머신러닝(Machine Learning)

앙상블 기법

목 차

01 앙상블 기법

▶ **앙상블(ensemble)는 여러 머신러닝 모델을 연결하여 더 강력한 모델을 만드는 기법**

▶ **랜덤 포레스트(Random Forest)와 그래디언트 부스팅(gradient boosting)**

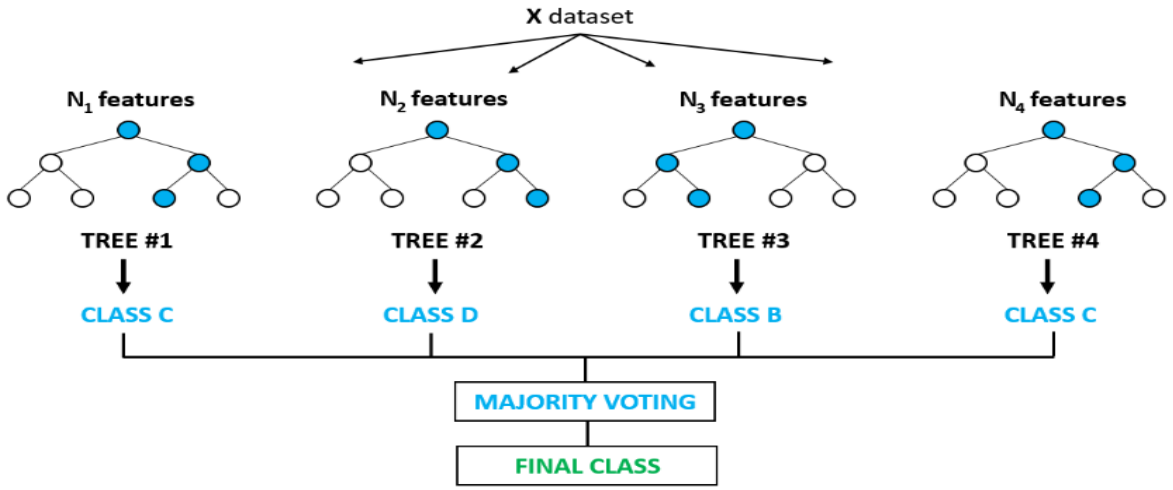
=> 둘 다 모델을 구성하는 기본 요소로 결정 트리를 사용.

02 앙상블 기법-랜덤 포레스트

- **결정 트리의 주요 단점 – 훈련 데이터에 과대 적합되는 경향이 있음.**

  A. 랜덤 포레스트 등장

- **아이디어 : 조금씩 다른 여러 결정 트리의 묶음.**



http://bitly.kr/s2GOAgZm0 참조

02 앙상블 기법-랜덤 포레스트

▶ 결정 트리의 주요 원리

(A) 잘 작동하되 서로 다른 데이터에 대해서 과대 적합된 트리를 많이 만들어 평균을 내면 과대적합을 줄어든다.

(B) 수학적으로 증명됨.

(1) 타깃 예측을 잘 해야 함.

(2) 다른 트리와 구별됨.

=> A. 데이터 포인트를 무작위로 선택

=> B. feature(특성)을 무작위로 선택

03 앙상블 기법 – Gradient Boosting 기법

- 여러 개의 결정 트리를 묶어 강력한 모델을 만든다.
- 분류(Classification)과 회귀(Regression)에 모두 사용 가능.
- 랜덤 포레스트(random forest)와 달리 이전 트리의 오차 보완하는 방식
- 트리가 많을 수록 성능이 좋아짐.
- 랜덤 포레스트보다 매개 변수 설정에 더 민감하여 잘 조정하면 높은 정확도를 얻음.

03 앙상블 기법 – Gradient Boosting 기법

▶ 트리의 깊이가 5정도 깊지 않은 트리를 사용하며 메모리 사용이 적고 예측이 빠름.

▶ 메모리 사용이 적고 예측이 빠르다.

▶ 이전 트리 오차를 얼마나 강하게 보정할 것인가를 제어하는 파라미터 (learning_rate)

04 앙상블 기법 – 모델의 장단점

- 단점

  (1) 매개 변수를 잘 조정해야 한다.

  (2) 학습 시간이 길다.

- 장점 : feature 의 스케일을 조정하지 않아도 된다.

04 앙상블 기법 – 매개변수

- learning_rate : 이전트리의 오차 보정 정도
- n_estimator : 트리의 모델 수

(A) n_estimator : 크면 클수록 좋음(랜덤 포레스트)

(B) n_estimator : 과적합의 가능성(그래디언트 디센트)

- max_depth : 트리 모델의 복잡도

(A) max_depth를 매우 작게 설정하며 트리의 깊이가 5보다 깊어지지 않도록 한다.

- n_estimators을 맞춘 이후에 learning_rate를 찾음.

05 여러가지 모델

- KNN

  작은 데이터 셋, 기본 모델로서 좋고 설명하기 쉽다.

- 선형 모델

  대용량 데이터 셋 가능. 고차원 데이터에 가능

- 나이브 베이즈

  분류만 가능. 선형모델보다 훨씬 빠름. 선형 모델보다 덜 정확함.

- 결정 트리

  매우 빠르고, 데이터 스케일 조정이 필요 없음. 시각화하기 좋고, 설명하기 쉬움.

05 여러가지 모델

▶ 랜덤 포레스트

A. 결정 트리 하나보다 거의 항상 좋은 성능을 냄. 매우 안정적이고 강력.

B. 데이터 스케일 조정 필요 없음. 고차원 희소 데이터에 잘 안 맞음.

▶ 그래디언트 부스팅 결정 트리

A. 랜덤 포레스트보다 조금 더 성능이 좋음.

B. 랜덤 포레스트보다 학습은 느리나 예측은 빠름. 메모리를 조금 사용.

C. 매개변수 튜닝이 많이 필요.

▶ 신경망

특별히 대용량 데이터셋에서 매우 복잡한 모델을 만들 수 있음. 매개변수 선택과 데이터 스케일에 민감.

