19 대 국회 뽀개기

이홍주

Who am I?

- Research Engineer at fintech startup.
  - Fraud Detection System based on Machine Learning techniques.
- Previous Employment
  - Software Center at LG Electronics
  - IT headquarters at KT
  - Daum Communications
  - NURI Telecom

질문

- 내가 만든 Python 코드는 대게 일회용이다 .

질문

- Python 또는 다른 언어로 ML 프로그래밍을 해봤다 .

질문

- Numpy / Pandas 의 slicing, indexing 에 익숙하다 .

질문

- lambda expression 을 많이 사용하고 있다 .

Machine Learning

- Algorithm → learn data → model → prediction
  - “The ***ability*** to learn without being explicitly programmed”. (Authur Lee Samuel)
  - Set of ***algorithms*** that can ***learn*** from and perform ***predictive*** analysis on data.
  - Such ***algorithms*** operate by building a ***model*** from an example ***training set*** of input observations in order to make data-driven ***predictions*** or decisions expressed as outputs, rather than following strictly static program instructions.

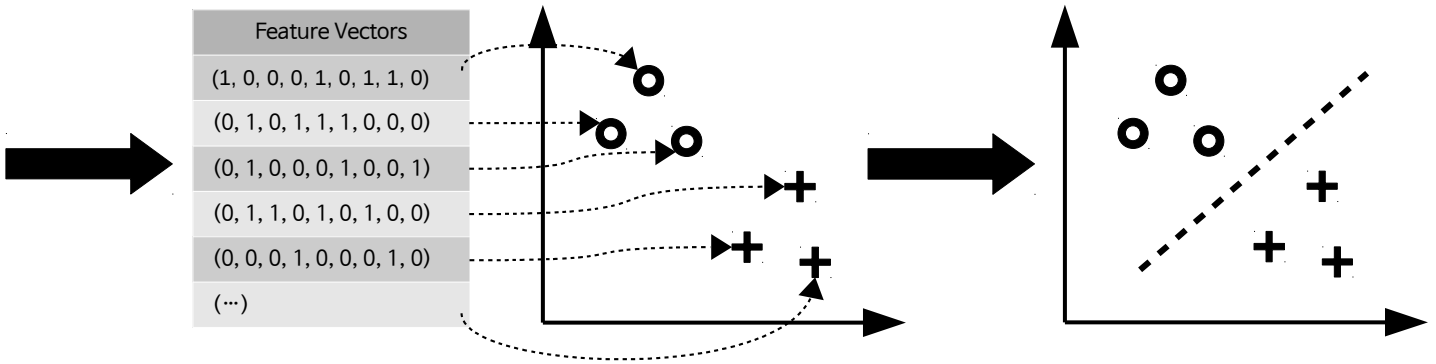
ML Process

- Featurization → Feature Vectors
  - Analyze Data
    - Understand the information available that will be used to develop a model.
  - Prepare Data
    - Discover and expose the structure in the dataset.
- Train Model → Model
- Evaluate Model
  - Develop a robust test harness and baseline accuracy from which to improve and spot check algorithms.
- Improve Results
  - Leverage results to develop more accurate models.

ML Process

이름	특징 (feature)	성별
철수	넥타이 , 안경 , 수염 , 대머리	남
영희	스커트 , 안경 , 핸드백 , 긴머리	여
근혜	스커트 , 핸드백 , 닭머리	여
두환	군복 , 대머리 , 안경	남
홍주	긴머리 , 수염	남
...	...	...

Feature Vectors
(1, 0, 0, 0, 1, 0, 1, 1, 0)
(0, 1, 0, 1, 1, 1, 0, 0, 0)
(0, 1, 0, 0, 0, 1, 0, 0, 1)
(0, 1, 1, 0, 1, 0, 1, 0, 0)
(0, 0, 0, 1, 0, 0, 0, 1, 0)
(…)



Data (featurizing) Feature vectors (training) Model

What we don’t

- !Crawling
  - already done and ready
  - Not in this scope
  - Thanks to 참여연대, 팀포퐁, OhmyNews
- !Evaluation
  - We are not focusing on results but the process.
- !Improving
  - Just one cycle with no turning back.

