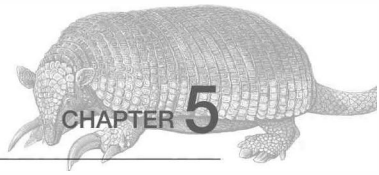


CHAPTER 5

학습 데이터 수집하기

비즈니스에 필요한 새로운 통찰을 얻기 위해 비지도 학습을 사용하기도 하지만, 분류와 회귀 같은 지도 학습이나 추천 시스템에서 높은 성능을 내려면 정답 정보가 포함된 데이터나 말뭉치, 사전처럼 양질의 데이터가 많이 필요하다. 이번 장에서는 지도 학습에 필요한 데이터를 수집하는 방법을 설명한다.

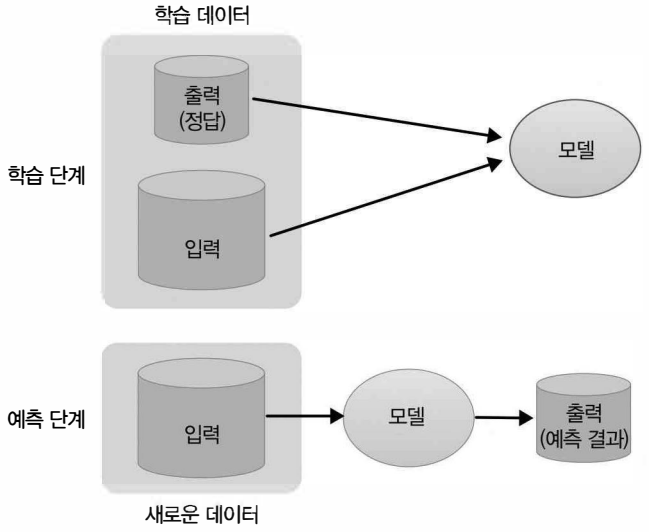
머신러닝을 실제 서비스에 적용하려면 실제 문제를 해결해야 하므로 공개된 데이터셋만으로는 부족할 것이다. 이번 장에서는 조금 깔끔하지는 못해도 머신러닝에서 중요한 훈련 데이터를 만드는 방법을 알아보겠다.

5.1 학습 데이터를 얻는 방법

지도 학습에서 빼놓을 수 없는 것이 훈련 데이터다. 그런데 훈련 데이터는 어떻게 구성될까? 크게 보면 다음의 2가지 정보로 이뤄진다.

- 입력: 열람 로그에서 추출한 특징
- 출력: 분류 레이블이나 예측값

그림 5-1 머신러닝(지도 학습)의 얼개([그림 1-1]과 같음)



특징을 찾는 데 필요한 시행착오는 일찍이 설명한 바 있다. 보통 발견법heuristic적인 방법으로 찾아 나간다. 출력 레이블 혹은 예측값은 다음과 같은 방법으로 부여한다.

- 서비스 도중 발생하는 로그를 수집하여 추출(완전 자동)
- 콘텐츠를 사람이 직접 보며 부여(수동)
- 기계적으로 정보를 부여한 뒤 사람이 확인(자동 + 수동)

이번 장에서는 훈련 데이터를 만드는 주체가 누구인가의 관점에서 설명하려고 한다.

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

5.2 공개된 데이터셋이나 모델 활용

학습까지 완료된 공개 모델을 사용하거나, 경진대회용 데이터셋으로 기본 학습을 마친 모델을 전용하는 방법이다. 이번 절에서는 이러한 데이터를 수집하는 방법을 알아보자.

공개된 데이터셋 소스로는 UCI 머신러닝 저장소[1]와 머신러닝 경진대회 플랫폼인 캐글[2]이 유명하다. 캐글에는 여러 경진대회에 사용됐거나 일반 사용자가 올린 데이터셋이 올라와 있다. 이미지 인식 분야에서는 일반 물체 인식에 쓰이는 이미지넷ImageNet[3] 등 레이블링된 데이터셋도 공개되어 있다. 이 외에도 (직접 학습에 쓰이는 데이터는 아니지만) 딥러닝 라이브러리인 카페Caffe에는 이미 학습된 모델을 공유하는 모델 주Model Zoo[4]라는 기능이 있다. 텐서플로도 물체 인식용으로 학습된 모델을 제공한다.[5]

일본어 텍스트 데이터로는 위키백과 덤프 데이터와 유료로 제공되는 신문 말뭉치가 있다. 이 데이터를 실제 활용한 예로는 위키백과의 형태소분석 사전을 사용한 해커돌[6]을 들 수 있다.

