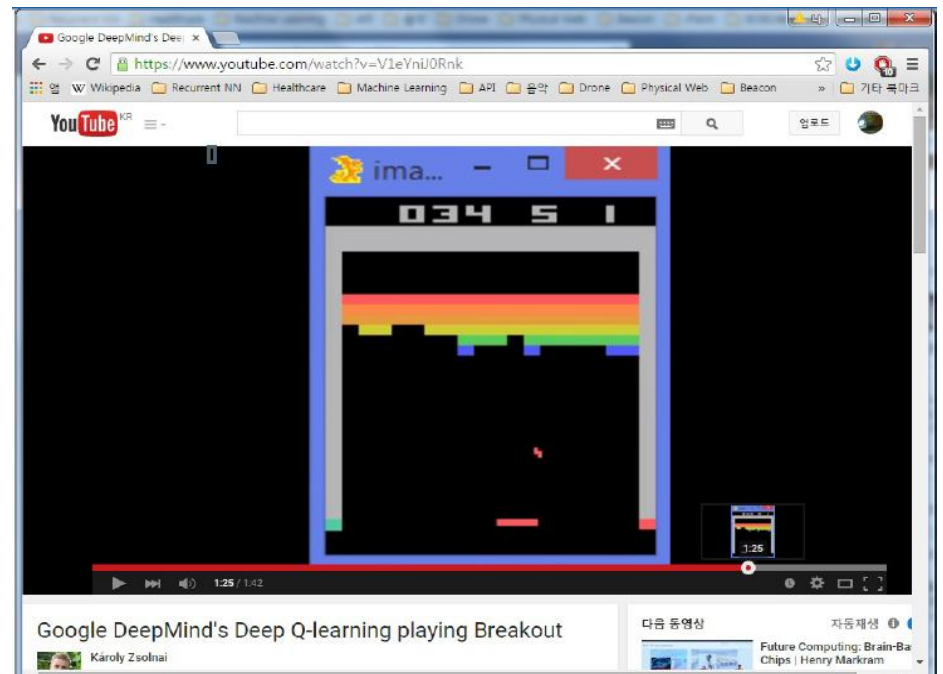
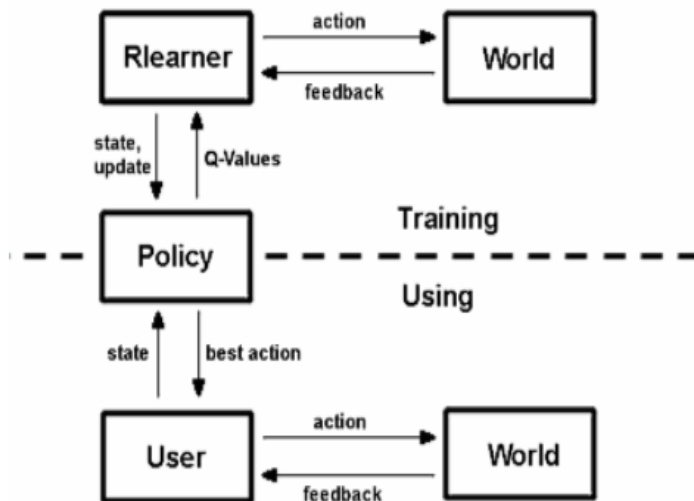
* 장바구니 분석을 하여 무슨 아이템이 무슨 아이템과 잘 어울리는 지 밝혀내는 것
(규칙을 생성하여 리스트를 압축)

