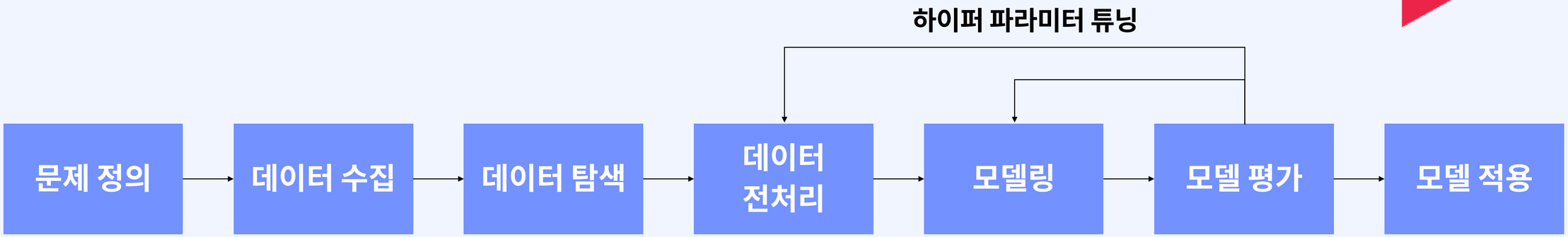
2. 문제 정의와 데이터 수집

1 머신러닝 프로세스 개요

머신러닝 프로세스



실전에서는 모델 평가에서 그 결과가 마음에 들지 않거나 하면 이전 단계로 되돌아가서 성능을 만족할 때까지 **반복적인 작업**을 함

머신러닝 프로세스 (계속)

(1) 문제 정의	어떠한 문제를 해결할지 정하고 대략적인 분석 계획을 수립
(2) 데이터 수집	정의한 문제를 해결하는데 필요한 데이터를 수집
(3) 데이터 탐색	수집한 데이터를 탐색하여 분석 계획을 구체화
(4) 데이터 전처리	원활한 모델링을 위해 데이터를 가공
(5) 모델링	사용자가 지정한 모델 종류와 하이퍼 파라미터에 따른 모델 학습
(6) 모델 평가	모델의 성능을 비롯한 다양한 측면에서의 모델 평가 수행
(7) 모델 적용	최종 모델이 선정되면 실제 환경에 적용

2. 문제 정의와 데이터 수집

2 문제 정의와 데이터 수집

개요

문제 정의와 데이터 수집은 많은 강의와 서적 등에서 설명을 생략하지만 실무에서 가장 중요한 단계라고 할 수 있음

문제 정의 단계에서는 어느 데이터로 어떤 과제를 해결할 것인지를 결정해야 합니다.

데이터 수집 단계에서는 어느 방법으로 데이터를 수집할 것인지 결정하고 수집한 데이터에는 문제가 없는지 등을 확인해야 합니다.

과제 유형 정의

해결하고자 하는 문제가 어느 유형에 속하는지를 정의해야만 분석 계획을 적절하게 수립할 수 있습니다.

과제 유형 정의를 위한 질문 리스트

반드시 머신러닝을 사용해야 하는 과제입니까?

데이터를 사용하지 않아도 되거나 도메인 지식만으로 풀 수 있는 과제가 의외로 많습니다. 그러므로 머신러닝을 반드시 사용해야 하는지를 먼저 판단하고 머신러닝을 사용할 필요가 없다고 판단되면 과감히 머신러닝 사용을 포기해야 합니다.

과제의 목표가 전문적인 지식이 없는 사람이 지금까지 하던 일을 자동화하는 것입니까?

"예"라고 대답한다면, 지도 학습 과제일 가능성이 큽니다. 이러한 과제는 누구나 라벨을 부착할 수 있고 어느 특징이 중요한지 알 수 있어 데이터 수집 및 전처리가 어렵지 않을 가능성이 큽니다.

새로운 인사이트를 발견하는 것이 목표입니까?

"예"라고 대답한다면, 비지도 학습, 통계 분석, 시각화 등과 관련된 과제일 가능성이 큽니다. 이러한 과제의 산출물은 모델보다는 모델을 해석한 결과가 담긴 보고서 형태일 가능성이 큽니다.

과제 유형 정의 (계속)

해결하고자 하는 문제가 어느 유형에 속하는지를 정의해야만 분석 계획을 적절하게 수립할 수 있습니다.

과제 유형 정의를 위한 질문 리스트

전문가가 지금까지 하던 일을 도와주는 것이 목표입니까?

"예"라고 대답한다면, 지도 학습 과제일 가능성이 큽니다. 그러나 전문가만 라벨을 부착할 수 있기에 데이터 수집이 어려울 가능성이 큽니다.

