FALCON

데이터 분석 교육 자료

with NLP

Natural Language Processing

이성희

2021-07-31

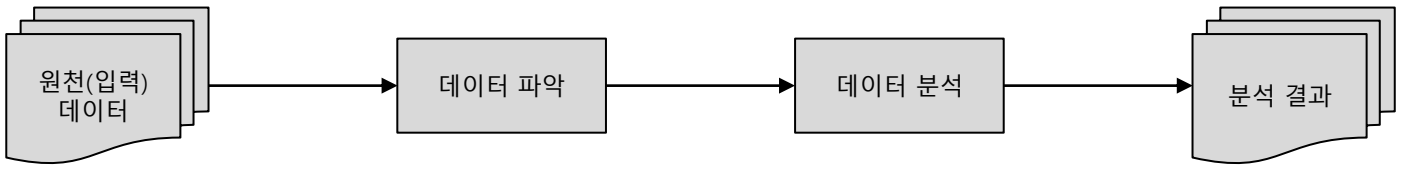
목차

1-1. 데이터 분석 개요 (1/3)

- 데이터 분석이란?
  - 빅 데이터 (big data) : 기존의 데이터베이스 관리 도구 능력을 넘어서는 대량의 정형 또는 데이터베이스 형태가 아닌, 비정형 데이터까지 포함하는 대용량의 데이터 집합

    Three V : Volume (크기), Variety (다양성), Velocity (속도)

  - 데이터 분석 (data analysis) : 다양하게 표현되어 있는 대용량의 데이터 집합으로부터 신속, 효율적으로 의미(데이터가 내포하고 있는 실제 내용)를 파악하고 분석한 뒤, 분석된 결과를 이용하여 가치를 도출하는 일련의 모든 작업



1-1. 데이터 분석 개요 (2/3)

- 데이터 분석을 위한 3가지 정의

① 목적 (goal) 또는 요구 사항 : 무엇을 할 것인가?

- 채팅 시스템, 리뷰 긍/부정 분류
- 스팸 분류, 뉴스 카테고리 분류
- 키워드 추출, 등...

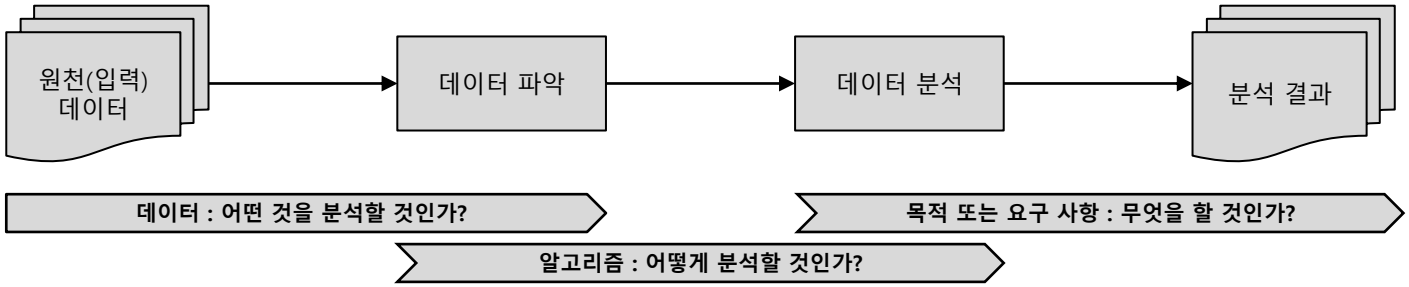
② 데이터 : 무엇을 분석할 것인가?

- 채팅 말뭉치 (발화와 응답의 쌍)
- 네이버 리뷰 데이터, 요기요 리뷰 데이터
- 위키피디아 문서 데이터, 뉴스 데이터, 은행 데이터, 등...

③ 알고리즘 : 어떻게 분석할 것인가?

- 규칙 기반 시스템을 이용할 것인가?
  - LSP : Lexico Semantic Pattern
  - Automata Theory, Other rules, 등...
- 기계 학습 기반 시스템을 이용할 것인가?
  - HMM, CRFs, SVM
  - CNN, RNN, LSTM, BERT, 등...

1-1. 데이터 분석 개요 (3/3)



① 목적 (goal) 또는 요구 사항 : 무엇을 할 것인가?

- 채팅 시스템, 리뷰 긍/부정 분류
- 스팸 분류, 뉴스 카테고리 분류
- 키워드 추출, 등...

