2020 D&A

MACHINE LEARNING SESSION

3. Machine Learning
Base Model

D&A

Contents

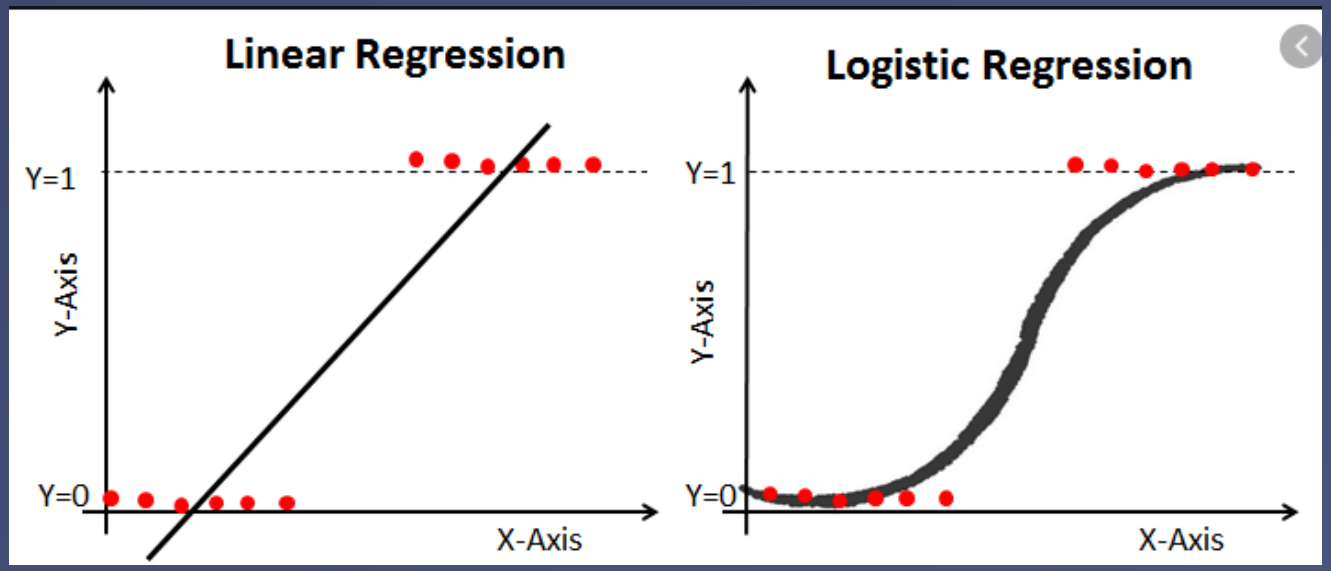
D&A

Logistic Regression (로지스틱 회귀)

선형회귀 모델 중 분류에 사용되는 회귀 모델

선형회귀 직선을 Sigmoid함수를 이용해 곡선으로 만들어 예측

확률값으로 예측해 0.5이상이면 1, 0.5미만이면 0으로 예측



장점

편의(편향)이 없음
간단한 알고리즘
예측의 신뢰도를 평가하는데 사용할 수 있는 가능성을 계산할 수 있음

단점

선형 관계에 있다고 가정해야함
학습을 제대로 하지 못하는 과소적합의 가능성이 높음

K-NN (K-최근접 이웃; K-Nearest Neighbors)