대표 알고리즘 : Apriori

Part 3 기계학습 - 소개

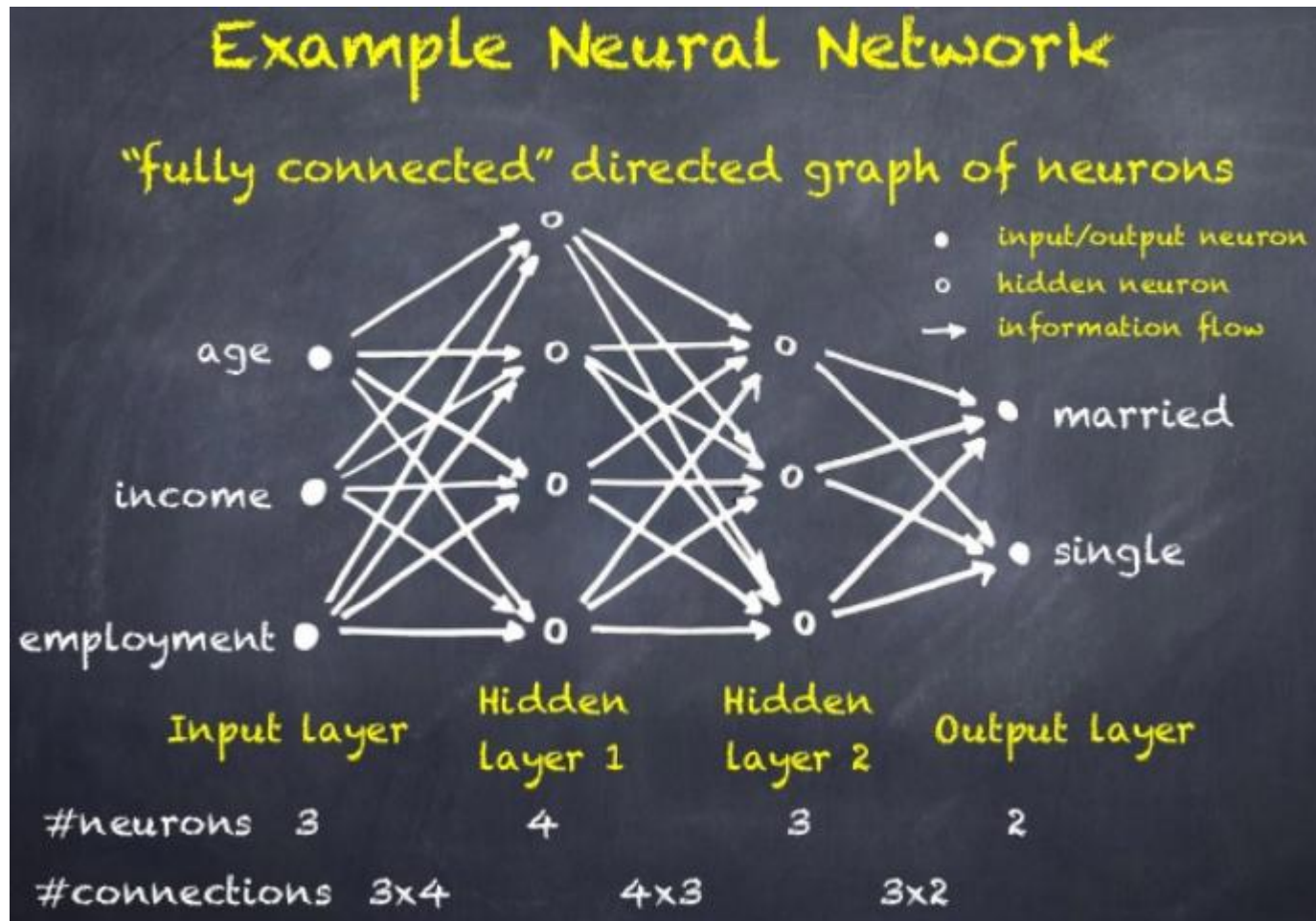
[3] 강화학습(Reinforcement Learning)

칭찬은 고래도 춤추게 한다는 말처럼 컴퓨터가 수행한 행동에 대해 보상을 주어 좋은 방향으로 반복하여 행동을 강화시키는 학습 방법



구글 딥마인드의 벽돌부시기 게임학습 방법(강화학습+딥러닝)