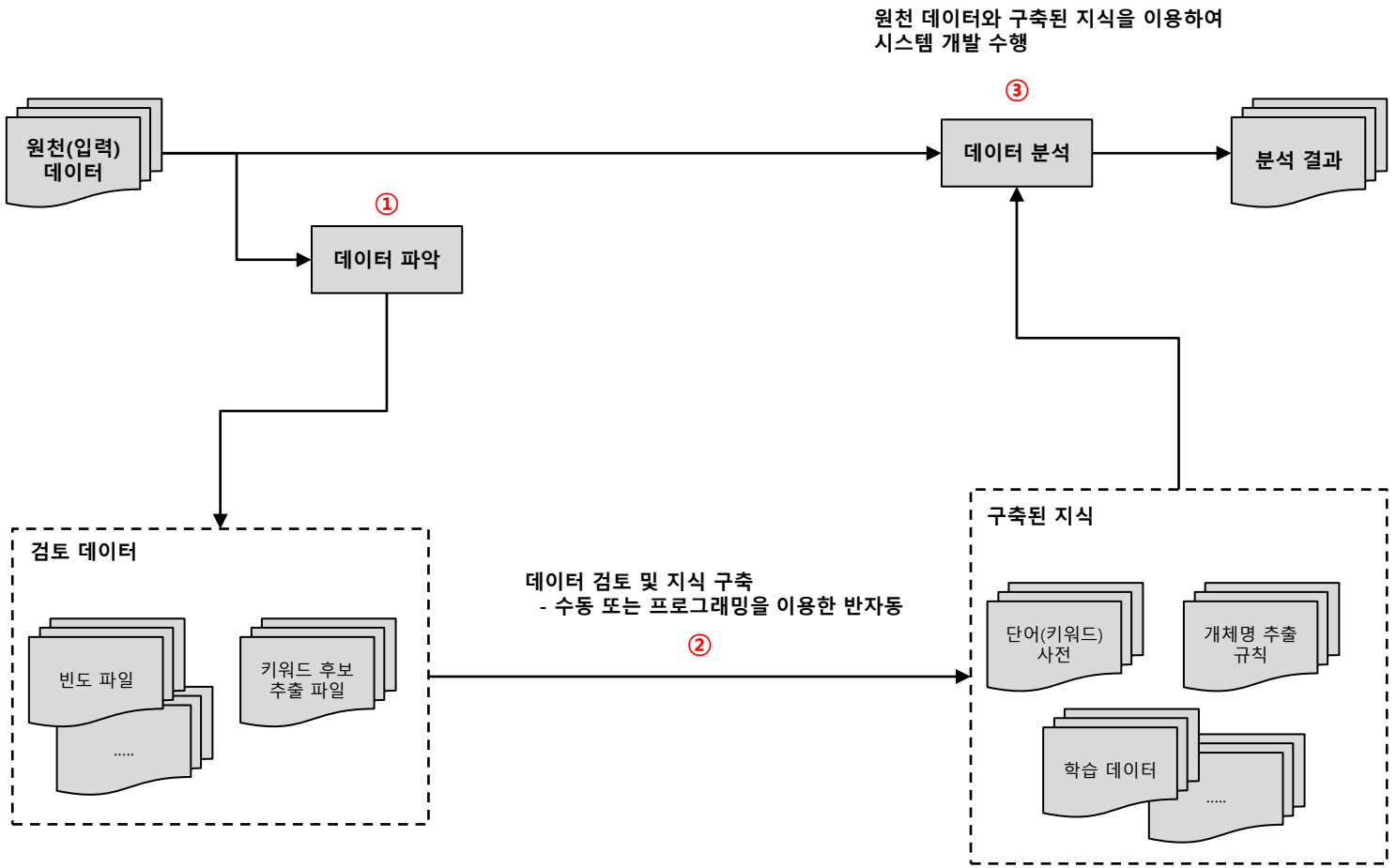
② 데이터 : 어떤 것을 분석할 것인가?

- 채팅 말뭉치 (발화와 응답의 쌍)
- 네이버 리뷰 데이터, 요기요 리뷰 데이터
- 위키피디아 문서 데이터, 뉴스 데이터, 은행 데이터, 등...

