3주차 모델 선택 및 평가

발표자 : 서기원 / 2018-04-11

목차

모델 선택과 평가

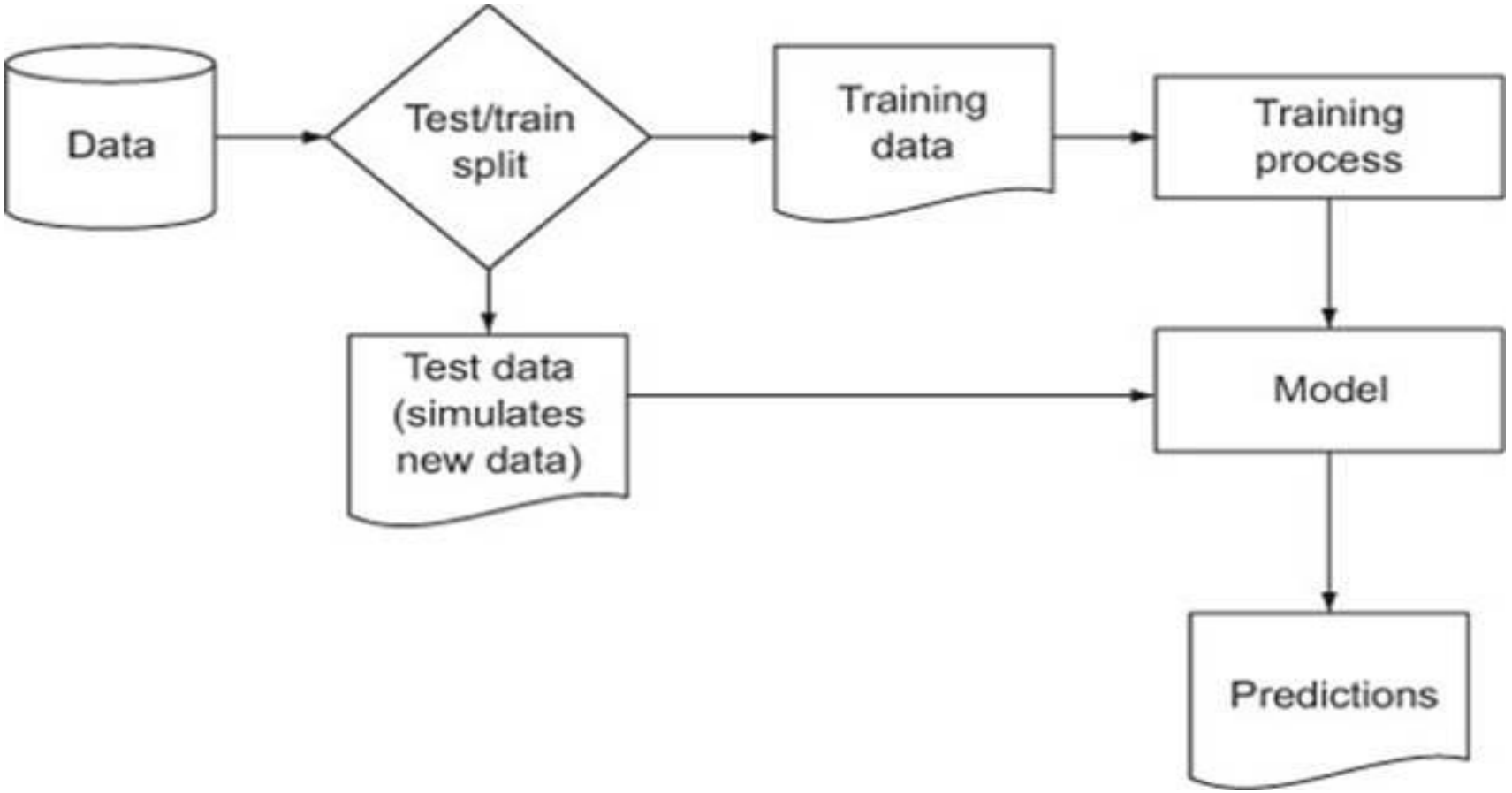
이 장에서 다룰 내용

1. 비즈니스 문제를 머신러닝 과제로 매핑하기
2. 모델 품질 평가하기
3. 모델 건전성 검증하기

모델 성능 평가와 유효성 검증

-> 5.1 과 같이 트레이닝 데이터와 테스트 데이터로 분리

모델 구축과 평가의 순서도



5.1 머신 러닝 과제에 문제 매핑하기

- 과거 구매 내역을 바탕으로 고객이 어떤 상품을 살지 예측하기
- 사기 구매인지 인식하기
- 다양한 상품 또는 상품군에 대한 가격 탄력성 예측하기(상품 가격의 인상이 상품 판매를 어떤 비율로 감소시키는지, 혹은 그 반대의 경우)
- 고객이 상품을 검색할 때 상품을 보여주는 최적의 방법 결정하기
- 고객 세분화: 같은 구매 패턴을 가진 고객 그룹핑
- 마케팅 캠페인에 대한 평가
- 새로운 상품들을 제품 카탈로그에 배열하기