What we do

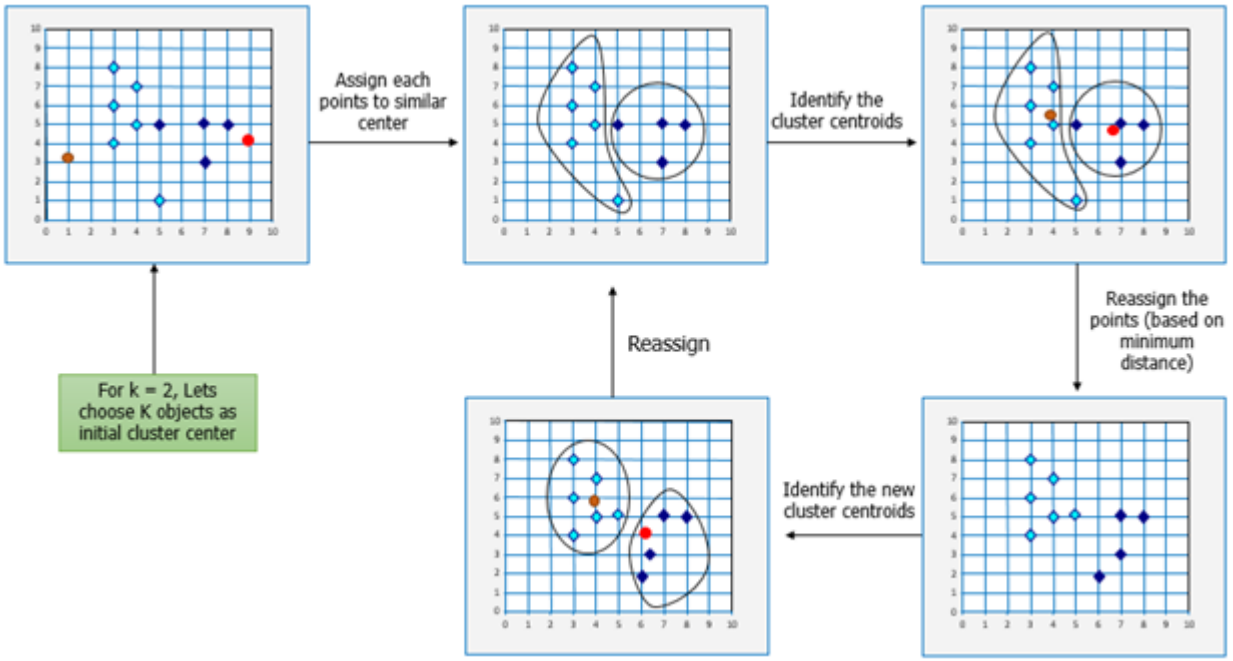
- Featurization
  - Analyze / Prepare data
  - Pandas
- Train a model
  - Unsupervised Learning : K-Means Clustering
  - Scikit Learn, PySpark
- Presentation
  - Matplotlib, LightingViz
- Subject
  - 19대 국회 표결 결과 / 정치자금 사용내역 / 국회 회의록

19 대 국회 의안표결

- Featurization
- Train model
  - K-Means Clustering
- Presentation

19 대 국회 의안표결

- Train model
  - K-Means Clustering



19 대 국회 의안표결

19 대 국회 정치자금 사용내역

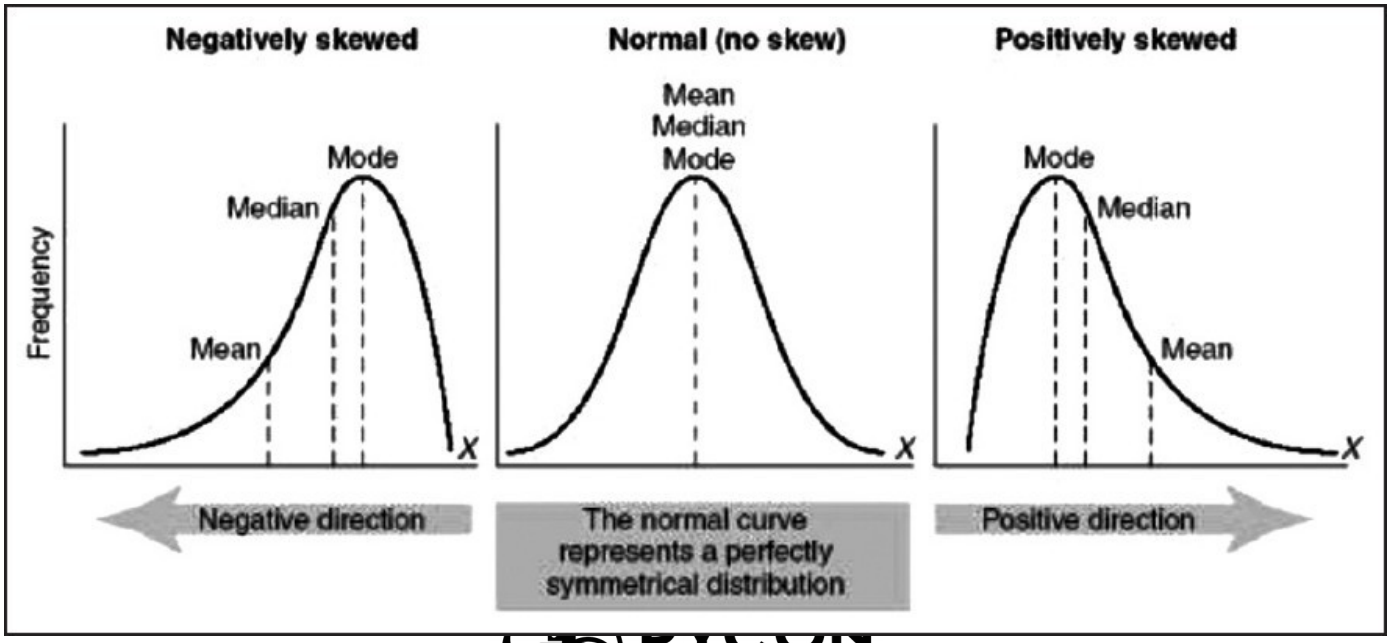
- Featurization
  - Normalization / Standardizing
  - Skewness
  - Outliers
- Train model
- Presentation