가장 간단한 머신러닝 알고리즘

train 데이터를 그냥 저장하는게 모델의 전부

train 데이터의 최근접 이웃을 찾아 예측

데이터 포인트 사이의 거리를 유클리디안 방식으로 계산

분류 / 회귀 모델에 전부 사용 가능함

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neighbors import KNeighborsRegressor
```

장점

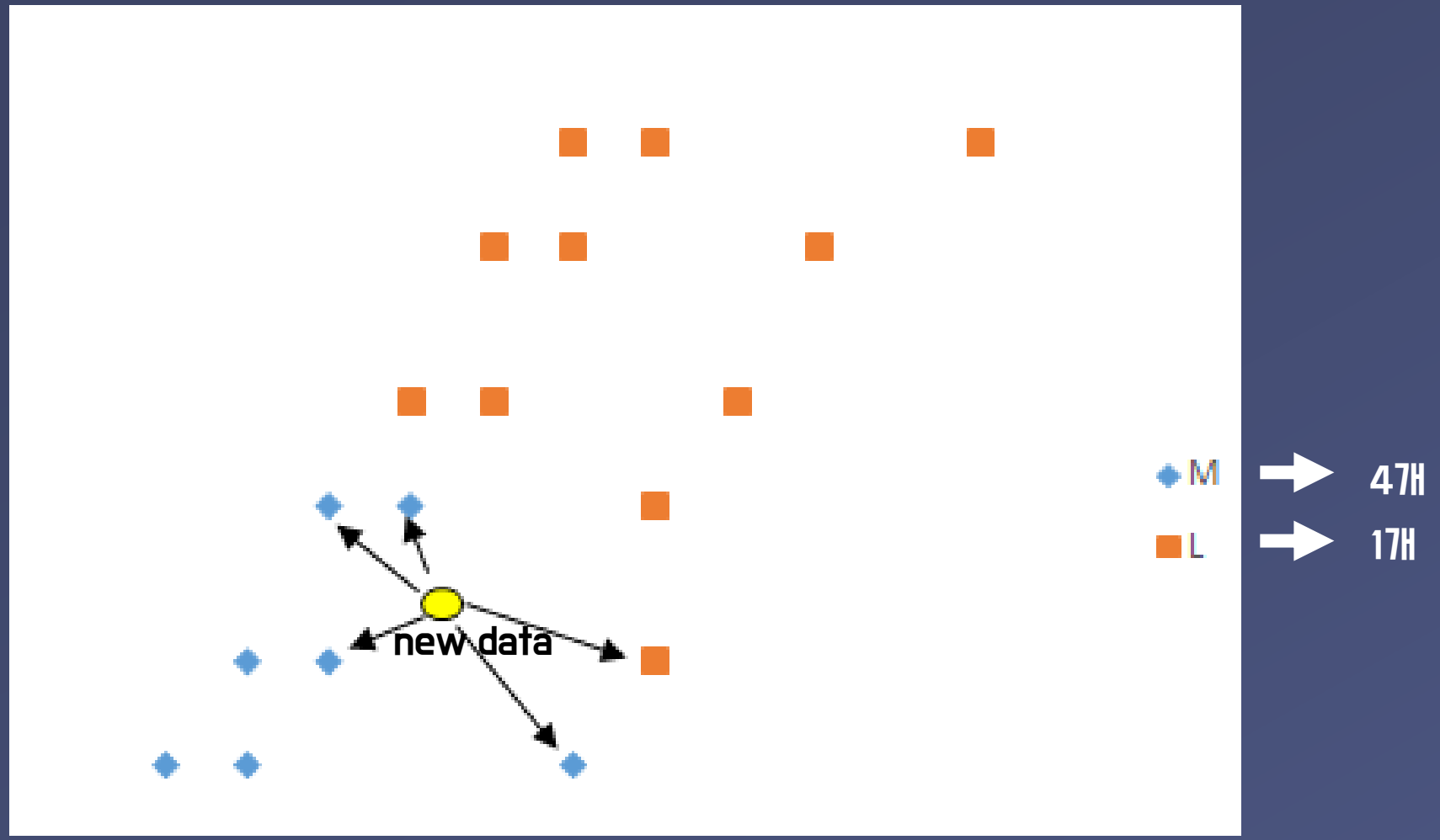
이해하기 매우 쉬운 모델
많이 조정하지 않아도 괜찮은 성능 발휘
파라미터가 많이 필요하지 않아 모델을 빠르게 만들 수 있음

단점

train 데이터가 매우 크면 예측이 느려짐
전처리 과정이 매우 중요!
(이웃간의 거리를 계산하기 데이터 정규화과정 필요)
희소데이터(sparse data)에서 잘 작동하지 않음

-> 현업에서 잘 사용하지 않음

02 K-NN



D&A

03 SVM

