**학습 데이터 수집하기**

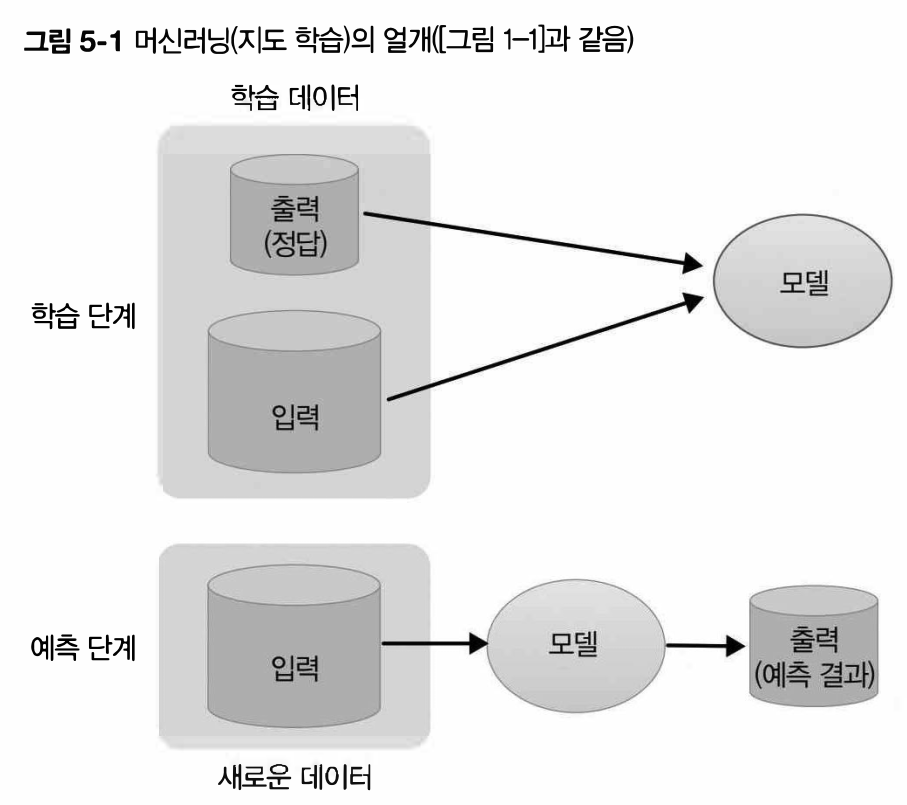
**학습 데이터 얻는 방법**

훈련데이터는 어떻게 구성되는 가? 크게 보면 2가지 정보.

1. 입력 : 열람 로그에서 추출한 특징
2. 출력 : 분류 레이블이나 예측값

출력 값의 경우 보통 heuristic적인 방법으로 찾아 나간다. 출력 레이블 혹은 예측값은 다음과 같은 방법으로 부여한다.

* 서비스 도중 발생하는 로그를 수집 (완전 자동으로 label을 붙인다)
* 사람이 직접 label을 붙인다. (수동)
* 기계적으로 label을 붙인 뒤, 사람이 확인 (자동 + 수동)



**훈련 데이터를 만드는 주체가 누구인지에 따라 설명**

1. 공개된 데이터셋이나 모델 활용
2. 개발자가 직접 작성
3. 동료나 친구에게 데이터 입력을 부탁
4. 크라우드소싱 활용
5. 서비스에 수집 기능을 넣고, 사용자가 입력하게 함.

* **공개된 데이터셋이나 모델 활용**

학습을 마친 공개 모델을 사용하거나, 경진대회용 데이터셋 사용하는 방법.

공개된 데이터 셋으로는 UCI 머신러닝 저장소와 머신러닝 경진대회 플랫폼인 캐글

이미지 인식 분야에서는 이미지넷등이 있다.

위키백과

데이터는 아니지만 모델을 사용하는 방법

딥러닝 라이브러리인 CAFFE에는 이미 학습된 모델을 공유하는 모델 ZOO라는 기능.

텐서플로도 물체 인식용으로 학습된 모델 제공.

이런 방법들을 사용할 때 주의할 점.

1. 모델이나 데이터셋의 라이선스가 상업적 이용을 허용하는가?

데이터는 주로 연구비를 투자하여 만들기 때문에 라이선스를 연구목적으로만 제한하는 경우가 많다. 이런 라이선스는 오픈소스 소프트웨어 라이선스처럼 표준화되어 있지 않다. 재배포를 하든 하지 않든, 원 정보명을 명확히 관리하는 것이 좋다.

1. 이미 학습된 모델이나 데이터셋이 원하는 분야에 적용가능한가?

배포된 데이터의 도메인이 사용 목적과 다르다면, 어떤 식으로든 가공을 거쳐야 한다. 반지도 학습(semi-supervised learning)이나 전이 학습(transfer learning)을 공부해보자.

* **개발자가 직접 만드는 데이터셋**

해결하는 문제가 분류 문제인지 회귀 문제인지 고려한다.

Ex) 소셜북마크 서비스에서 카테고리 예측 문제.