-> 회귀 모델 or 의사결정나무 or 랜덤 포레스트

5.1.1 분류 문제 해결하기

1. 상품 분류화는 상품의 속성과 상품의 설명서에 기반한 분류화의 예

2. 분류화 자체는 지도학습의 한 예

3. 분류화는 대상을 어떻게 분류화 할지 학습하기 위하여
   이미 분류화 된(트레이닝셋) 데이터 셋을 필요

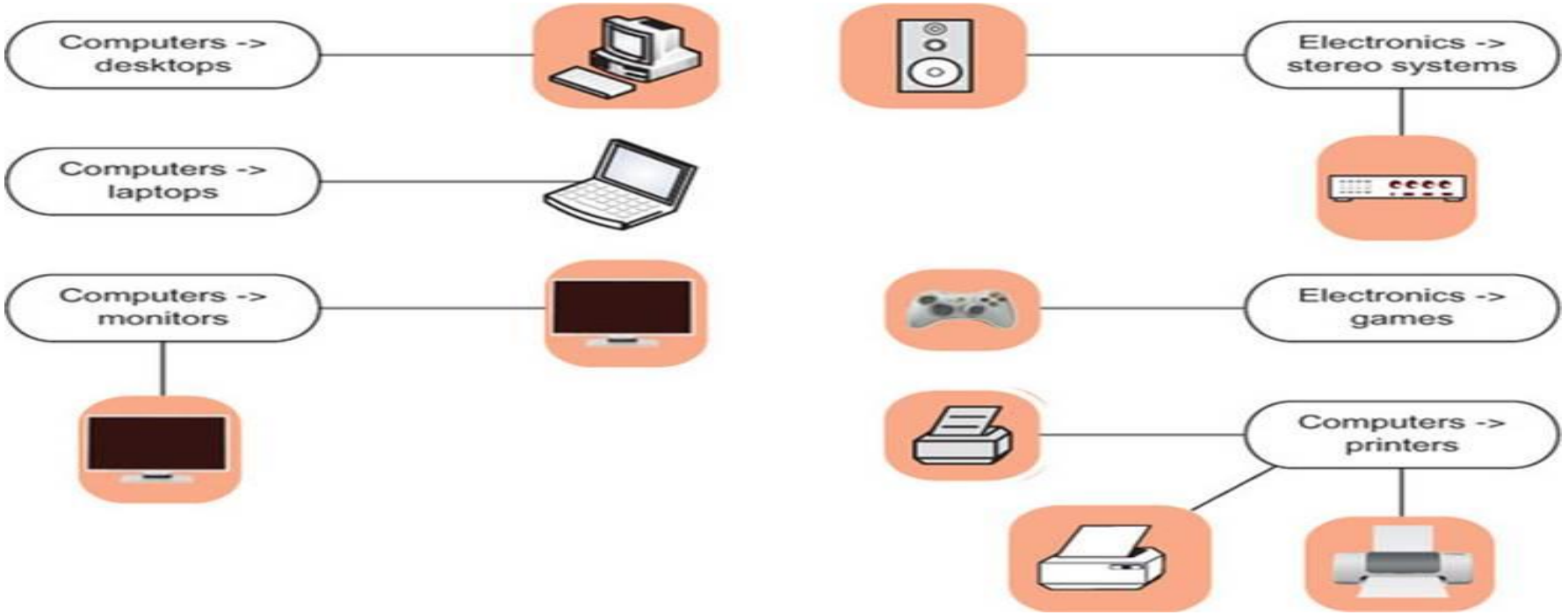


표 5.1 몇 가지 일반적인 분류화 방법

1. 나이브 베이즈
2. 의사결정 나무
3. 로지스틱 회귀
4. 서포터 벡터 머신

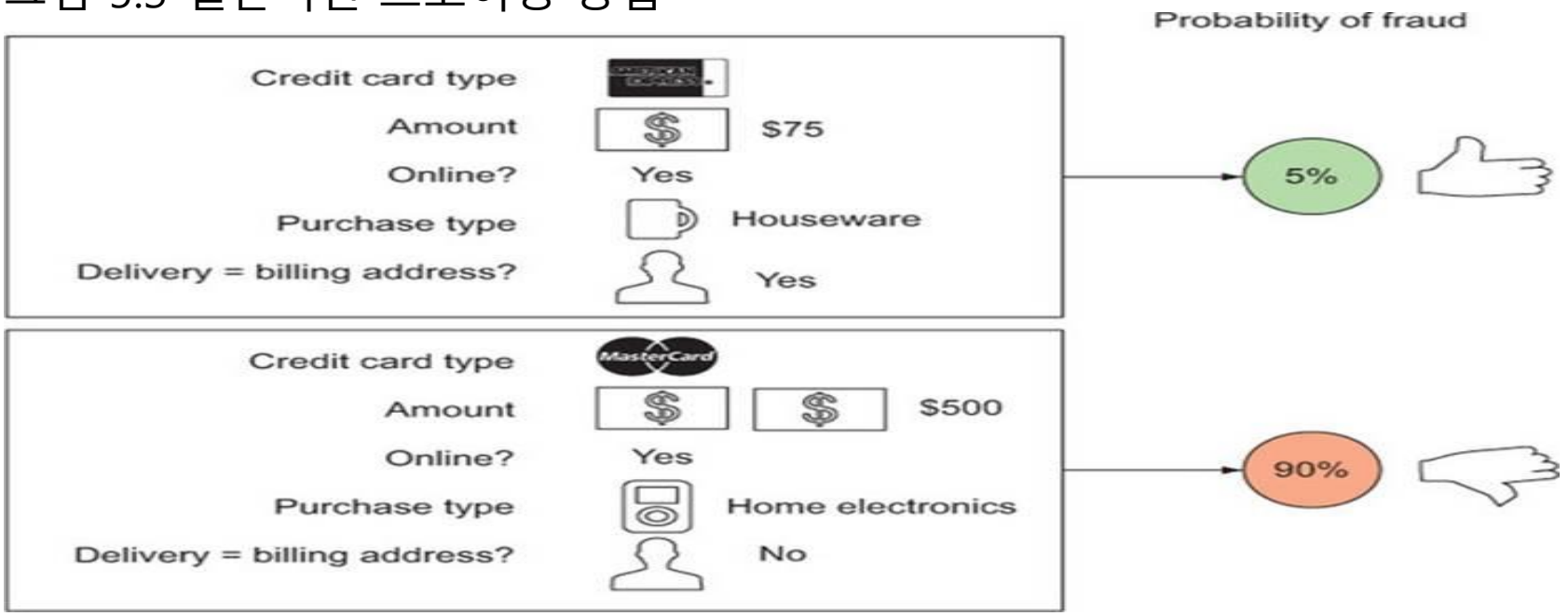
5.1.2 스코어링 문제 해결하기

스코어링

스코어링 예제에서 우리의 임무는 각기 다른 마케팅 캠페인이 웹 사이트에서 가치 있는 트래픽을 일으키는지 평가하는 것이라고 가정

스코어링은 지도 학습의 예

그림 5.3 일반적인 스코어링 방법



5.1.2 스코어링 문제 해결하기

선형 회귀

입력값의 선형 가산 함수인 숫자형 결과값을 예측하는 모델을 만듬

결과모델은 각각의 입력 변수가 결과에 상대적으로 미치는 효과의 예측을 제공

숫자 값을 예측할 때 가장 적합한 최선의 모델

로스스틱 회귀

확률이나 비율을 예측할 때 적합한 모델

항상 0과 1사이의 값을 가짐

사기 검출 문제에 적합한 방법

5.1.3 예측 결과 없이 일하기

앞서 다룬 예측 방법들은 예측 결과를 가지는 상황에 대한
트레이닝 데이터셋을 필요로 함

But -> 예측하고자 하는 특정한 결과가 없을 때도 있다 !!