19 대 국회 정치자금 사용내역

- Featurization
  - Normalization / Standardizing
    - Distance computation in k-means weights each dimension equally and hence care must be taken to ensure that unit of dimension shouldn’t distort relative nearness of observations.

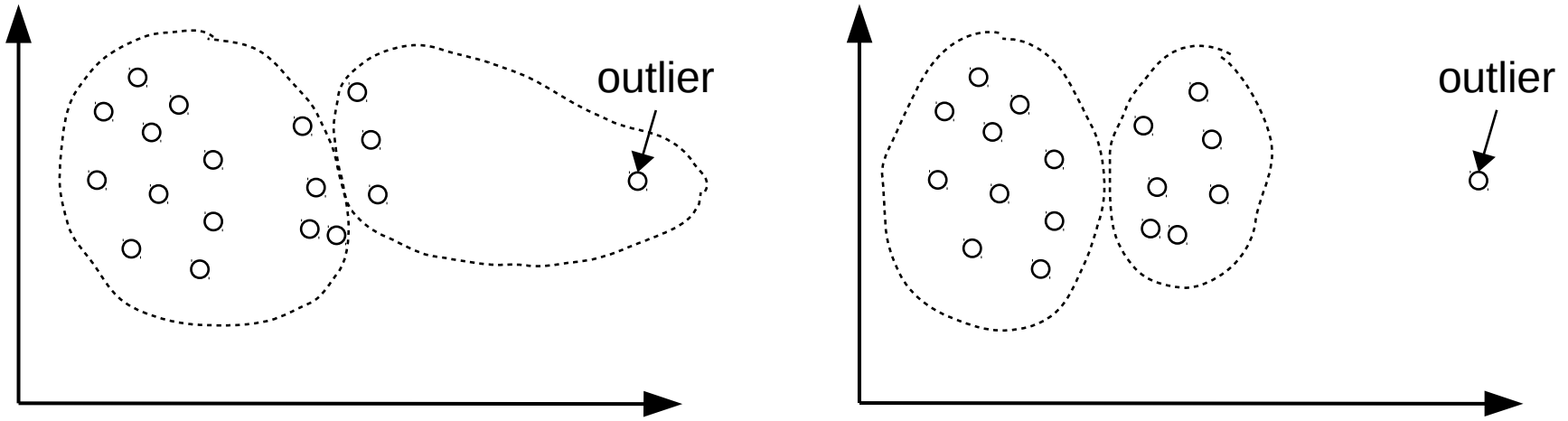
19 대 국회 정치자금 사용내역

- Featurization
  - Skewness
    - Leads bias



19 대 국회 정치자금 사용내역

- Featurization
  - Outliers
    - Weakness of centroid based clustering (k-means)



19 대 국회 정치자금 사용내역

- Featurization
  - Outliers
    - IQR (Inter Quartile Range)
    - 2-5 Standard Deviation
      - Chicken and Eggs
    - MAD (Median Absolute Deviation)
      - Christophe Leys (2014). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology

19 대 국회 정치자금 사용내역

19 대 국회 회의록

- Featurization
  - Preprocessing
    - Sentence Segmentation
      - Can skip to tokenziation depending on use case
    - Tokenization
      - Find individual words
    - Lemmatization
      - Find the base of words (stemming)
    - Removing stop words
      - “the”, “a”, “is”, “at”, etc
    - Some other refinements are needed occasionally.
    - Vectorization
      - We take an array of floating point values

