데이터마이닝팀

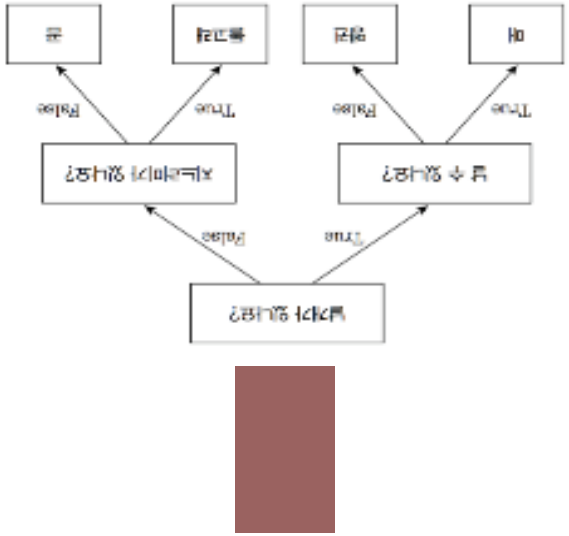
4팀

이진모
이은서
임주은
박지민
장이준

CONTENTS

Basic concept

Decision Tree



의사 결정 나무: 특정 기준이나 질문에 따라 데이터를 구분해주는 알고리즘

Decision Tree Regressor

Decision Tree Regressor

$$RSS = min_{Cm} \sum_{i=1}^{N} (y_i - f(x_i))^2 = min_{Cm} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} c_m I(X \in R_m) \right)^2$$

y_i: i번째 데이터 실제 종속변수값

$\sum_{m=1}^{M} c_m I(X \in R_m)$: i번째 데이터 예측값 = i번째 데이터가 속한 영역의 평균값

c_m = m번째 terminal 노드(영역)의 결과값

→ 회귀문제이므로 예측값 평균

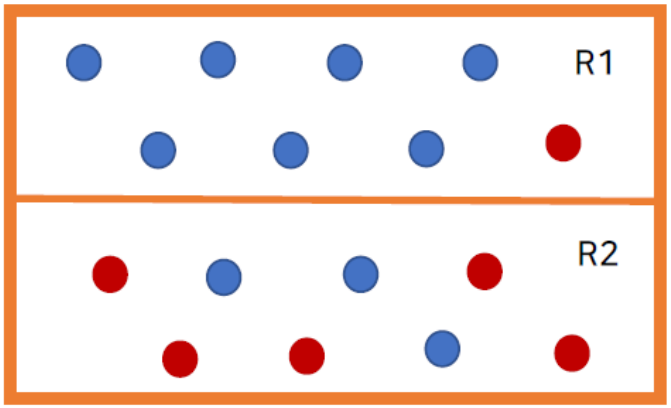
N = 전체 관측값 개수

M = 전체 node의 개수

Decision Tree Classifier

Decision Tree Classifier: Entropy

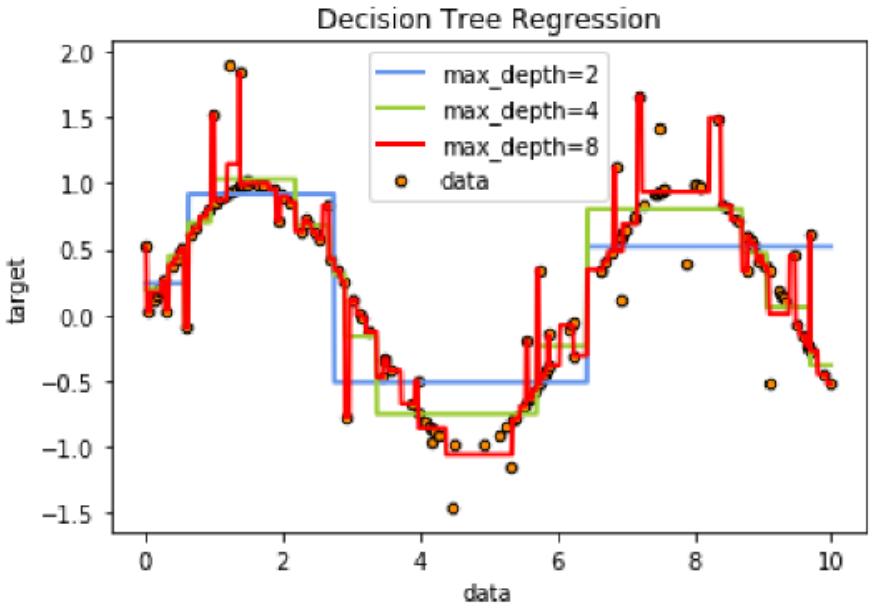
$$\text{Entropy} = -\sum_{k=1}^{k} \hat{p}_{mk} \log_2(\hat{p}_{mk}) \qquad where\, \hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R} I(y_i = k)$$