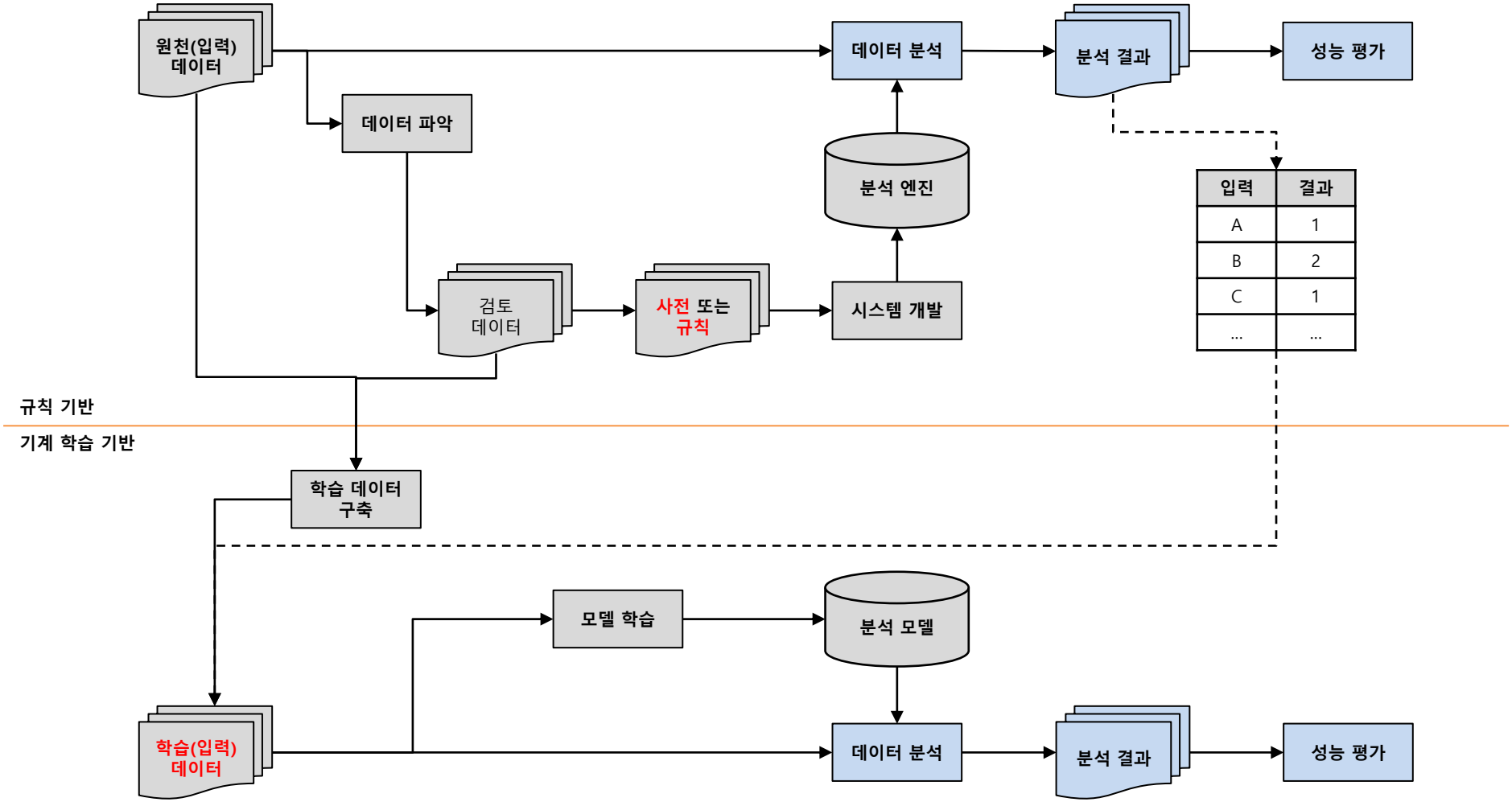
③ 알고리즘 : 어떻게 분석할 것인가?

- 규칙 기반 시스템을 이용할 것인가?
- 기계 학습 기반 시스템을 이용할 것인가?

1-2. 데이터 분석 프로세스 (1/2)



1-2. 데이터 분석 프로세스 (2/2)



목차

2-1. 규칙 기반 시스템을 이용한 데이터 분석 (1/2)

- 규칙 기반 데이터 분석 (rule based data analysis)
  - 데이터 파악 과정을 통해 프로그래밍이 가능한 규칙을 찾아내고,
    찾아낸 규칙을 이용하여 데이터 분석을 수행할 수 있는 시스템을 직접 개발