Therefore ->지도학습 보다는 비지도 학습으로 불리는 방법이 적절!!

-일반적인 클러스터링 방법-

1. **k-means 클러스터링**
2. **연관 규칙을 찾기 위한 Apriori**
3. **최근접 이웃**

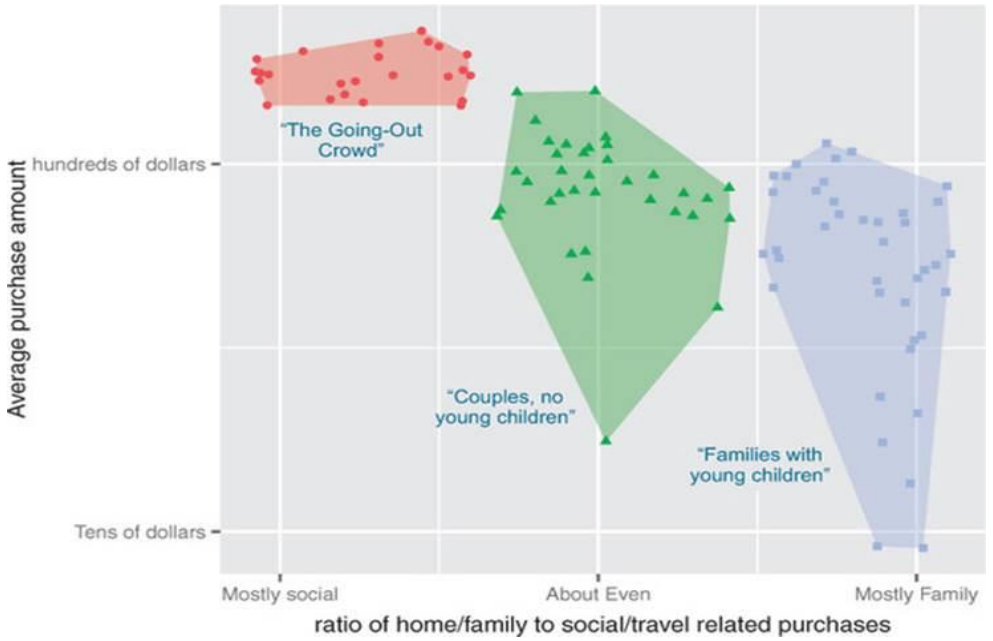
표 5.1 몇 가지 일반적인 분류화 방법

1. k-means 클러스터링

Ex) 동일한 구매패턴을 가진 고객을 일반적인 범주로 묶고 싶으나 지금은 이 그룹이 어떻게 묶이는지 모를때

2. 연관 규칙을 찾기 위한 Apriori

제품을 다른 제품과 함께 같이 구매하는 경향이 있는지에 대해서도 관심이 있을 때




Bikini, sunglasses, sunblock, flip-flops


Swim trunks, sunblock


Tankini, sunblock, sandals


Bikini, sunglasses, sunblock


One-piece, beach towel

80% of purchases include both a bathing suit and sunblock.