그러나 이 방법들을 사용하는 데는 몇 가지 주의할 점이 있다.

- 모델이나 데이터셋의 라이선스가 상업적 이용을 허용하는가?
- 이미 학습된 모델이나 데이터셋이 원하는 분야에 적용 가능한가?

첫째, 특히 레이블 포함 데이터는 주로 대학 등이 연구비를 투자하여 만들기 때문에 라이선스를 연구 목적으로 제한하는 경우가 많다. 이런 데이터에 쓰이는 라이선스는 오픈소스 소프트웨어 라이선스처럼 표준화되어 있지 않다. 따라서 웹에 공개된 데이터라도 상업적 이용은 불가능한 경우가 종종 있으니 반드시 확인하도록 한다. 또, 데이터를 이용하여 만든 모델을 재배포하는 데도 원 데이터에 따라 제한을 받는 경우가 있다. 재배포를 하든 하지 않든, 원 정보원을 명확히 관리하는 것이 좋다.

둘째, 배포된 데이터의 도메인이 사용 목적과 다르다면 어떤 식으로든 가공을 거쳐야 할 수 있다. 이 주제는 이 책에서 다루지 않으니 관심 있는 독자는 **반지도 학습**semi-supervised learning이나

1 http://archive.ics.uci.edu/ml/
2 https://www.kaggle.com/
3 http://www.image-net.org/
4 https://github.com/BVLC/caffe/wiki/Model-Zoo
5 http://bit.ly/2KeJuIE
6 http://www.slideshare.net/mosa_siru/ss-40136577

전이 학습transfer learning[7]에 관해 공부해보기 바란다. 특히, 물체 인식에서는 전이 학습을 이용하여 기존 학습 모델에 자신이 풀려는 문제의 정답 데이터를 추가로 학습시켜 비용을 절약할 수 있다. 드라마 〈Silicon Valley〉에 나온 핫도그 인식기가 이런 방법으로 만들어졌다.[8]

그러나 기존 데이터셋만으로 풀 수 있는 문제는 비교적 제한적이니, 이제부터 데이터셋을 직접 꾸리는 방법을 알아보도록 하자.

5.3 개발자가 직접 만드는 데이터셋

활용할 수 있는 기존 데이터가 없다면 가장 먼저 개발자가 직접 만들어 내는 방법을 생각해볼 수 있다. 어떤 데이터를 특징으로 삼느냐에 따라 성능이 크게 좌우되니, 직접 수고를 들여 훈련 데이터를 만드는 것이 매우 중요하다.

우선 해결하려는 문제가 분류 문제인지 회귀 문제인지를 고려한다.

예를 들어 소셜 북마크 서비스에서 카테고리를 예측하는 문제를 푼다고 가정하자.[9] 소셜 북마크는 흥미 있는 웹 사이트를 공유하는 서비스로, 공유 아이템을 정리하기 쉽도록 카테고리를 자동으로 부여하는 기능을 제공한다.

이 문제는 고정된 카테고리를 예측하는 문제이므로 분류 문제에 해당한다. 먼저 "정치", "연예", "기술", "생활" 등의 카테고리를 정의한다. 그런 다음 각 카테고리에 속하는 콘텐츠를 1,000건 정도씩 수집한다. 그리고 사람이 직접 이 콘텐츠를 분류한다. '수집'에도 여러 방법이 있는데, 기존 콘텐츠가 있는 경우라면 특정 키워드를 포함하는 콘텐츠를 정답 데이터로 삼을 수 있다. 이렇듯 어떤 기준을 정해 콘텐츠를 정답 카테고리로 분류하면 된다. 이상의 과정으로 첫 번째 개발 데이터를 작성한다.

여기서 '1,000건 정도'는 어디까지나 대강의 기준이다. 데이터가 더 적어도 풀 수 있는 문제도 있지만, 일단 카테고리당 이 정도 데이터를 갖춘다면 첫 단계로서는 무난한 수준이 된다.

