Data Mining with Python

Data Mining

Prof. Sujee Lee

Department of Systems Management Engineering

Sungkyunkwan University

scikit-learn

- https://scikit-learn.org/stable/index.html

Install | User Guide | API | Examples | More | Go

scikit-learn

Machine Learning in Python

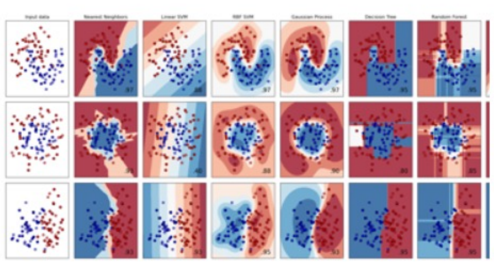
Getting Started | Release Highlights for 1.0 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

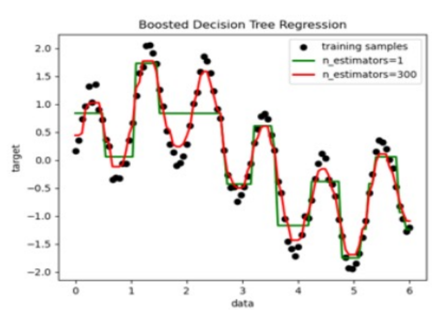
Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...



Regression

Predicting a continuous-valued attribute associated with an object.

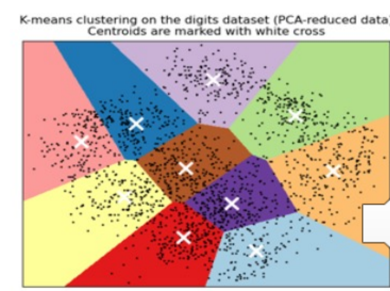
Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...



Clustering

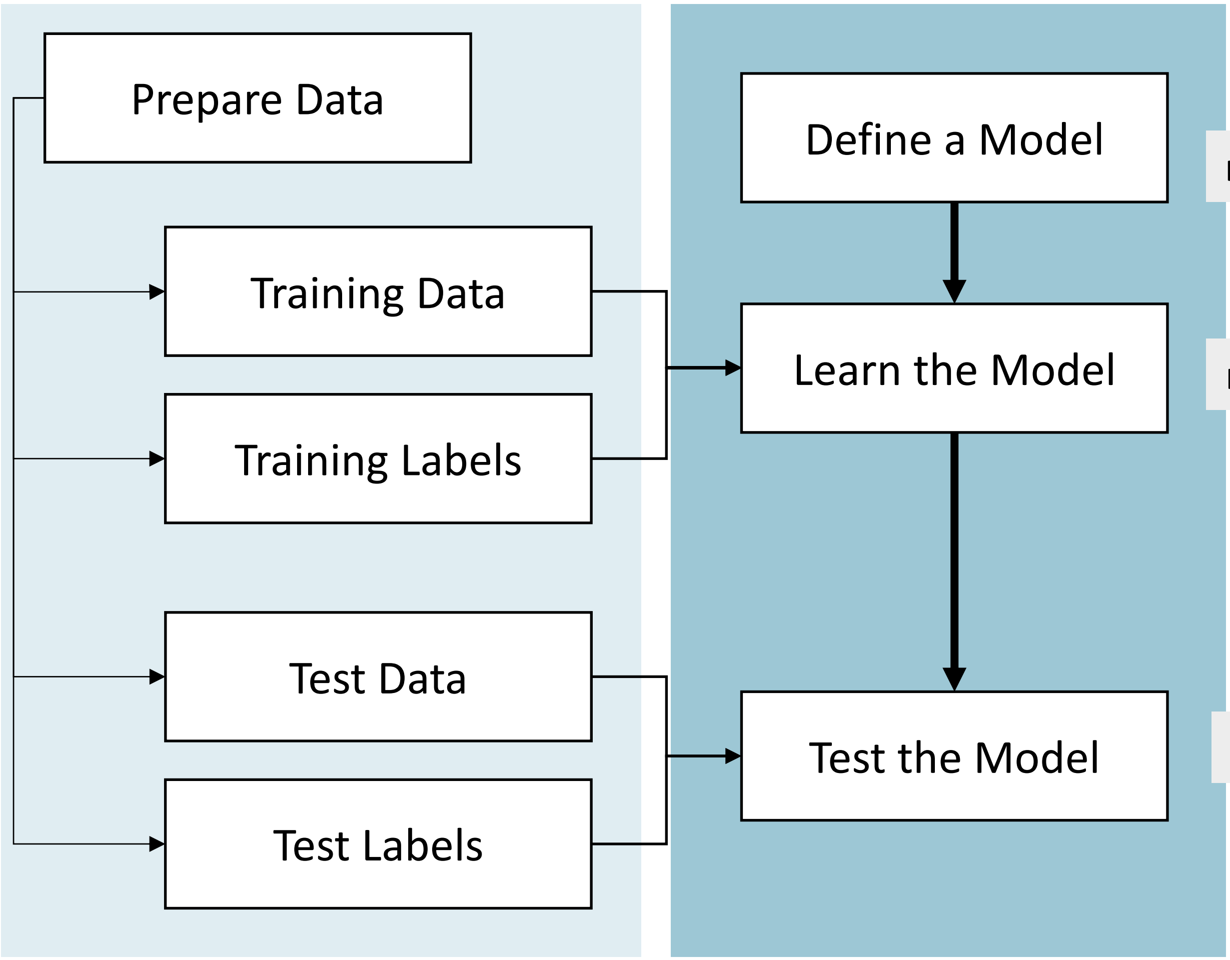
Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...