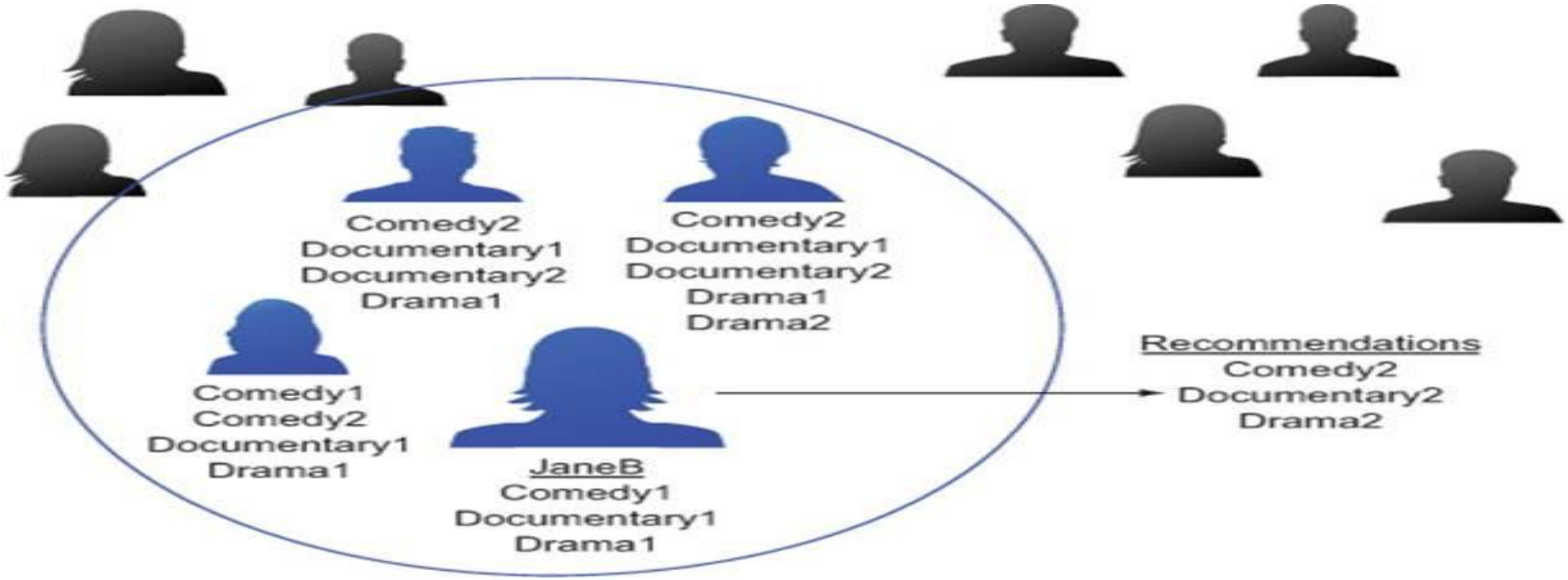
80% of purchases that include a bathing suit also include sunblock.

So customers who buy a bathing suit might also appreciate a recommendation for sunblock.

표 5.1 몇 가지 일반적인 분류화 방법

3. 최근접 이웃 메서드

상품 추천의 또 다른 방법은 고객 내에서 유사성을 찾는 것



5.1.4 문제와 방법 매핑하기

표 5.2 **문제 해결 접근법**

예제 업무	머신러닝 용어	일반적인 알고리즘
스팸 메일 식별 상품 카탈로그 내 상품 정렬 채무불이행 대출 인식하기 고객을 고객 클러스터에 할당하기	분류화: 객체에 알려진 라벨 할당하기	의사결정나무
나이브 베이즈 로지스티 회귀 서포트 벡터 머신 AdWords 값 예측하기 대출이 채무불이행될 확률 측정 마케팅 캠페인이 구매 거래 또는 매출을 얼마나 증가시킬지 예측하기	회귀: 숫자형 값 추측하거나 추론하기	선형 회귀

5.1.4 문제와 방법 매핑하기

표 5.2 문제 해결 접근법 (계속)

예제 업무	머신러닝 용어	일반적인 알고리즘
로지스틱 회귀 함께 구매하는 상품 찾기 같은 세션 내에서 방문하는 웹 페이지 식별 성공적인 웹 페이지와 애드워즈 조합 찾기	연관 규칙: 데이터 내에서 함께 노출되는 객체 경향 찾기	Apriori
같은 구매 패턴 내의 고객 그룹 식별 같은 지역 또는 고객 클러스터에서 인기 있는 상품 그룹 식별 모든 유사한 사안을 논의하고 있는 뉴스 항목 식별	클러스터링: 다른 그룹 내의 객체보다 비슷한 객체 그룹을 찾는 것	K-means
다른 비슷한 고객의 구매 내역을 바탕으로 상품 추천하기 유사 상품의 최종 가격을 바탕으로 경매 낙찰가 예측하기	최근접 이웃: 가장 비슷한 데이터를 바탕으로 데이터 속성 예측하기	최근접 이웃

5.2 모델평가

모델을 만들 때 첫 번째로 확인해야 하는 것

모델이 최소한 **트레이닝 데이터 내에서 잘 작동**하는가 ?!