19 대 국회 회의록

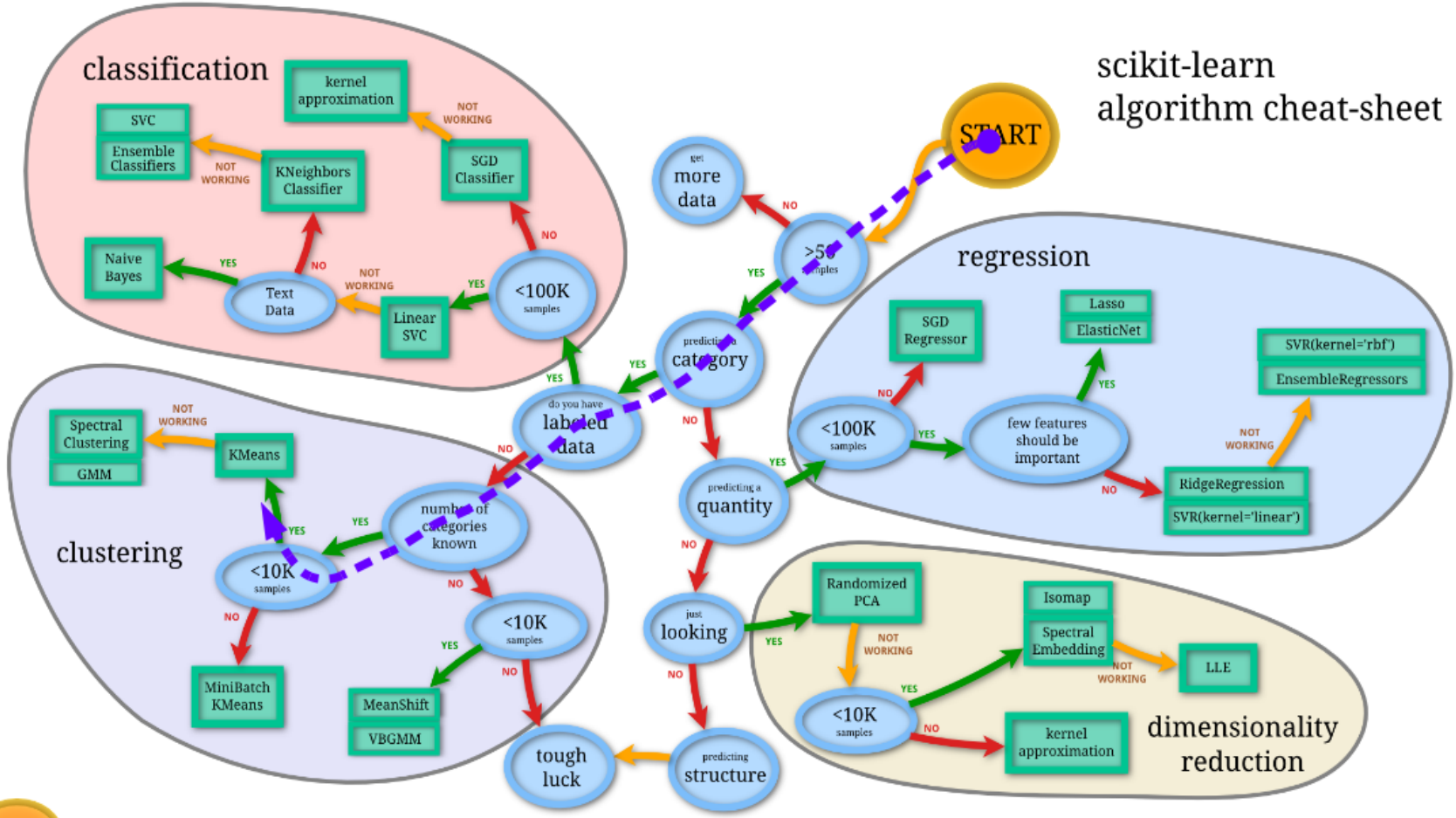
- Featurization
  - Text vectorization
    - How can we compose a feature vector of numbers representing a text document of words?
      - Bag of Words
      - TF-IDF
    - Each word in a text would be a dimension.
    - How big would that vector be?
      - Would it be as big as vocabulary of an article?
      - Would it be as big as vocabulary of “all” article?
      - Why not all vocabulary in the language?

19 대국회 회의록

- Featurization
  - Word vectorization
    - Vector representation of words in a latent context
    - Learn semantic relations of words in documents
    - ex) “I ate ____ in Korea”
      - with a context “eat” it can be DimSum, Spagetti, or KimChi without considering the place Korea.
      - with a context “Korea” it can be Seoul, K-pop, which represents the place better without considering context “eating”.

19 대 국회 회의록

Ending...


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
YES
get more data
NO
>50 samples
YES
predicting a category
YES
NO
do you have labeled data
NO
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression
SVR(kernel='linear')
predicting a quantity
YES
NO
clustering
Spectral Clustering
NOT WORKING
GMM
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
<10K samples
YES
NO
MiniBatch KMeans
MeanShift
VBGMM
just looking
YES
NO
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
tough luck
predicting structure
Back
scikit learn

Contact me!

- lee.hongjoo@yandex.com
- http://www.linkedin.com/in/hongjoo-lee