---

7 카미시마 토시히(神嶌敏弘), 「전이 학습(転移学習)」, 인공지능학회지(人工知能学会誌) 25.4, 2010, p572-580, http://www.kamishima.net/archive/2010-s-jsai_tl.pdf

8 https://hackernoon.com/ef03260747f3

9 http://bit.ly/2EFjlcC

머신러닝으로 풀 수 있는 문제는 대개 사람이 직접 하면 쉽게 풀 수 있는 수준의 문제다. 그러니 훈련 데이터를 작성할 때는 사람이 직접 푼다면 어떤 정보를 사용할지를 주의깊게 생각해봐야 한다.

데이터 중에는 사람이 보기에도 선뜻 판단하기 어려운 데이터가 있게 마련이다. 예를 들어 뉴스 기사의 제목이 "아이돌이 장관과 선수 응원 이벤트에 갔다"라면 이 뉴스를 "연예"로 분류해야 할지 "정치"로 분류해야 할지 판단하기 어렵다. 이런 경우에는 카테고리가 배타적인 문제로 볼지, 여러 카테고리에 속할 수 있도록 할지를 고민해야 한다. 어떤 결정을 내리냐에 따라 사용하게 될 알고리즘과 예측 방법이 달라지게 된다. 데이터를 훑어보기 전에 만든 카테고리를 그대로 사용할 수 있는지, 아니면 수정을 가해야 할지 생각하면서 진행하도록 한다.

그리고 데이터를 사람이 직접 분류하다 보면 '기사 제목에 포함된 단어로 카테고리를 분류하면 효과적이겠다' 같은 통찰을 얻게 된다. 이 유용할 듯한 정보(기사 제목에 포함된 단어)를 특징에 포함시키면 성능이 개선되기도 한다.

이런 식으로 모든 데이터에 카테고리를 매길 즈음이 되면 이 분류 작업의 정의에 대해 다른 사람에게 설명할 수 있을 만큼 여러 사실을 깨닫게 될 것이다. 이렇게 찾아낸 분류 기준이나 분류가 어려운 콘텐츠 등은 따로 정리해두도록 한다. 마치 소스코드처럼, 정리해두지 않으면 한 달 후 전혀 새로운 것처럼 보일 것이다. 그러니 그때의 자신에게 설명할 수 있을 정도로 정리해두는 것이 좋다.

이 방법은 학습에 사용할 데이터셋을 만드는 첫 단계라고 보아도 좋다. 그러나 카테고리를 부여할 콘텐츠가 너무 많아지면 일의 진척이 느려지게 된다. 또, 혼자 자신만의 느낌대로 작업하게 되면 편견이 개입되기 쉬워 사용자가 느끼는 것과 어긋나는 결과물이 될 수 있으니 주의해야 한다.

다음으로는 이런 단점을 극복할 방법을 알아보겠다.

5.4 동료나 친구에게 데이터 입력을 부탁

데이터가 대량으로 필요하다면 가장 쉬운 해결책은 역시 동료나 친구에게 도움을 받는 것이다. 물론 외주로 맡기는 방법도 있지만, 그보다는 비용을 걱정하지 않아도 되는 방법이 우선일 것이다.

가장 단순한 방법은 스프레드시트에 대상 데이터를 복사한 다음 레이블을 달도록 하는 것이다. 구글 스프레드시트처럼 협업하기 좋은 도구를 사용하면 중복 작업의 위험도 적다.

물론 가능하다면 레이블 부여에 특화된 도구를 만들어 사용하는 것이 이상적이다. 그러면 중복 작업이나 작업자 간 간섭 문제를 줄일 수 있다. 이미지의 영역을 선택하는 것처럼 애초에 스프레드시트로는 불가능한 문제라면, 먼저 기존에 누군가가 만들어 놓은 어노테이션 도구를 찾든가 직접 만드는 수밖에 없다.