모델 성능을 측정하는 계량적인 수단 (모델유형)

1. 분류화
2. 스코어링
3. 확률 예측
4. 랭킹
5. 클러스터링

5.2 모델평가

표 5.3 보정을 위한 이상적 모델

1. 널 모델

결과를 도출할 수 있는 가장 간단한 모델

하나의 상수값을 가지는 모델(모든 상황에 대해 동일한 결과를 냄) or 독립적인 모델

간단하나 항목의 전체 분포를 파악할 수 있으므로 최고의 널 모델로 동작 할 수 있음

항상 우리가 비교하고 잇는 널 모델이 모든 가능한 널 모델 중 최선의 것이라고 가정

5.2 모델평가

표 5.3 보정을 위한 이상적 모델

2. 베이즈율 모델

주어진 데이터에서 가장 최선의 모델

완벽한 모델이고 알려진 사실이 다른결과의 정확히 동일한 세트를 가진 여러 예제가 있는 경우에만 실수를 함

현실적이지 않으나 모델의 평가 스코어 상한으로 작용

5.2 모델평가

표 5.3 보정을 위한 이상적 모델

3. 단일 변수 모델

최적의 단일 변수 모델로 모델을 비교하는 것

훈련 데이터로 얻을 수 있는 최선의 단일변수 모델의 성능을 능가하지 못하면 우리가 만든 복잡한 모델의 성능을 인정할 수 없다.

5.2.1 분류 모델 평가하기

분류 모델은 2개 또는 그 이상의 범주를 갖는다

분류기의 품질을 측정하는 방법은 '정확도' 이다

분류기의 품질 성능을 측정하기 위해 '혼동행렬' 사용

R 코드 예제 5.1~

5.2.1 분류 모델 평가하기

***정밀도와 재현율**

정밀도는 확인을 위한 측정 수단

재현율은 유용성에 대한 측정 도구

***F1 (정밀도와 재현율의 조합)**

2*정밀도*재현율/(정밀도+재현율)

***민감도와 특이도**

모두 효율을 측정하기 위한 도구로 양성으로 분류된 분류 요소의 비율과 음성으로 분류된 분류 요소의 비율을 의미함

라벨을 바꾸면 민감도와 특이도를 바꾸게 된다

널 분류기는 항상 0

표5.5 분류기 성능 측정 예

일반적인 분류 성능 측정

Measure	Formula	Email spam example	Akismet spam example
Accuracy	(TP+TN)/(TP+FP+TN+FN)	0.9214	0.9987
Precision	TP/(TP+FP)	0.9187	0.9999
Recall	TP/(TP+FN)	0.8778	0.9988
Sensitivity	TP/(TP+FN)	0.8778	0.9988
Specificity	TN/(TN+FP)	0.9496	0.9965

표5.6 비즈니스 스토리를 측정하는 분류기 성능(책 참고)

표 5.6 비즈니스 스토리를 측정하는 분류기 성능

측정 도구	일반적인 비즈니스 목표	추가 질문
정확도	우리는 대부분 정확한 판단을 해야 한다.	5%의 오류를 감내할 수 있을까? 그리고 사용자들은 스팸이 아닌데 스팸으로 되어 있는 것이나, 스팸인데 스팸이 아니라고 되어 있는 것을 동등한 오류로 간주할까?
정밀도	스팸으로 인지된 메일의 대부분은 스팸이어야 한다.	즉, 스팸 폴더 안에 있는 메일의 대부분이 실제로 스팸이어야 한다. 하지만 사용자의 정상적인 이메일이 손실되는 비율을 측정하는 것은 최선의 방법이 아니다. 이 목표를 달성하기 위해 모든 사용자에게 쉽게 스팸으로 인식할 수 있는 많은 이메일을 전송함으로써 속일 수 있다. 좋은 명확한 기준이 필요할 것이다.