SVM (Support Vector machines)

선형 함수를 모델로 이용하는 것

SVM, C-SVM 커널-SVM으로 나눠짐
- SVM : margin내의 오차를 허용하지 않는 SVM
- C-SVM : margin내의 오차를 허용하는 SVM
- 커널-SVM : 평면에서 정의되지 않는 복잡한 모델을 만들 때 사용됨

분류와 회귀 둘 다 사용 가능

데이터 포인트 사이의 거리를 가우시안 커널에 의해 계산함

특성이 비슷하고 스케일이 비슷할 때 성능이 잘 나옴

L2규제를 사용 -> (C 파라미터로 조정)

from sklearn.svm import LinearSVC -> 분류
from sklearn.svm import LinearSVR -> 회귀

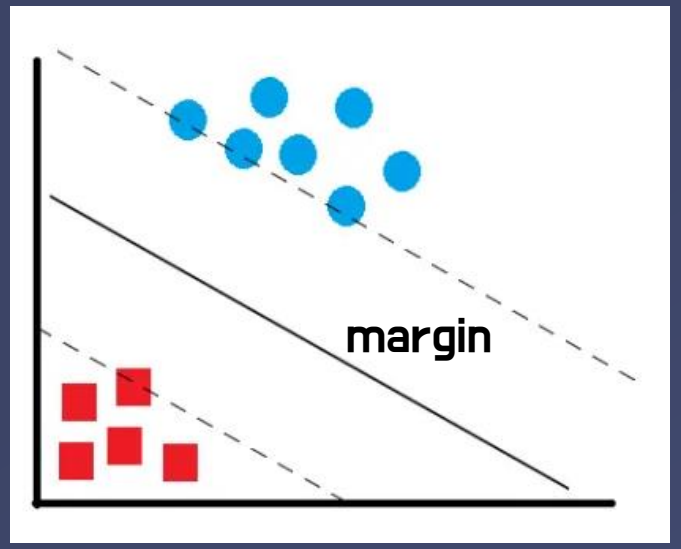
장점

강력한 모델이며 다양한 데이터에서 잘 작동함
피쳐가 얼마 없어도 복잡한 결정경계를 만들 수 있음
학습속도가 빠르고 예측도 빠름