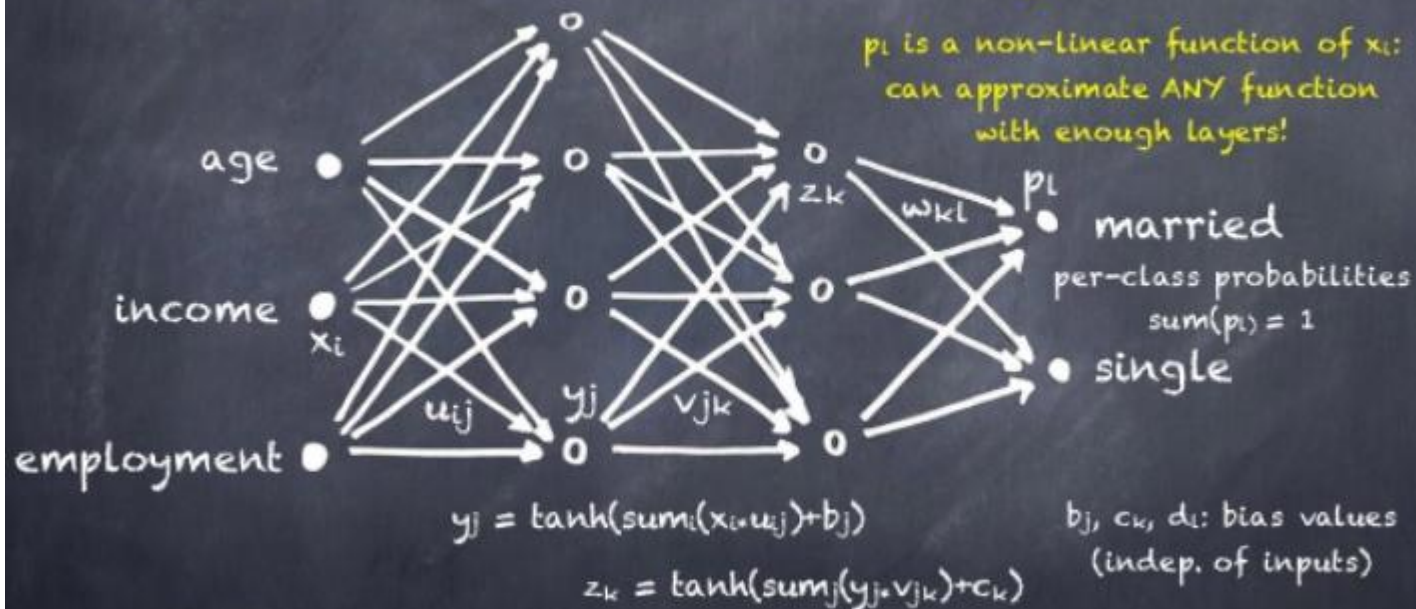
Part 3 기계학습 - 방법



Part 3 기계학습 - 방법

Prediction: Forward Propagation

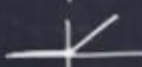
"neurons activate each other via weighted sums"



activation function: \tanh

alternative:

$x \rightarrow \max(0, x)$ "rectifier"



$p_l = \text{softmax}(\sum_k (z_k \cdot w_{kl}) + d_l)$

$\text{softmax}(x_k) = \exp(x_k) / \sum_k (\exp(x_k))$

Part 3

기계학습 - 방법

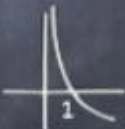
Training: Update Weights & Biases

For each training row, we make a prediction and compare with the actual label (supervised learning):

predicted	actual	
0.8	1	married
0.2	0	single

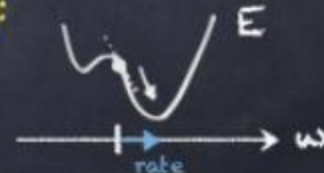
Objective: minimize prediction error (MSE or cross-entropy)

Mean Square Error = $(0.2^2 + 0.2^2)/2$ "penalize differences per-class"