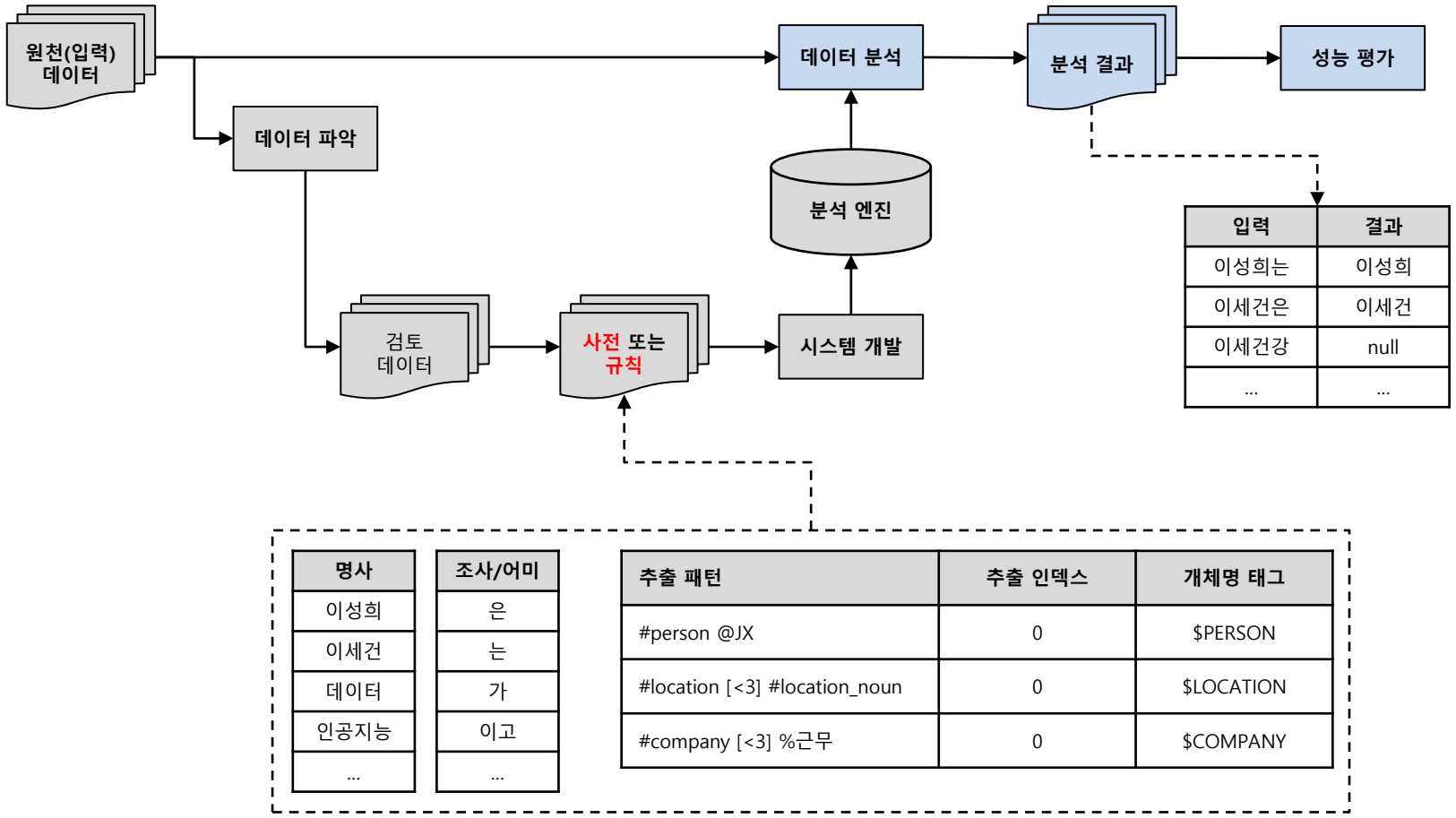
  - 예를 들어, 데이터 분석 과정에서 사람의 이름을 찾아야 한다면 다음과 같은 간단한 규칙을 생각해 볼 수 있다.

    Rule example).
    - 형태소 분석 결과에서 '3'음절의 명사 뒤에 조사 또는 어미가 왔을 때, 명사의 첫 번째 음절이 '성'을 나타내는 경우
      → '이성희는'
      → '이성희/NNP + 는/JKB'
      → '이성희' = 사람의 이름

  - 위와 같이 프로그래밍이 가능한 규칙을 데이터로부터 찾아내고, 이를 이용한 규칙 기반 시스템을 개발함으로써
    데이터 분석 업무를 수행할 수 있다.

2-1. 규칙 기반 시스템을 이용한 데이터 분석 (2/2)

- 규칙 기반 데이터 분석 프로세스



입력	결과
이성희는	이성희
이세건은	이세건
이세건강	null
...	...

명사
이성희
이세건
데이터
인공지능
...