매우 큰 데이터나 희소한 데이터에서도 잘 작동함

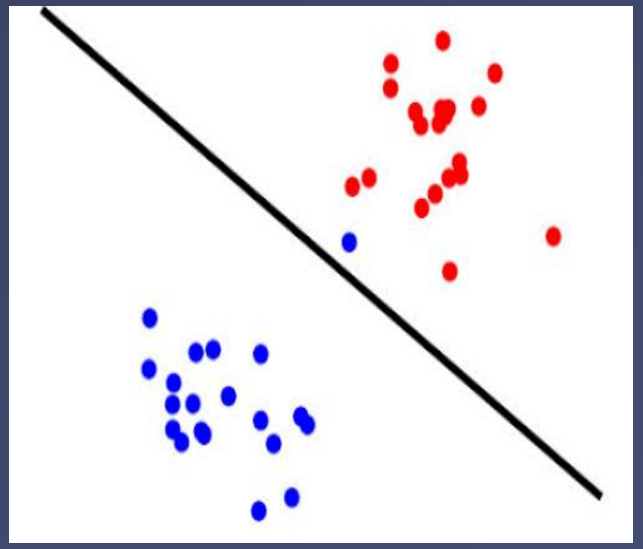
단점

샘플이 많을 때에는 비용이 많이 들기 때문에 사용하기 힘듦
고차원 데이터에서 매우 강력해지지만 피쳐가 많으면 과대적합되기 쉬움
저차원 데이터에서는 사용하기 힘듦

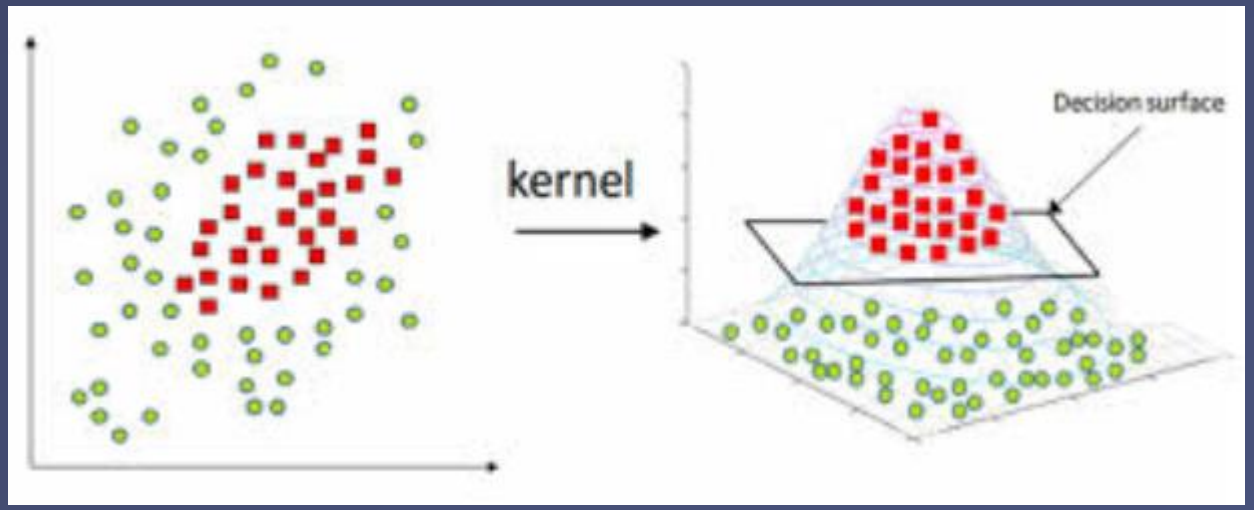
03 SVM



SVM



C - SVM



Kernel - SVM

D&A

DT (결정트리; Decision Tree)

분류와 회귀문제에 가장 널리 사용되는 모델

예/아니오 질문으로만 문제를 해결해나감

노드(node),리프(leaf),엣지(edge)등으로 트리의 모양을 구분

사후 가지치기, 사전 가지치기 등으로 과대적합을 막음

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import DecisionTreeRegressor
```