시간만 지나면 정답을 알 수 있는 과제입니까?

주가 예측과 로또 당첨 번호 예측처럼 현재 시점에서는 알 수 없지만, 시간이 지나면 정답을 알 수 있는 문제가 있습니다. 이러한 문제 대부분은 지도 학습 가운데에서도 시간 특성에 영향을 받는 시계열 지도 학습 과제이므로 특징 공학과 모델 선택 시에 시간적 특성을 고려해야 합니다.

지도 학습 과제 같지만 실제 정답을 알기 어려운 과제입니까?

이러한 과제 대부분은 이상 탐지 혹은 추천과 관련됩니다. 이상 탐지는 이상하다고 판단하는 객관적인 근거가 없기에 정답을 알 수 없고, 추천은 추천 대상(사용자)의 선호를 정확히 알 수 없기에 정답을 알 수 없습니다.

사용 데이터 정의

과제 유형이 정의됐다면 특징과 라벨을 중심으로 사용할 데이터를 정의해야 합니다. 만약 과제 유형이 지도 학습 과제라면 아래 질문 리스트를 참고하여 특징과 라벨을 가능한 구체적으로 정의해야 합니다.

사용 데이터 정의를 위한 질문 리스트

데이터를 정말 수집할 수 있습니까?

수집할 수 없는 데이터를 활용해야 하는 과제를 기획하는 경우가 의외로 많습니다. 예를 들어, 사용자별 검색 기록을 바탕으로 범죄 여부 및 시점을 예측하는 과제는 사생활과 관련된 검색 기록을 수집할 수 없어 정상적으로 진행하기 어려울 것입니다.

새로운 데이터의 라벨을 정말 알 수 없습니까?

의외로 많은 과제에서 라벨이 특징과 같이 수집됨에도 불구하고 라벨을 예측하는 모델을 만듭니다. 이러한 모델 개발은 실제로 진행이 필요한 과제라기보다 과제를 위한 과제일 가능성이 큽니다.

특징이나 라벨이 구체적입니까?

모델을 학습하려면 모든 상황이 데이터화돼야 합니다. 따라서 가능한 구체적이고 객관적인 특징을 사용해야 합니다.

사용 데이터 정의 (계속)

과제 유형이 정의됐다면 특징과 라벨을 중심으로 사용할 데이터를 정의해야 합니다. 만약 과제 유형이 지도 학습 과제라면 아래 질문 리스트를 참고하여 특징과 라벨을 가능한 구체적으로 정의해야 합니다.

사용 데이터 정의를 위한 질문 리스트

사용하고자 하는 특징이 샘플마다 차이가 있습니까?

자주 바뀌지 않는 특징(예: 한 팀의 직원 수)을 사용한다면 해당 특징은 사실상 모델 학습에 전혀 도움이 되지 않을 수도 있습니다. 따라서 변동성이 있는 특징을 정의하는 것이 바람직합니다.

특징별 수집 주기가 같습니까? 예를 들어, 어떤 특징은 일별로 수집되는데 어떤 특징은 월별로 수집되지는 않습니까?

특징별 수집 주기가 다르다면, 실제적인 모델 활용이 어려울 수 있습니다. 수집 주기를 가장 긴 주기로 일치시켰을 때 문제가 있지 않을지를 검토해야 합니다.

분석 목표 및 계획 수립

사용 데이터까지 정의했다면 대략적인 분석 목표와 계획을 수립해야 합니다.

문제 정의

- 특정한 수치가 아니라 방향을 나타내는 분석 목표 수립
  (예) 정확도 90% 이상의 모델 X, 예측력이 우선하는 모델 O
- 이 단계에서 구체적인 목표를 세우는 것은 불가능하거나
  가능하더라도 계획대로 진행될 가능성이 매우 희박함

데이터 탐색

- 구체적인 전처리, 모델링, 모델 평가 방안 수립

문제 정의 단계에서 대략적으로 분석 목표와 계획을 수립하고, 데이터 탐색 단계에서 목표와 계획을 구체화해야 합니다.

데이터 수집

데이터 수집 단계에서는 문제 정의 단계에서 정의한 데이터를 수집하고 수집한 데이터에 문제가 없는지 확인합니다.

데이터 수집

- 기존에 수집한 데이터 베이스 활용
- 센서 및 로그 활용
- 설문 조사
- 실험 및 평가 등

수집한 데이터 확인

- 측정 오류 확인
- 문제랑 무관한 데이터 수집했는지 확인
- 특정한 상황이 데이터에 누락됐는지 확인

데이터 수집 단계에서 중요한 것은 데이터 수집 자체가 아니라 수집한 데이터의 문제를 확인하는 것입니다.