조사/어미
은
는
가
이고
...

추출 패턴	추출 인덱스	개체명 태그
#person @JX	0	$PERSON
#location [<3] #location_noun	0	$LOCATION
#company [<3] %근무	0	$COMPANY

2-2. 기계 학습 기반 시스템을 이용한 데이터 분석 (1/8)

- 기계 학습 기반 데이터 분석 (machine learning based data analysis)
  - 사람의 힘으로 처리하기 힘든, 많은 규칙이 필요한 경우에는 프로그래밍이 어려워진다.
  - 개발자가 직접 규칙을 정의하고 프로그래밍하는 것이 아니라, 모델 스스로가 규칙들을 학습
    - 모델 : 데이터에서 자동으로 규칙들을 학습하는 프로그램

  - 기계 학습을 위해서는 모델이 학습을 수행하기 위한, 학습 데이터(training data)가 반드시 필요하다.
    - 학습 데이터의 형식에 따라 지도 학습(supervised learning)과 비지도 학습(unsupervised learning) 방법이 있다.

<table>
<tr><th></th><th>지도 학습 (supervised learning)</th><th>비지도 학습 (unsupervised learning)</th></tr>
<tr><td>학습 방법
목표</td><td>• 정답을 알려주면서 학습을 시키는 것
• 입력 데이터와 정답 간에 매핑할 함수를 학습</td><td>• 정답을 알려주지 않고 학습을 시키는 것
• 주어진 데이터가 가지고 있는 숨겨진 패턴을 학습</td></tr>
<tr><td>활용</td><td>• 회귀 (regression), 분류 (classification)</td><td>• 클러스터링 (clustering), 특성 학습 (feature learning)</td></tr>
<tr><td>학습 데이터
예시</td><td>
<table>
<tr><th>입력 (input data)</th><th>정답 (label)</th></tr>
<tr><td>너무 재밌어요. 또 보고 싶네요.</td><td>긍정</td></tr>
<tr><td>완전 대박 스토리 연기 아주 만족만족</td><td>긍정</td></tr>
<tr><td>딱히 재밌진 않았던 것 같아요</td><td>부정</td></tr>
<tr><td>스토리가 좀 너무 뻔하달까?</td><td>...</td></tr>
</table>
</td><td>
<table>
<tr><th>입력 (input data)</th></tr>
<tr><td>너무 재밌어요. 또 보고 싶네요.</td></tr>
<tr><td>완전 대박 스토리 연기 아주 만족만족</td></tr>
<tr><td>딱히 재밌진 않았던 것 같아요</td></tr>
<tr><td>스토리가 좀 너무 뻔하달까?</td></tr>
</table>
</td></tr>
</table>

2-2. 기계 학습 기반 시스템을 이용한 데이터 분석 (2/8)

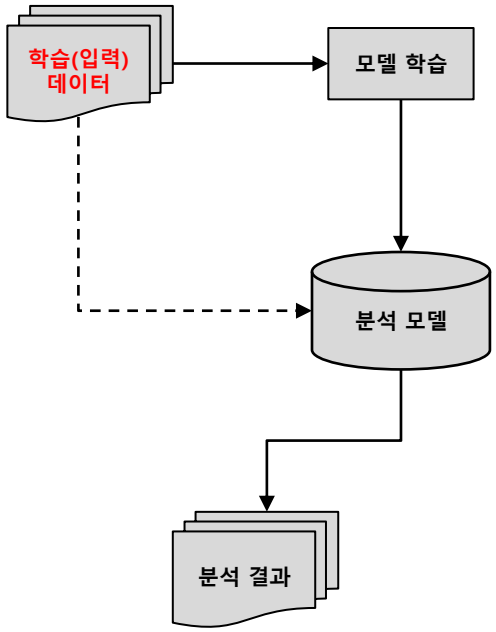
- 기계 학습 기반 데이터 분석 (machine learning based data analysis)
  - 일반적으로 비지도 학습으로는 요구 사항에 맞는 분석 결과를 얻기 힘들다.
  - 대부분의 기계 학습을 이용한 데이터 분석 과제는 지도 학습으로 해결하는 경우가 많다.
  - 지도 학습으로 문제를 해결하려면, 반드시 정답이 부착된 학습 데이터가 필요하다.
    - 정답이 부착된 학습 데이터를 이용하여, 기계 학습을 수행하는 것을 지도 학습이라고 부른다.
    - 학습 데이터에 정답이 부착되어 있지 않다면, 지도 학습을 수행하는 것은 불가능하다.
  - 긍/부정 분류 모델을 위한 학습 데이터의 정답은 긍정과 부정 두 가지 경우로 매우 단순하지만, 뉴스 카테고리 분류 모델을 위한 학습 데이터라면 정답의 경우의 수는 훨씬 많아질 것이다.
    - 뉴스 카테고리의 수는 긍정과 부정 두 가지 경우보다 훨씬 다양하다. (정치, 경제, 사회, 생활, 문화, 스포츠, IT 등...)
    - 정답의 경우의 수가 많아지면, 학습을 위한 과정도 복잡해지며 모델의 성능도 평균적으로 내려간다.
      - 일반적으로 2개 중에서 1개를 맞추는 것보다, 10개 중에서 1개를 맞추는 것이 훨씬 어렵기 때문
  - 그렇다면, 지도 학습을 위한 정답이 부착된 학습 데이터가 없는 경우에는 어떻게 분석을 수행할까?