$$\text{Entropy(X)} = 0.5\ (-\tfrac{7}{8}\log_2\left(\tfrac{7}{8}\right) - \tfrac{1}{8}\log_2\left(\tfrac{1}{8}\right)) + 0.5\ (-\tfrac{5}{8}\log_2\left(\tfrac{5}{8}\right) - \tfrac{3}{8}\log_2\left(\tfrac{3}{8}\right)) \approx 0.75$$

Avoid overfitting in Tree Based Models

controlling depth

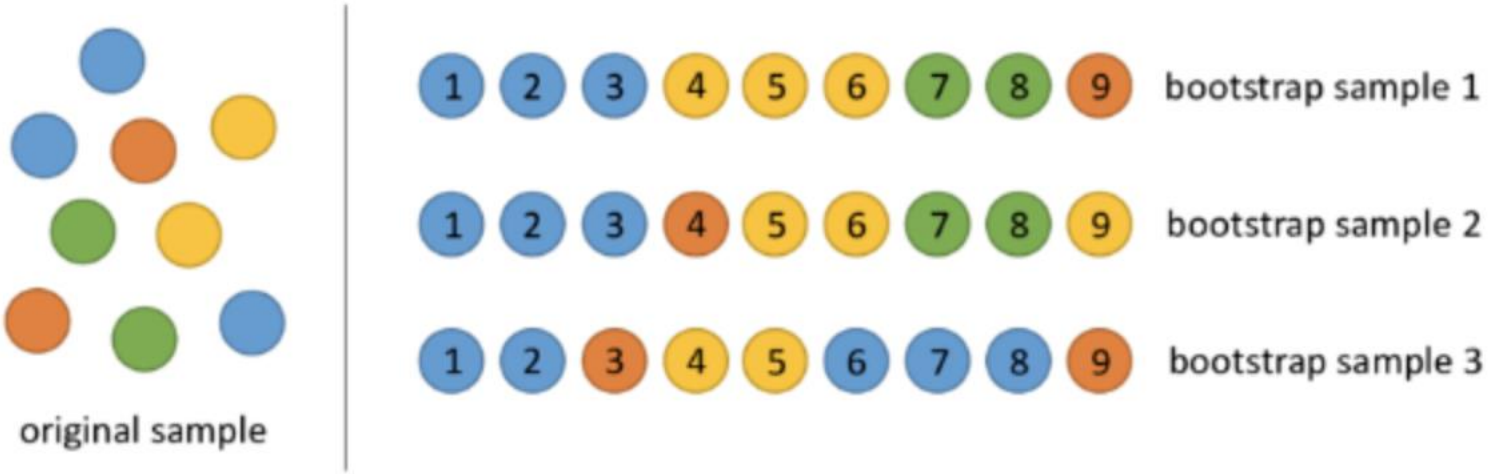


Max_Depth:

분기할 수 있는 최대 깊이를 지정해 줌으로써 트리 기반 모델의 과적합을 방지

Ensemble Methods

Ensemble Methods - Bagging에서의 Bootstrapping



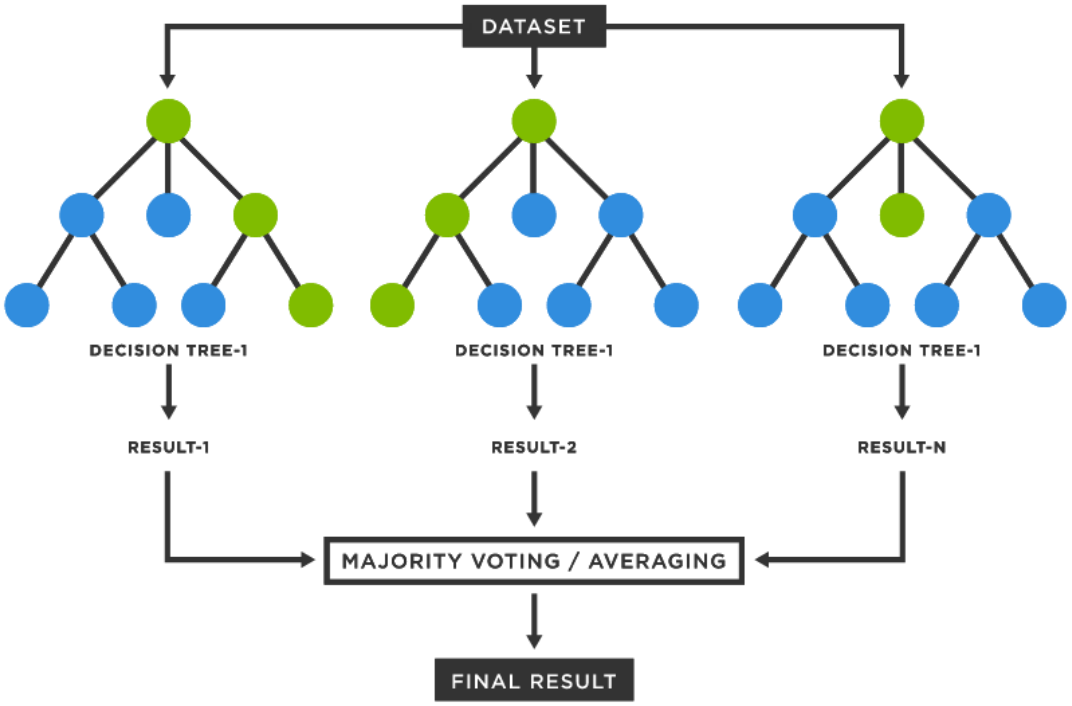
Bootstrapping:

데이터셋으로부터 샘플을 추출할 때

복원추출이 가능하게 해 일부러 샘플마다 중복되는 관측값이 있게 함

Ensemble Methods

Ensemble Methods - 랜덤 포레스트

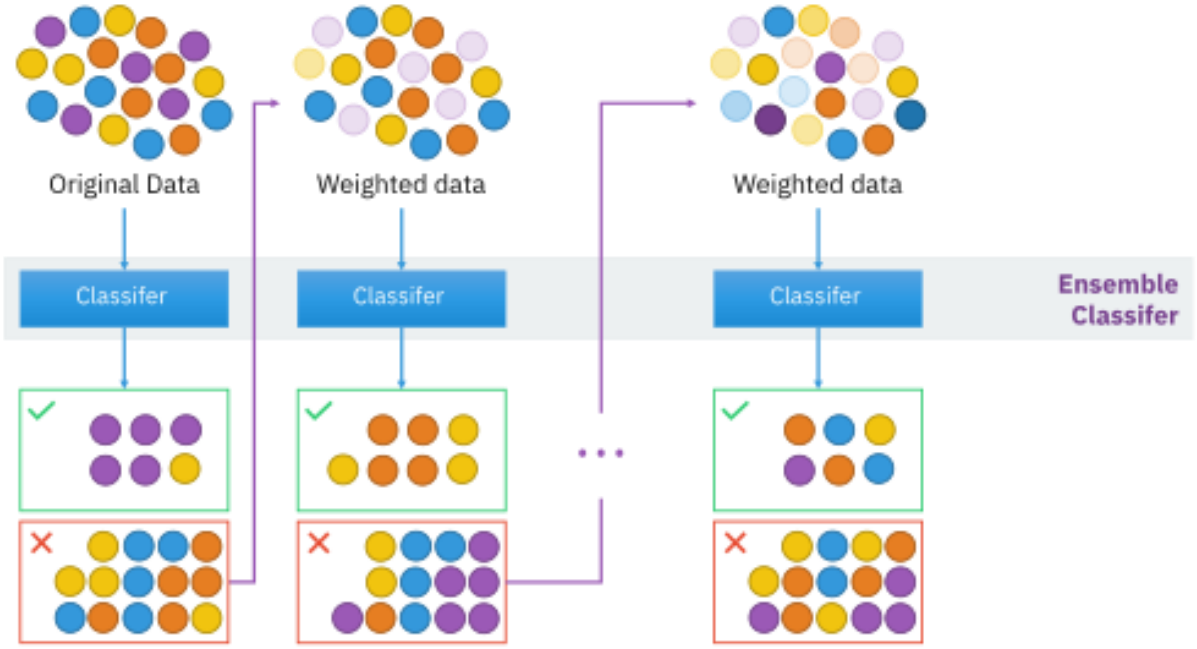


모델링마다 사용되는 feature의 개수를 랜덤하게 선택함으로써

Bagging에서 샘플 간의 높은 상관관계 문제를 해결!

