13장 방법론 결합: 앙상블과 업리프트모델링

I. 앙상블 (ensembles)

여러 지도학습 기반의 모델들을 하나의 '슈퍼모델'로 결합한 방법

- 왜 앙상블이 예측력을 높일 수 있는가?

방법론 결합은 일반적으로 위험부담을 줄이기 위해 사용

예측 모델에서의 '위험' = 예측오차의 변동 → 예측오차의 변동이 심할수록 모델이 불안정

ex) 크기가 n인 데이터에 대해 2개의 서로 다른 예측모델이 있다고 가정

$e_{1,i}, e_{2,i}$ → 모델 1, 모델 2에 대해 레코드 i의 오차

각 모델이 가지는 예측오차의 평균 : 0 으로 가정 $E(e_{1,i}) = E(e_{2,i}) = 0$

↳ 2가지 방법을 결합해 예측치의 평균값을 택한다면 $\hat{y} = (\hat{y}_{1,i} + \hat{y}_{2,i})/2$

오차의 기댓값도 0 $E(y_i - \hat{y}_i) = E\left(y_i - \frac{\hat{y}_{1,i}+\hat{y}_{2,i}}{2}\right) = E\left(\frac{y_i - \hat{y}_{1,i}}{2} + \frac{y_i - \hat{y}_{2,i}}{2}\right) = E\left(\frac{e_{1,i}+e_{2,i}}{2}\right)$

⇒ 앙상블이 개별적인 모델들과 동일한 평균오차를 가짐

$Var\left(\frac{e_{1,i}+e_{2,i}}{2}\right) = \frac{1}{4}\left(Var(e_{1,i}) + Var(e_{2,i})\right) + \frac{1}{4} \times 2Cov(e_{1,i}, e_{2,i})$ ← 여기에 의존

↳ 두 예측오차 간의 상관관계가 없다면 개별분산 보다 작은 분산을 가지게 됨.

⇒ 두 예측의 평균값을 사용하면 잠재적으로 예측오차의 변동을 줄일 수 있고,

그 결과 예측력을 높일 수 있다.

- 단순 평균

예측, 분류, 성향 모델에서 앙상블을 만들기 위한 가장 간단한 방법은 여러 모델들을 결합.

↳ 결과값들을 결합.

- 예측결합

결과값이 수치로 표현되는 예측문제에서 예측치들의 평균 을 계산하는 방식으로 결합.

단순평균을 대체하는 다른 방법 중 하나는 중앙값 → 극단적인 예측치의 영향을 덜 받음

사용자의 관심도에 비례하는 가중평균 계산

↳ 가중치는 모델의 정확도에 비례할수도 있고, 데이터의 질에 비례할수도 있음.

- 분류결합

여러 분류기로부터 결과들을 결합할 때 '투표' 방식으로 결합

가장 간단한 규칙 : 가장 [illegible] 사용

가장 단순한 개념: 가장 우세한 범주로 분류

모델의 정확도나 데이터의 질 등에 근거하여 특정 모델의 결과값에 더 큰 가중치 부여 가능

- 성향결합

성향 또한 단순 (혹은 가중) 평균을 구하는 방식으로 결합.

↳ 나이브 베이즈와 같은 몇몇 알고리즘은 편향된 성향을 만들기 때문에

다른 방법으로 구한 성향들과 단순평균을 구해서는 안된다.

· 배깅

여러 랜덤 데이터 샘플들의 평균을 기반으로 한다.

배깅 (bagging ; bootstrap aggregating)

① 다수의 랜덤표현들을 생성 (훈련 데이터로부터 복원추출) - 이 방식을 bootstrap sampling

② 각 표본에 대해 알고리즘을 적용하고 분석을 생성

→ 배깅은 모델의 안전성을 향상시키고, 서로 다른 데이터 표본을 개별적으로 모델링

⇒ 과적합 현상을 방지하고 그 결과들을 결합 (나무 모형이나 신경망 모형에 특히 유용)