2-2. 기계 학습 기반 시스템을 이용한 데이터 분석 (3/8)

- 지도 학습을 위한 정답이 부착된 학습 데이터를 구축하는 방법
  - ① 사람이 직접 모든 학습 데이터를 검토하여, 정답을 부착
    - 일반적으로 대용량의 데이터가 주어지므로, 이를 사람이 전부 확인하여 정답을 부착하는 것은 매우 어려운 작업
    - 데이터의 수가 100만 건이고 정답의 수가 10개라도 소수의 인원이 직접 검토하여 정답을 부착하는 것은 사실상 불가능
  - ② 비지도 학습을 수행하여, 비지도 학습의 결과를 사람이 검토하고, 검토한 내용을 토대로 정답을 일괄 부착
    - 비지도 학습의 결과로 10개의 그룹이 생성되었다면, 그룹 별로 정답을 할당
    - 여기서 문제는 정답의 수와 그룹의 수가 다른 경우에는, 검토에 많은 비용이 들 수 있다.
      - 정답의 수는 10개인데, 그룹은 2개가 생성되었다면? → 2개의 그룹을 10개로 다시 나누는 작업 필요
      - 정답의 수는 2개인데, 그룹은 10개가 생성되었다면? → 10개의 그룹을 2개로 다시 나누는 작업 필요
  - ③ 규칙 기반 시스템의 분석 결과를 지도 학습의 학습 데이터로 사용
    - 규칙 기반 시스템의 분석 성능이 일정 수준 이상일 때, 이 시스템을 신뢰할 수 있다고 가정하며, 분석 결과를 정답으로 사용

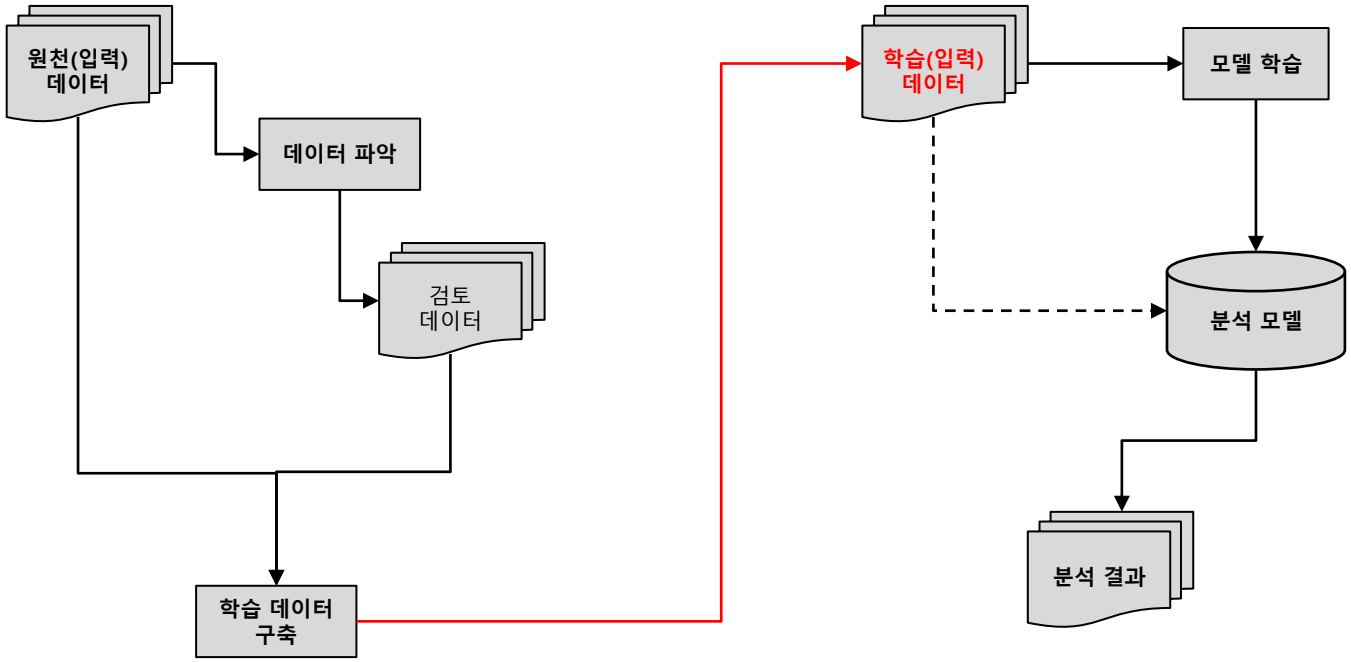
2-2. 기계 학습 기반 시스템을 이용한 데이터 분석 (4/8)