Cross-entropy = $-\log(0.8)$  "strongly penalize non-1-ness"

Stochastic Gradient Descent: Update weights and biases via gradient of the error (via back-propagation):

$$w \leftarrow w - \text{rate} * \partial E / \partial w$$



Part 3

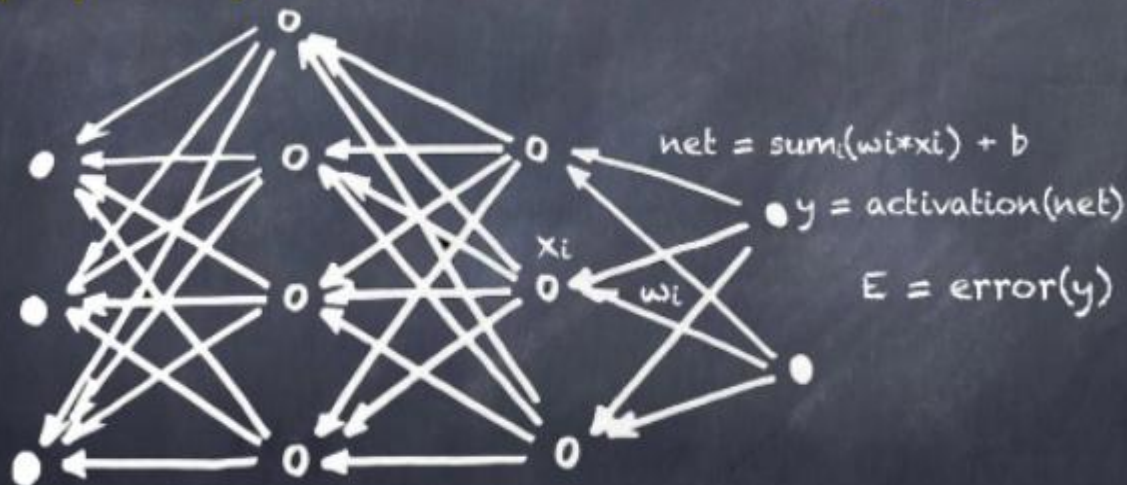
기계학습 - 방법

Backward Propagation

How to compute $\partial E / \partial w_i$ for $w_i \leftarrow w_i - \text{rate} * \partial E / \partial w_i$?

Naive: For every i , evaluate E twice at $(w_1, \dots, w_i \pm \Delta, \dots, w_N) \dots$ Slow!

Backprop: Compute $\partial E / \partial w_i$ via chain rule going backwards



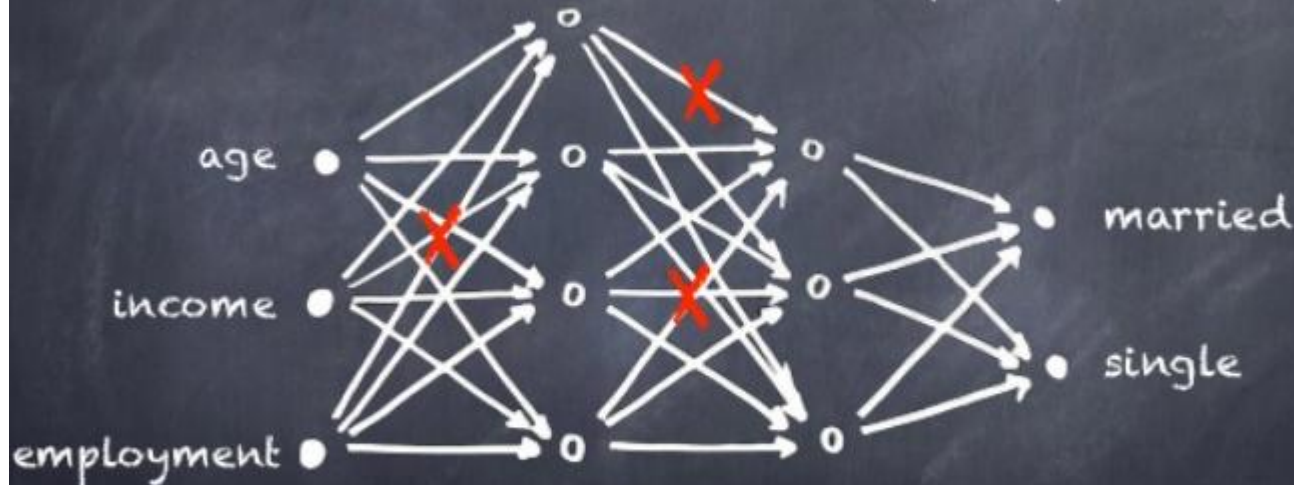
$$\begin{aligned} \partial E / \partial w_i &= \partial E / \partial y * \partial y / \partial \text{net} * \partial \text{net} / \partial w_i \\ &= \partial (\text{error}(y)) / \partial y * \partial (\text{activation}(\text{net})) / \partial \text{net} * x_i \end{aligned}$$