scikit-learn Practice

- **Supervised Learning in scikit-learn**



model = *algorithm*(parameters)

model.**fit**(X_train,y_train)

y_test_hat = model.**predict**(X_test)

Regression Example: Boston Housing

- **Boston House Prices Dataset**
  - https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data
  - Information on various factors influencing housing prices in the Boston area, based on data collected in 1978

Data Set Characteristics:

Number of Instances:	506
Number of Attributes:	13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.
Attribute Information (in order):	• CRIM per capita crime rate by town • ZN proportion of residential land zoned for lots over 25,000 sq.ft. • INDUS proportion of non-retail business acres per town • CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) • NOX nitric oxides concentration (parts per 10 million) • RM average number of rooms per dwelling • AGE proportion of owner-occupied units built prior to 1940 • DIS weighted distances to five Boston employment centres • RAD index of accessibility to radial highways • TAX full-value property-tax rate per $10,000 • PTRATIO pupil-teacher ratio by town • B 1000(Bk - 0.63)^2 where Bk is the proportion of black people by town • LSTAT % lower status of the population • MEDV Median value of owner-occupied homes in $1000's

Classification Example 1: Personal Loan Offer

- **Personal Loan Offer Dataset**
  - Predict which customers with existing debt are more likely to accept personal loan offers through targeted marketing
  - Target variable: accept bank loan (0/1)
  - Predictors: Demographic info, and info about their bank relationship

Age	캠페인 완료 당시 고객의 나이
Experience	경력 연수
Income	고객의 연간 수입(단위: 1,000달러)
Family Size	고객의 가족 수
CCAvg	월평균 신용카드 지출액(단위: 1,000달러)
Education	교육 수준 (1: Undergrad, 2: Graduate 3: Advanced/Professional)
Mortgage	주택 모기지 가치 (해당하는 경우) (단위: 1,000달러)
Securities Account	고객이 은행에 증권 계좌가 있는 경우 1로 코딩
CD Account	고객이 은행에 CD 계좌가 있는 경우 1로 코딩
Online Banking	인터넷 뱅킹 이용 시 1로 코딩
Credit Card	유니버설 은행에서 발급한 신용카드를 사용하는 경우 1로 코딩

Classification Example 2: Riding Mowers

- **Riding Mowers Dataset**
  - 24 households classified as owning or not owning riding mowers
  - Target variable: Ownership of a riding mower
  - Predictors: Income, Lot Size

가구 번호	소득 (1,000달러 단위)	주택 대지 크기 (1,000제곱피트 단위)	승차식 잔디깎이 기계 소유
1	60.0	18.4	Owner
2	85.5	16.8	Owner
3	64.8	21.6	Owner
4	61.5	20.8	Owner
5	87.0	23.6	Owner
6	110.1	19.2	Owner
7	108.0	17.6	Owner
8	82.8	22.4	Owner
9	69.0	20.0	Owner
10	93.0	20.8	Owner
11	51.0	22.0	Owner
12	81.0	20.0	Owner
13	75.0	19.6	Nonowner
14	52.8	20.8	Nonowner
15	64.8	17.2	Nonowner
16	43.2	20.4	Nonowner
17	84.0	17.6	Nonowner
18	49.2	17.6	Nonowner
19	59.4	16.0	Nonowner
20	66.0	18.4	Nonowner
21	47.4	16.4	Nonowner
22	33.0	18.8	Nonowner
23	51.0	14.0	Nonowner
24	63.0	14.8	Nonowner
25	60.0	20.0	?

Classification Example 3: Flight Delays

Flight Delays Dataset

- All flights from Washington D.C. to New York during January 2004.
- Target variable: Flight status (Ontime/Delayed)
  - A delay is defined as being more than 15 minutes late
  - Out of 2,201 flights, the percentage of delayed flights is 19.5%
- Predictors: 6 variables below

Day of Week	1=월요일, 2=화요일, …, 7=일요일
Departure Time	오전 6시와 오후 10시 사이를 18개 구간으로 나눈 출발 시간
Origin	3개의 출발 공항 코드: DCA(레이건 국립공항), IAD(댈러스 국제공항), BWI(볼티모어–워싱턴 국제공항)
Destination	3개의 도착 공항 코드: JFK(케네디 국제공항), LGA(라구아디아 공항), EWR(뉴어크 국제공항)
Carrier	8개의 항공사 코드: CO(컨티넨탈 항공), DH(아틀란틱 코스트 항공), DL(델타 항공), MQ(아메리카 이글 항공), OH(컴에어 항공), RU(컨티넨탈 익스프레스 항공), UA(유나이티드 항공), US(US 에어웨이 항공)
Weather	악천후로 연착된 경우 1로 표기