Ensemble Methods

Ensemble Methods - Boosting

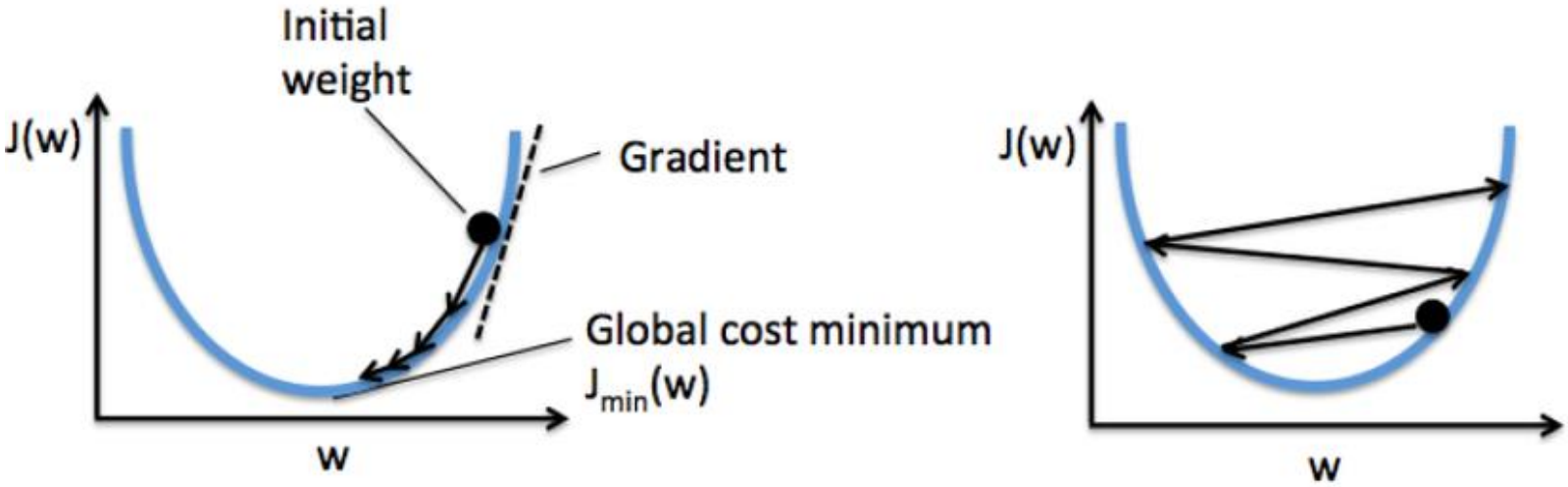


이전 단계에서 잘못 분류된 관측치들에 대해 다음 단계에서 큰 가중치 부여

→ 최종 분류기에 다다를수록 정확한 분류 가능

Ensemble Methods

Ensemble Methods - GBM



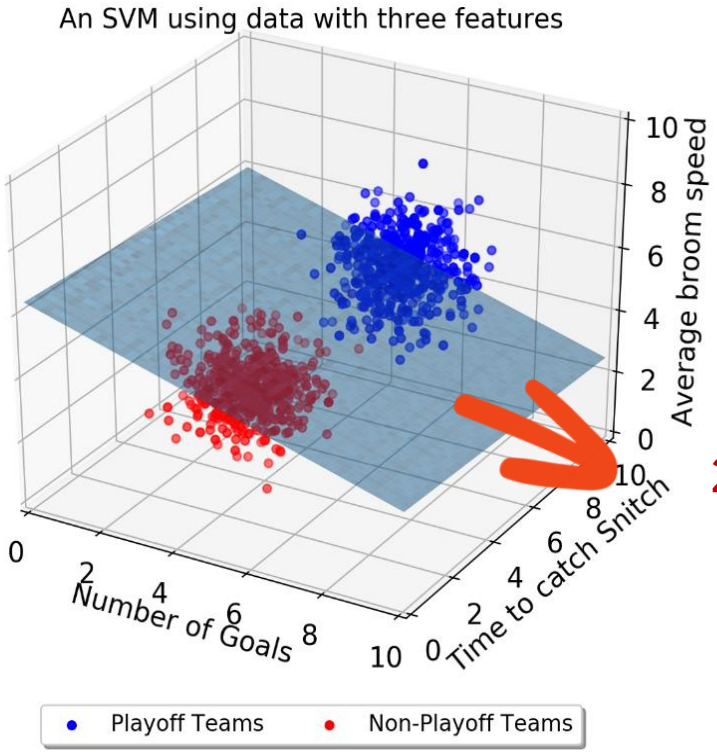
학습률:

각 단계마다 업데이트되는 잔차에 곱해지는 값(아주 작은 값 사용)

GBM이 최솟값을 찾는 과정에서의 '한 걸음의 폭'