- 기계 학습 기반 데이터 분석 프로세스
  - 정답이 부착된 학습 데이터가 있는 경우



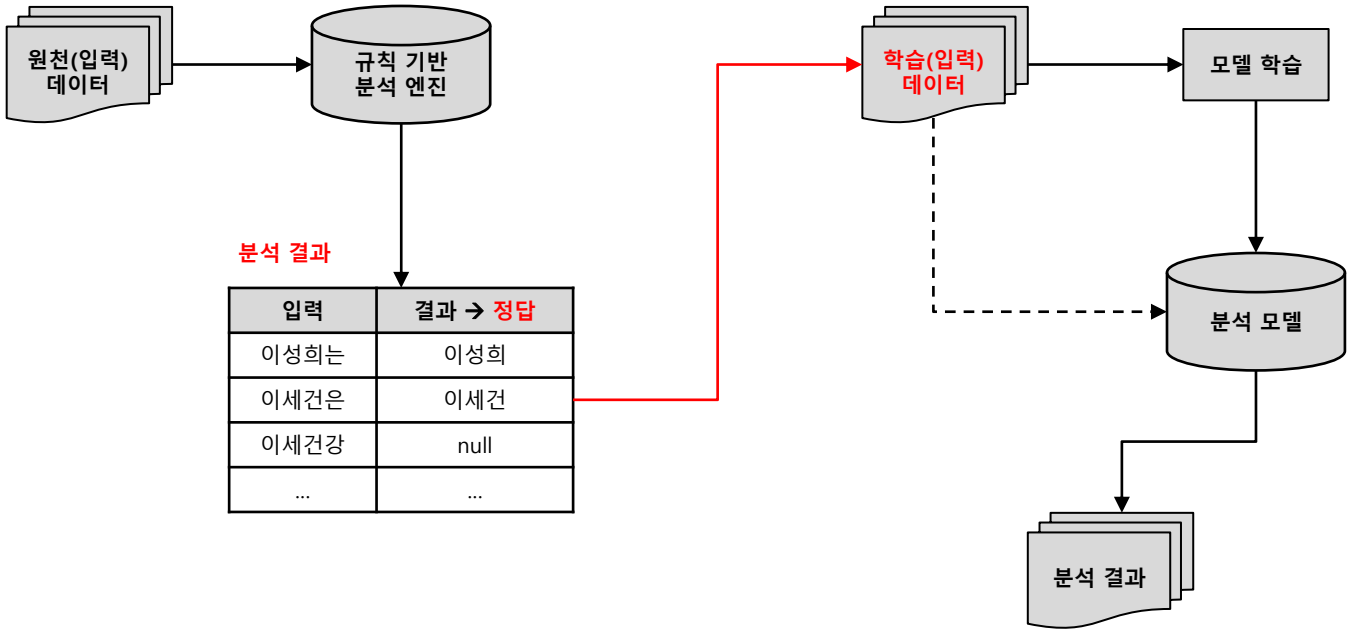
2-2. 기계 학습 기반 시스템을 이용한 데이터 분석 (5/8)

- 기계 학습 기반 데이터 분석 프로세스
  - 정답이 부착된 학습 데이터가 없는 경우
    - ① 사람이 직접 모든 학습 데이터를 검토하여, 정답을 부착
    - ② 비지도 학습을 수행하여, 비지도 학습의 결과를 사람이 검토하고, 검토한 내용을 토대로 정답을 일괄 부착



2-2. 기계 학습 기반 시스템을 이용한 데이터 분석 (6/8)