Part 3 기계학습 - 방법

Detail: Dropout Regularization

Training:

For each hidden neuron, for each training sample, for each iteration, ignore (zero out) a different random fraction p of input activations.

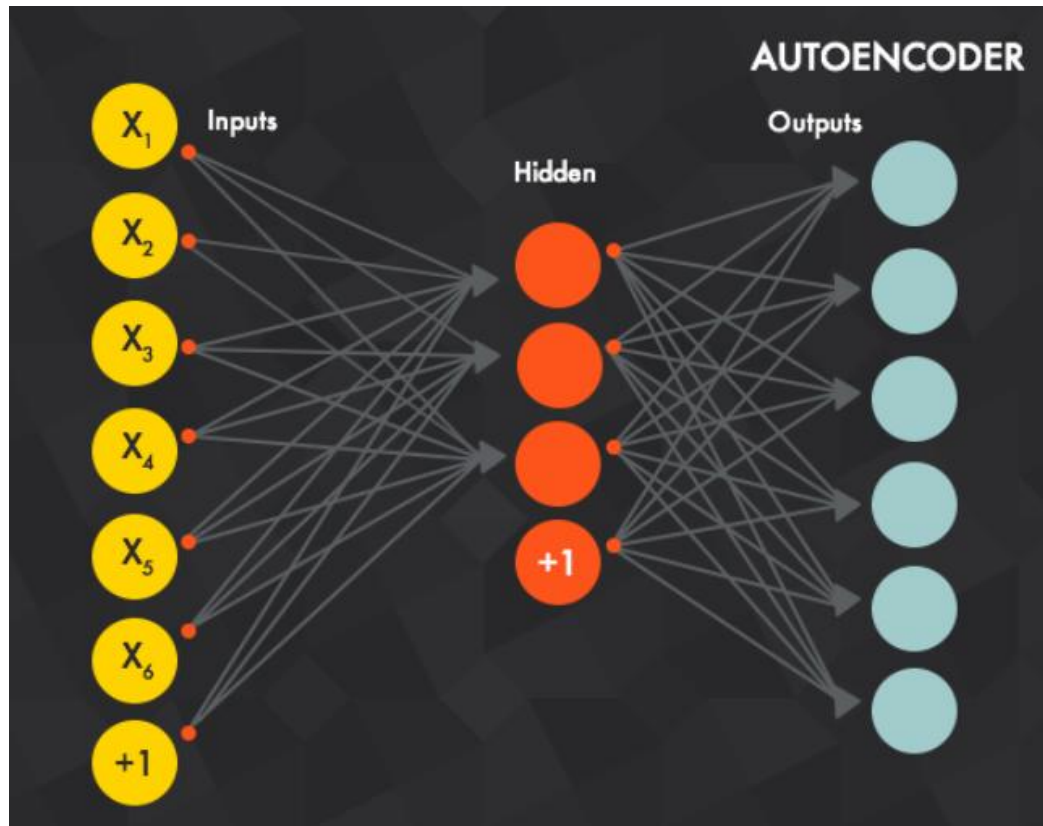


Testing:

Use all activations, but reduce them by a factor p (to "simulate" the missing activations during training).

cf. Geoff Hinton's paper

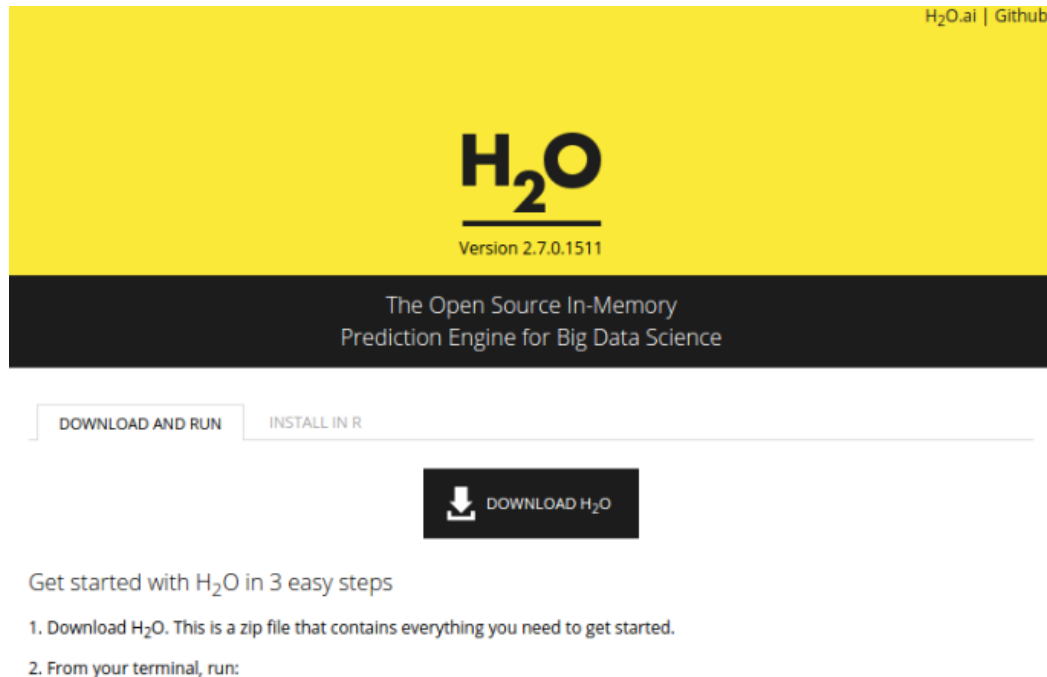
Part 3 기계학습 - 방법



Output Layer을 Input과 동일하게 두고 Back Propagation을 사용하는 형태.
Input \rightarrow Hidden (압축된 표현으로 학습함) \rightarrow Output으로 재구성(Reconstruct)
Input = decoder(encoder(input)) 훈련방법 : RBM

Part 3

기계학습 - 사용한 툴



H2O is the open source math & machine learning engine for big data that brings distribution and parallelism to powerful algorithms while keeping the widely used languages of R and JSON as an API.
사이트 (<http://learn.h2o.ai/>)

Part 3 기계학습 - 사용한 툴



Learn H₂O at learn.h2o.ai
Learn more [@gitbook](#)

A unique way to explore H₂O

Use H₂O from R



H₂O supports both R and R Studio.

[Try it!](#)

Random Forest

Random Forest is a classical machine learning method for classification and regression. Learn how to use it with H₂O for better predictions.

[Try it!](#)

GBM

GBM uses gradient boosted trees for classification and regression, and is one of the most powerful machine learning methods in H₂O.

[Try it!](#)

GLM

Generalized linear model is a generalization of linear regression. Experience its unique power and blazing speed on top of H₂O.