Decision Boundary

라벨에 따른 그룹을 구분해주는 선



2차원 decision boundary

3개의 feature, 3차원 공간에 표현

Maximal Margin Classifier

마진을 최대로 갖는 hyperplane 찾아 줌

$$Find\ \beta_0\ and\ \beta\ with\ Max\ M,$$

$$subject\ to\ \beta^T\beta = 1\ and\ y_i(\beta_0 + x_i^T\beta) \geq M$$

M: 마진

β: $(\beta_1, \cdots, \beta_p)'$

x_i: $(x_{i1}, \cdots, x_{ip})'$

x값이 hyperplane을 기준으로

마진에 걸쳐 있거나, 마진 바깥에 위치해야 함

Maximal Margin Classifier

마진을 최대로 갖는 hyperplane 찾아줌

$$Find\ \beta_0\ and\ \beta\ with\ Max\ M,$$

$$subject\ to\ \beta^T\beta = 1\ and\ y_i(\beta_0 + x_i^T\beta) \geq M$$

M: 마진

β: $(\beta_1, \cdots, \beta_p)'$

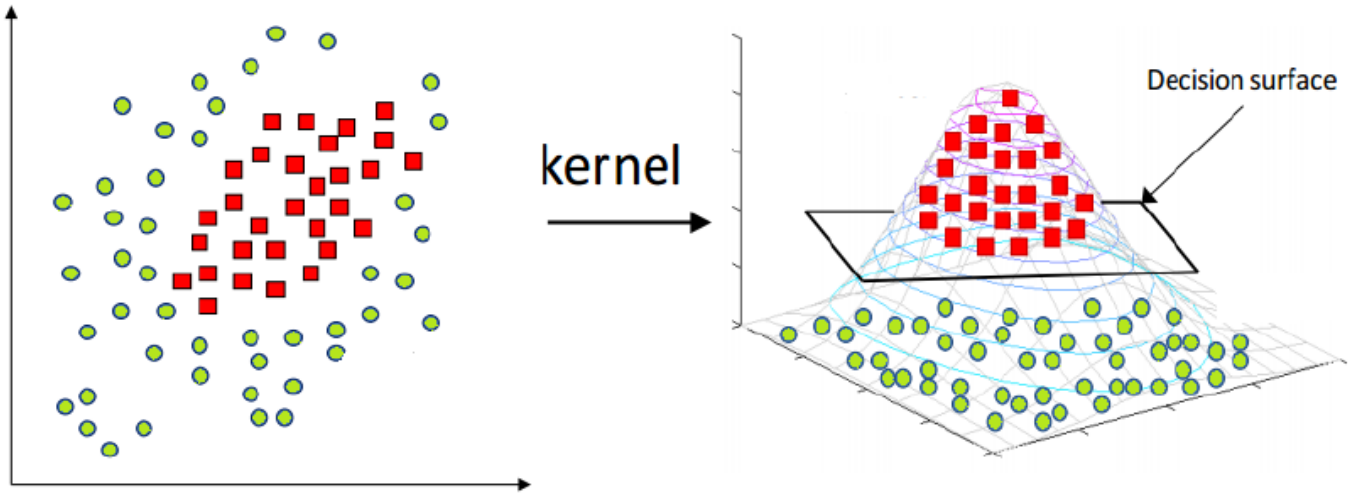
x_i: $(x_{i1}, \cdots, x_{ip})'$

그룹이 완전히 분리되어 분류될 때만 적용 가능

과적합의 위험 높음

Support Vector Machine

RBF Kernel(or 가우시안 커널)



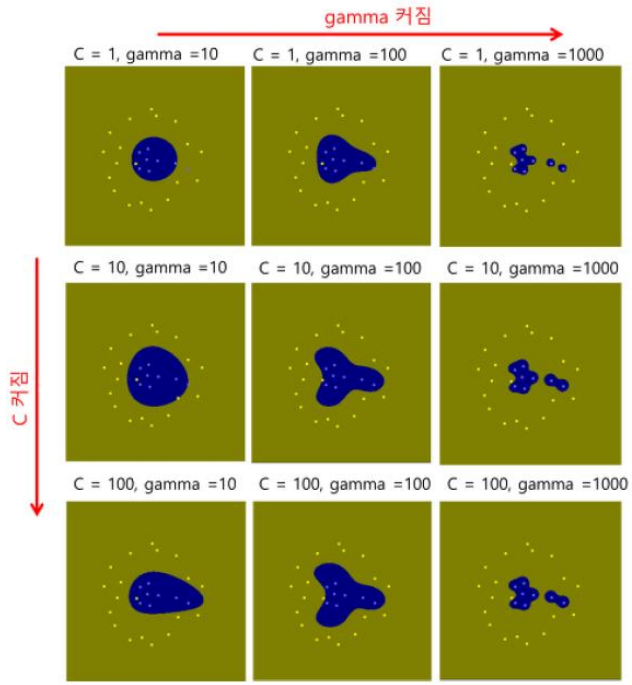
Polynomial Kernel

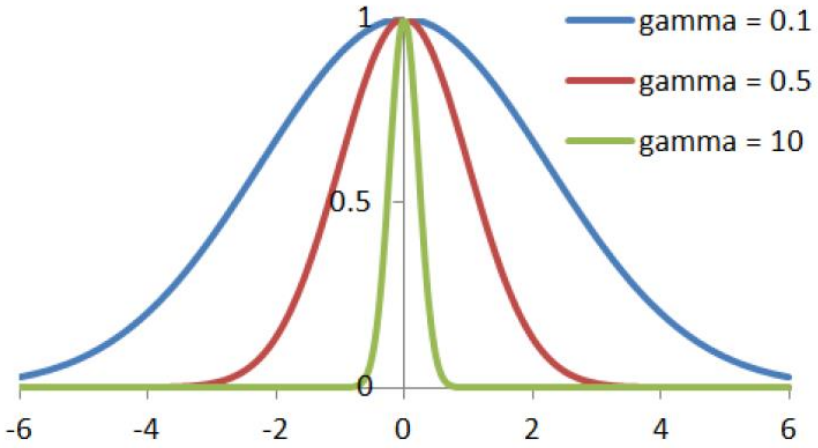
RBF 커널:

차수가 무한대인 polynomial 커널

Support Vector Machine

RBF Kernel 하이퍼 파라미터





Gamma 값이 높으면 데이터가 영향을 미치는 거리가 짧아짐

과적합 위험 ↑