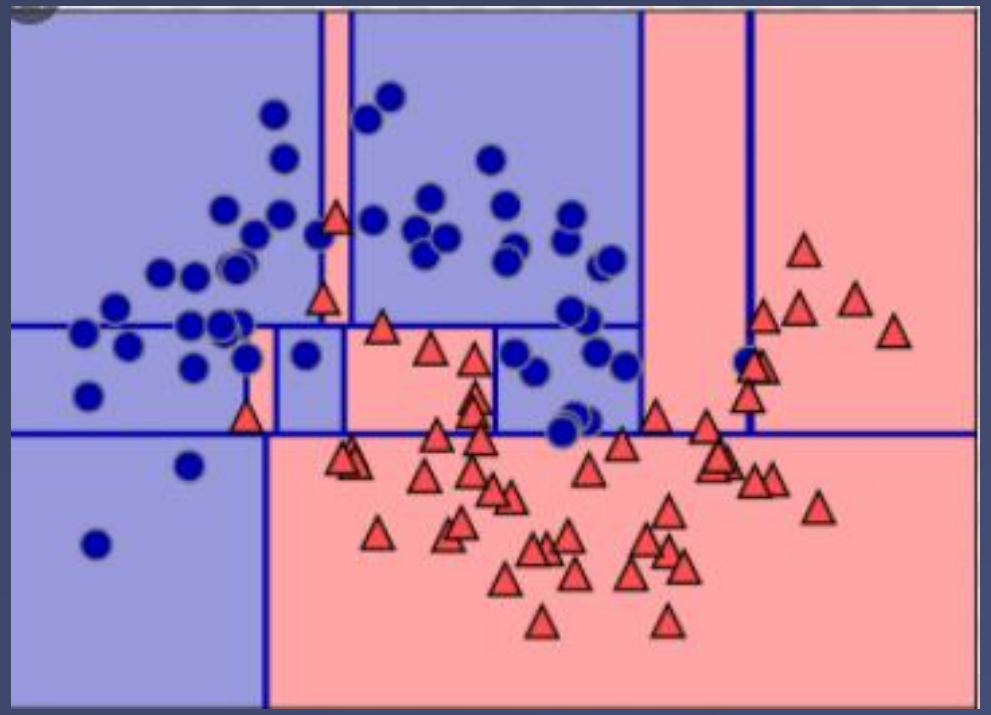
장점

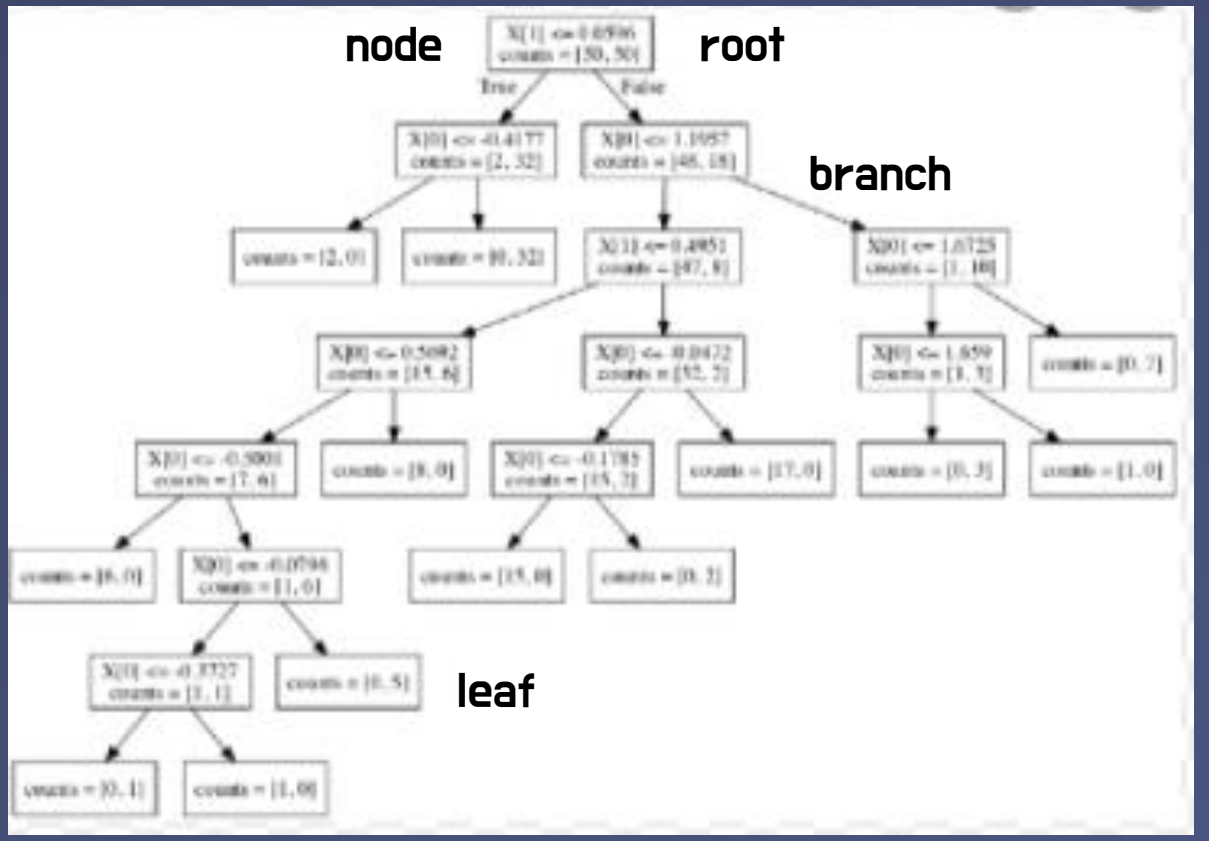
트리를 시각화할 수 있어 설명력이 높음
데이터의 스케일에 구애받지 않음
각 피쳐가 개별적으로 처리되어 각 피쳐의 정규화나 표준화가 필요없음
각 피쳐들에 이진 특성이나 연속적인 특성이 혼합되어 있어도 잘 작동함

단점

사전 가지치기를 사용해도 과대적합됨
(이의 대안으로 앙상블 모델을 사용)

04 DT





과제

각자 만든 피쳐 KNN, SVM, DT에 넣어보고
최고 성능 내기 위해 노력하기

test 데이터 예측한 값 csv파일로 제출

제출 기회 총 3번 있음!!
(9/23 09:00까지)

D&A