[Try it!](#)

K-Means

Perform clustering analysis with H₂O. K-means is a highly scalable clustering algorithm for unsupervised learning on big data.

[Try it!](#)

Deep Learning

H₂O's distributed Deep Learning gives you the power of deep neural networks for highest accuracy for classification and regression.

[Try it!](#)

Part 3

기계학습 - 사용한 툴

```
Sys.setenv(JAVA_HOME="C:/Program Files/Java/jdk1.8.0_45")
```

```
library(h2o)
```

```
#start an H2o cluster on local pc at with 4gs of memory and access to all cores
```

```
localh2o <-
```

```
h2o.init(ip="localhost",port=54321,startH2O=T,max_mem_size='6g',nthreads  
= -1)
```

```
head(sfcrime_train4)
```

```
str(sfcrime_train4)
```

```
dim(sfcrime_train4)
```

```
dat_h2o <- as.h2o(localh2o,sfcrime_train4,key='train')
```

```
sol_h2o <- as.h2o(localh2o,sfcrime_test3,key='test')
```

Train 데이터 정제 [시간 데이터 + 주소 데이터 + 나머지 Factor 변수]

(Address 정보에서 요약된 접미사(Suffix 값을 뽑아낸 뒤 dummy 변수화)

(http://pe.usps.gov/text/pub28/28apc_002.htm)

Part 3

기계학습 - 사용한 툴

```
model<-h2o.deeplearning(x= 1:45,
                        classification=T,
                        nfolds = 5,
                        y= 46,
                        data=dat_h2o,
                        activation="RectifierWithDropout",
                        hidden_dropout_ratio=c(.2,.3,.2),
                        l1=1e-5,
                        hidden = c(500,500,500),
                        epochs = 100)

model@model$train_class_err
model@model$confusion
model@model$valid_class_error
str(model@model)
h2o_predicted<-h2o.predict(model,sol_h2o)
predicted<-as.data.frame(h2o_predicted)
final <- data.frame(Id = sfcrime_test$Id , predicted[,-1])
colnames(final) <- c("Id",levels(sfcrime_train$Category))
write.csv(final,file = "h2o_suffixadd.csv",row.names = FALSE,quote = F)
```

Part 3 기계학습 - 예측력

16	—	willjvr	2.38558	16	Fri, 26 Jun 2015 16:36:58 (-2.4d)
17	—	sd.groeve	2.39022	8	Fri, 03 Jul 2015 18:49:06 (-2.1d)
18	13	lenguyenthedat	2.39884	23	Thu, 16 Jul 2015 02:59:39 (-0.3h)
19	11	Richard Giles	2.40376	7	Tue, 30 Jun 2015 13:33:26
20	11	Smerity	2.41087	3	Mon, 08 Jun 2015 07:43:31 (-2.2d)
21	11	Denchik	2.41208	2	Sat, 13 Jun 2015 08:22:42
22	—	Sledge Hammer!	2.41804	9	Thu, 02 Jul 2015 00:10:12 (-4.5h)
23	—	Devin	2.42173	1	Thu, 09 Jul 2015 02:08:51
24	—	Vladimir Nekrasov	2.42312	14	Fri, 26 Jun 2015 19:18:38 (-1.2h)
25	new	이상열	2.42769	2	Thu, 16 Jul 2015 03:05:02

Your Best Entry ↑

You improved on your best score by 0.12504.

You just moved up 26 positions on the leaderboard.

 Tweet this!

Reference

1. 요약의 기술 데이터마이닝 (고려대학교 산업경영공학과 김성범 교수님)
2. Data Science – 왜 ‘과학’인가? (김형진님)
3. Kaggle 포럼 (<https://www.kaggle.com/c/sf-crime/forums>)
- 4, H2ODeepLearningThroughExamples021215
(<http://www.slideshare.net/0xdata/h2odeeplearning>)

Thank you

들어주셔서 감사합니다.