- 기계 학습 기반 데이터 분석 프로세스
  - 정답이 부착된 학습 데이터가 없는 경우
    - ③ 규칙 기반 시스템의 분석 결과를 지도 학습의 학습 데이터로 사용

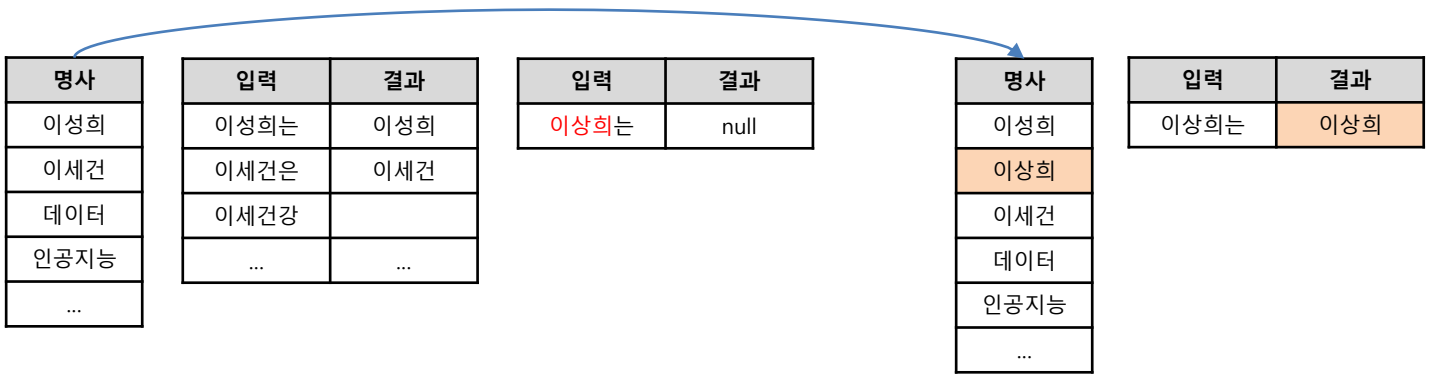


2-2. 기계 학습 기반 시스템을 이용한 데이터 분석 (7/8)

- 정답이 부착된 학습 데이터가 없는 경우
  - ③ 규칙 기반 시스템의 분석 결과를 지도 학습의 학습 데이터로 사용

- 이미 규칙 기반 시스템의 분석 성능이 일정 수준 이상일 때, 이 시스템을 신뢰할 수 있다고 가정했다.
- 그런데 굳이 왜 규칙 기반의 분석 결과를 이용하여 기계 학습을 수행하는가?

- 규칙 기반 시스템의 신뢰도는 현재의 규칙과 입력 데이터에 한해서만 신뢰할 수 있다.
  - 기존 규칙에 해당하지 않는 새로운 데이터가 들어오면, 새로운 규칙을 반드시 정의
  - 규칙의 수는 점점 증가할 것이며, 지속적인 관리 필요 (유지 보수)



2-2. 기계 학습 기반 시스템을 이용한 데이터 분석 (8/8)

- 완성된 규칙 기반 시스템과 비교했을 때, 기계 학습의 가장 큰 장점은 지속적인 관리의 필요성을 어느정도 해결
  - 기계 학습도 새로운 데이터가 너무 많이 발생하면, 새로운 학습 데이터를 구축하고 기존 학습 데이터와 병합하여 재 학습 수행

- 기계 학습은 학습 데이터에서 무수히 많은 규칙들을 스스로 학습하기 때문에,
  학습 데이터에서 관찰되지 않은 입력에 대해서도 분석할 수 있는 가능성을 가진다.

입력	정답
이성희는 학교에 다닌다	이성희
이성희는 바이브에 다닌다	이성희
이성희는 멘토이다	이성희
이성희는 멘티들을 도와준다	이성희
이세건은 학교에 다닌다	이세건
이세건은 바이브에 다닌다	이세건
이세건은 멘토이다	이세건
이세건은 멘티들을 도와준다	이세건
최성희가 학교에 다닌다	최성희
최성희가 바이브에 다닌다	최성희
…	…

입력	예측
이성희는 학교에 다닌다	이성희
이성희는 바이브에 다닌다	이성희
이세건은 멘토이다	이세건
최성희는 멘티들을 도와준다	최성희
최성희가 멘티들을 도와준다	최성희
…	…
이상희는 학교에 다닌다	**이상희**
이상희는 멘토다	**이상희**
이상희 멘티들 도와줌	**이상희**
…	…
…	…