· 부스팅

boosting : 앙상블을 생성하기 위해 다소 다른 방식으로 접근

↳ 잘못 분류된 레코드들에 대해 모델을 개선함으로서 순차적으로 그 레코드에 대한 성능을 개선

① 데이터에 모델을 적용

② 잘못 분류된 레코드들 (혹은 예측오차가 큰 레코드들)이 더 높은 확률로 선택되도록

데이터로 부터 표본 추출

③ 새로운 표본에 모델을 적용

④ 이 단계들을 2-3번 반복

· 앙상블의 장점과 단점

여러 모델의 점수를 결합하는 이유 → 보다 정확한 (예측 오차의 분산이 작은) 예측을 하기 위해서

앙상블은 결합된 모델들이 음의 상관관계를 가질 때 가장 유용하지만

상관관계가 낮을 때도 유용

단순평균, 가중평균, 투표, 중앙값 등을 이용

모델들은 동일 또는 서로 다른 알고리즘에 기반할 수 있으며, 서로 다른 데이터 표본도 사용가능

⇒ 다수의 데이터 크런처 (data chunchers) 가 힘을 합치고 결과를 종합함으로서

높은 예측력을 가진 해결책에 도달하는 실질적 방법 제공

서로 다른 데이터 표본에 기반한 앙상블은 과적합 방지에 기여.

↳ 그러나 앙상블을 약간만 오용해도 과적합 발생 가능

주된 단점은, (분석자의 숙련도와 시간이 요구된다는 것. → 서로 다른 모델 전부 개발 필요
(예측변수와 출력변수의 관계가 불분명한 '블랙박스' 모델

2. 업리프트 (상승) 모델

- A-B 검정

마케팅 산업에서 각 개인에 대한 결과가 추적될 수 있는 하나의 표준적인 과학적 실험

핵심은 하나의 처리를 다른 처리, 혹은 대조군에 대해 검정하는 것

↳ 처리 (treatment) – 검정 시 우리가 개입하게 되는 것.

A-B 검정의 중요한 구성요소는 랜덤할당으로 처리들은 개인에게 랜덤하게 할당되거나 된다

⇒ 처리 A와 처리 B의 차이점은 (우연이 작용하지 않는 한) 처리에 기인한 것이 됨.

- 업리프트

A-B 검정은 어떤 처리가 평균적으로 뛰어난지는 알려주지만, 특정 개인에 대해 어떤 처리가 가장 적합한지는 알려주지 않는다.

– 개별적 업리프트 모델링

책의 유권자 예제에서

Uplift = 메시지를 받은 후에 호의적인 의견의 성향 (확률) 증가 ← 각각의 유권자에 대해

업리프트 모델을 세울 때, 처리 (ex) 메시지)를 받은 결과 발생하는 '성공 (경향)'의 확률을

예측하기 위해 다음의 과정 진행

① 임의로 표본을 처리를 받은 그룹과 대조군으로 나눈 후 A-B 검정을 시행하고

결과를 기록

② 표본을 다시 합한 후, 학습 데이터와 검증데이터로 분할.

이 결과를 가지고 결과변수와 처리상태를 보여주는 예측변수를 포함하여 예측모형 세운다.

③ 검증데이터를 사용하여 각 유권자에 대해 예측모델의 점수를 매긴다.

이것은 처리가 주어졌을 때, 각 검증레코드에 대해 성공확률을 예상한다

④ 처리변수의 값을 반대로 취하고, 검증데이터에 대해 같은 모델로 점수를 다시 구한다.

⇒ 각 검증레코드에 대해 다른 처리를 받았을 때의 성공성향을 산출.

⑤ 개개인에 대한 uplift는 다음에 의해 추정

$P(Success \mid 처리=1) - P(Success \mid 처리=0)$

⑥ 실험이 행해지지 않은 새로운 데이터에 대해서는 처리에 대한 예측변수를 생성하고, 처음에 '1'을 부여, 점수 계산, '0' 부여, 점수 계산. 위와 같은 방법으로 새로운 레코드에 대한 uplift 성향을 추정

- 업리프트 모델 결과의 이용

각 개인에 대한 성향변화정도가 '업리프트'로 추정되면 그 결과는 업리프트 정도에 따라 정렬.

↳ 주로 마케팅, 정치선전활동등에 사용. (누구에게 설득 메세지를 보낼지, 내버려둘지 결정
보내기로 했다면 몇몇 후보들에게는 어떤 메세지를 보낼지

3. 요약

실제로는 위 두가지 방법이 단독으로 쓰이는 경우는 드물며, 주로 정보를 얻고 통찰력을 제공하는 것을 목표로 분석과정의 밑바탕을 이루는데 사용

앙상블 : 예측성능을 향상시키기 위해 여러 모델들을 가중치를 사용해서 결합.

업리프트 모델링 : A-B 검정의 결과를 제안 또는 설득하는 메세지를 보낼지의 여부뿐만 아니라 누구에게 보내야 할 지와 같은 선택들을 안내하기 위한 예측변수로서 사용 → 예측모델링 과정에 포함.