여러 명에게 작업을 나누어 맡기는 경우라면 사전에 데이터 내용을 말로 표현할 수 있도록 한 다음, 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 한 명확히 나타내야 고품질 데이터를 얻을 수 있으니 분류 기준을 문서화해야 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 나게 마련이다. 예를 들어 '음색만으로 감정을 판단하는 문제'처럼 사람이 직접 판단해도 어려울 때가 있다. 따라서 각각이 판단한 정답이 얼마나 일치하는지를 잘 파악해야 한다. 단순히 작업자들의 판단이 얼마나 일치하는가만 훑어봐도 문제의 난이도를 짐작할 수 있다. 사람끼리도 50% 이상 일치하지 않는 문제라면 머신러닝으로도 풀 수 없을 가능성이 매우 크다. 또, 우연히 일치할 가능성을 고려한 기준인 **카파 계수**Kappa coefficient로도 문제의 난이도를 판단할 수 있다.

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

5.5 크라우드소싱 활용

대량의 데이터를 수집할 수 있는 또 다른 방법으로 크라우드소싱crowd sourcing이 있다. 크라우드소싱이라고 하면 랜서즈[10]나 크라우드웍스[11]처럼 불특정 다수의 사람이 서로 경쟁하는 방

10 http://www.lancers.jp/
11 http://crowdworks.jp/

식을 떠올리기 쉽지만, 아마존 메커니컬 터크Amazon Mechanical Turk[12]나 야후! 크라우드소싱[13], CROWD[14] 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 방식이어서 국내외에서 많이 사용된다. 또 기업에서 크라우드소싱으로 훈련 데이터를 만드는 경우도 늘고 있다.

크라우드소싱을 활용하여 훈련 데이터를 만들 때의 장점은 다음과 같다.

- 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.

데이터양이 적고 간단한 과제라면 한두 시간 안에 처리되기도 하며, 더 나은 데이터를 얻기 위해 반복할 수 있다는 점이 매력이다.

다음은 주의할 점이다.

- 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.

작업 결과의 품질 높이려면 시행착오와 노하우가 필요하다. 같은 작업을 여러 명에게 맡겨 중복된 데이터를 얻은 뒤 다수결 정보 레이블을 따로 부여하거나, 미리 연습문제를 풀게 하거나, 설문조사를 통해 작업자를 걸러내는 방법도 사용한다. 이런 식으로 어느 정도 품질을 확보하면서 많은 양의 데이터를 얻을 수 있다.

또, 모든 데이터를 직접 확인할 수 없으므로 확보된 데이터의 품질을 평가하는 방법도 미리 생각해둬야 한다. 예를 들어 분류 학습에 이용할 데이터라면 '카테고리별 표본을 추출하여 분류 정확도를 확인'하는 표본 추출 방법이 많이 사용된다.

12 https://www.mturk.com/mturk/welcome

13 http://crowdsourcing.yahoo.co.jp/

14 http://www.realworld.jp/crowd/

5.6 서비스에 수집 기능을 넣고 사용자가 입력하게 함

훈련 데이터를 꼭 직접 만들어낼 필요는 없다. 예를 들어 정답 데이터를 서비스 사용자로부터 입력받는 방법이 있다. 넓은 의미에서는 크라우드소싱으로 볼 수도 있지만, 운영 중인 자사 서비스를 잘 이해한 사용자의 협력을 얻을 수 있다는 점이 큰 매력이다.

일반 사용자 대상 서비스라면 직접 간단한 설문조사를 하거나, 콘텐츠에 적절한 태그나 카테고리를 사용자가 부여하게 하거나, 검색 결과나 추천 결과에서 부적절한 콘텐츠를 신고하게 하는 등의 데이터 수집 기능을 서비스에 포함시킬 수 있다. 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 reCAPTCHA[15]도 좋은 예다.

이 방식으로는 새로 추가된 콘텐츠에 대해서도 지속적으로 정답 데이터를 얻을 수 있으므로 트렌드를 좇기 쉽다는 장점도 따라온다.

5.7 정리

이번 장에서는 지도 학습용 학습 데이터를 수집하는 방법을 알아보았다. 구체적으로는 공개된 데이터셋이나 모델을 사용하기, 직접 만들기, 동료나 지인에게 부탁하기, 크라우드소싱 활용하기, 서비스에 데이터 수집 기능 포함시키기 등이 있다.

품질 좋은 데이터를 충분히 확보하는 일은 머신러닝에서 중요한 핵심 중 하나다. 여러분의 프로젝트에도 적절히 활용해보기 바란다.

15 https://ja.wikipedia.org/wiki/ReCAPTCHA