예측하는 문제이므로 분류 문제. “정치”, “연예”, “기술”등의 카테고리 정의. 각 카테고리에 속하는 컨텐츠 1000건 정도 수집. 사람이 직접 분류.

훈련데이터를 작성할 때는 사람이 직접 푼다면 어떤 정보를 사용할지를 주의깊게 생각해봐야 한다. 예를 들어, 뉴스 기사의 제목이 “아이돌이 장관과 선수 응원 이벤트에 갔다”라면 이 뉴스를 “연예”로 분류할 지, “정치”로 분류할 지 어렵다. 어떤 결정을 내리냐에 따라 사용하게 될 알고리즘과 예측 방법이 달라지게 된다.

이렇게 찾아낸 분류 기준이나 분류가 어려운 콘텐츠 등은 따로 정리해두도록 한다.

자신 혼자만 느낌대로 작업하게 되면 편견이 개입되기 쉬워 사용자가 느끼는 것과 어긋나는 결과물이 될 수 있으므로 주의해야 한다.

* **동료나 친구에게 데이터 입력을 부탁**

데이터가 대량으로 필요하다면 동료나 친구에게 도움을 받을 수 있다. 가장 단순한 방법은 구글 스프레드시트를 통해 협업하는 것이다.

여러 명에게 작업을 나누어 맡기는 경우라면 사전에 데이터 내용을 말로 표현할 수 있도록 한다음, 작업 내용과 분류 기준을 잘 설명해야 한다. 혼자 작업할 때와 달리 여럿이서 할 때는 암묵적인 기준이 일관되게 적용되는 경우가 드물기 때문이다. 특히 분류 문제는 대상을 가능한 명확히 나타내야 고품질 데이터를 얻을 수 있으니 분류 기준을 문서화 한다.

여럿이 작업할 때는 같은 데이터에 여러 명이 정답을 부여하도록 하는 것도 중요하다. 정답 레이블의 방향성이 일치하도록 기준을 만든다 해도, 여전히 사람에 따라 판단에 차이가 나기 마련이다. 사람끼리 50%이상 일치하지 않는 문제라면 머신러닝으로도 풀 수 없을 가능성이 매우 크다. 또, 우연히 일치할 가능성을 고려한 기준인 카파계수로도 문제의 난이도를 판단할 수 있다.

한편, 다른 작업자가 매긴 정답 데이터를 서로 보지 못하게 해야 한다. 작업자들이 서로의 결과를 보게 되면 그 자체가 편견(편향)으로 작용할 우려가 있다.

* **크라우드 소싱 활용**

크라우드 소싱이라고 하면 랜서즈나 크라우드 웍스처럼 불특정 다수의 사람이 서로 경쟁하는 방식도 있지만, 아마존 매커니컬 터크나 야후! 크라우드소싱 같은 마이크로 태스크 방식도 있다. 마이크로 태스크 방식은 데이터 입력처럼 단시간에 할 수 있는 단순 작업을 의뢰하는 형태인데 불특정 다수의 일반인이 모여 협동을 한다는 점이 경쟁 방식과 가장 큰 차이다.

특히 머신러닝용 훈련 데이터를 만드는 작업과 잘 맞는 방식이다.

장점

* 전문가를 고용하는 것보다 작업이 훨씬 빠르며 비용도 비교적 낮다.
* 작업 속도가 빠르므로 그만큼 시행착오를 여러 번 반복할 수 있다.
* 비용이 낮으므로 여러 사람에게 같은 일을 맡겨 중복성 있는 데이터를 만들 수 있다.

주의할 점

* 작업자가 단시간에 끝낼 수 있어야 하므로 작업을 설계하기 까다롭다.
* 높은 전문성이 요구되는 작업은 절차를 잘 세분화하고 자세히 설명해야 한다.
* 작업 결과의 품질을 높이려면 결과를 주의해서 가공해야 한다.

같은 작업을 여러 명에게 맡겨 중복된 데이터를 얻은 뒤 다수결 정보 레이블을 따로 부여하거나, 미리 연습문제를 풀어보도록 하게 하는 방법도 사용할 수 있다.

또 모든 데이터를 직접 확인할 수 없으므로 확보된 데이터의 품질을 평가하는 방법도 미리 생각해둬야 한다.

* **서비스에 수집기능을 넣고 사용자가 입력하게 함.**

훈련 데이터를 꼭 직접 만들어낼 필요는 없다. 정답데이터를 서비스 사용자로부터 입력받을 수 있다. 넓은 의미에서는 크라우드 소싱이라고 볼 수 있지만, 서비스를 잘 이해하는 사용자의 협력을 받는 다는 점이 매력이다.

이용자 수가 일정 규모 이상이어야 하고 참여자에게 줄 보상도 설계해야 하지만, 정답 데이터로 사용할 데이터를 얻을 수 있다. 아마존은 오래전부터 검색 결과에 대한 피드백을 받는 페이지를 두고 적극적으로 사용자 피드백을 활용해왔다. 로봇 여부를 판별하는 reCAPTCHA도 좋은 예다.