표 5.6 비즈니스 스토리를 측정하는 분류기 성능 (계속)

측정 도구	일반적인 비즈니스 목표	추가 질문
재현율	사용자가 보는 스팸의 양을 10배 정도 줄이고 싶다(90%의 스팸을 제거).	만일 스팸의 10%가 통과된다면 사용자는 대부분의 스팸이나 스팸 아닌 메일을 확인하는지? 이 결과가 사용자 경험 측면에서 충분한가?
민감도	우리는 많은 스팸을 제거하고 싶다. 그렇게 하지 못한다면 사용자에게 이익이 없을 것이다.	만일 우리가 1% 정도까지 스팸을 줄일 수 있다면 사용자 경험 측면에서 충분할까?
특이도	정상적인 메일의 최소한 99.9%가 사용자에게 전달되어야만 한다.	0.1%의 손실에 대해 사용자가 감내할 수 있는지? 그리고 스팸 폴더로 정상적인 메일이 가는 경우 사용자가 확인할 수 있도록 스팸 폴더의 메일을 유지해야 할까?

5.2.2 스코어링 모델 평가하기

시각적인 작업

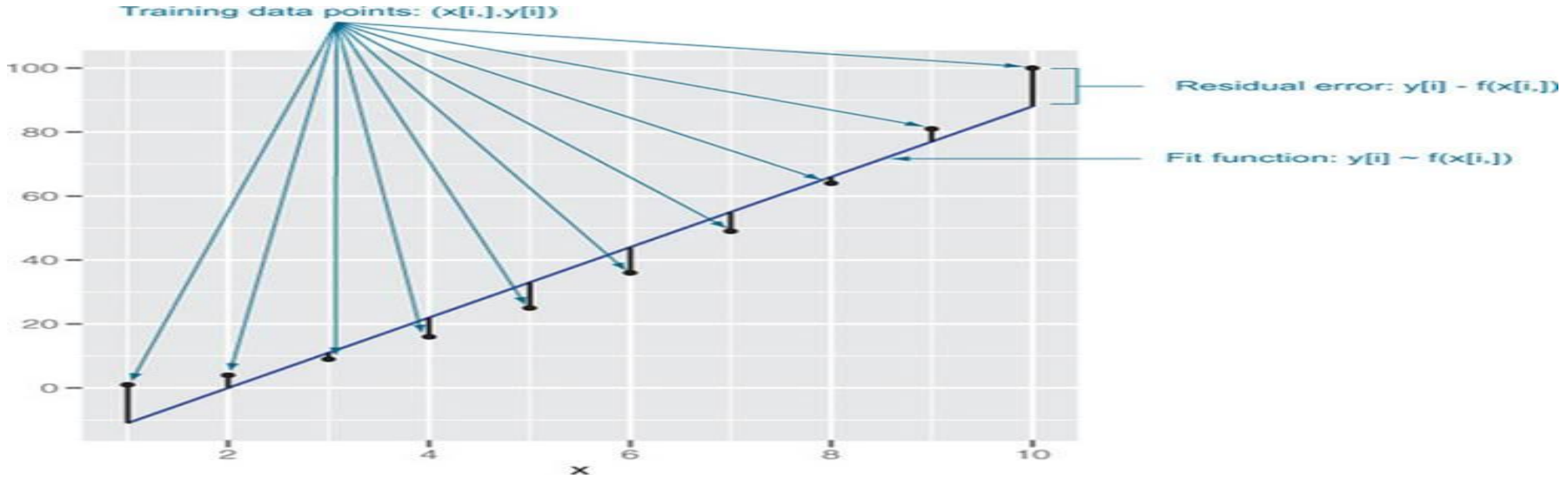
주 개념은 잔차!!

R 예제 5.5 잔차 그래프로 표현하기

5.2.2 스코어링 모델 평가하기

1. 평균 제곱근 오차
2. R 스퀘어드
3. 상관
4. 절대오차

그림5.7 잔차 스코어 표현하기



5.2.3 확률 모델 평가하기

확률 모델은 분류와 스코어링 업무에 유용

확률 모델은 요소가 주어진 분류에 속해 있는지 확인하고 분류에 속하는 확률을 얻을 수 있는 모델

이중 밀도 플롯을 만드는 것이 유용

R 예제 5.6

5.2.3 확률 모델 평가하기

1. 수신자 조작 특성 곡선
예제 5.7 수신자 조작 곡선 플롯 그리기(이중 밀도 함수 플롯의 대안)

2. 로그 우도
예제 5.8 로그 우도 계산하기
예제 5.9 널 모델의 로그 우도 계산하기

3. 편차

4. AIC

5.엔트로피
예제 5.10 엔트로피의 조건부 엔트로피 값 계산하기

5.2.5 클러스터 모델 평가하기

R 예제 5.11~14

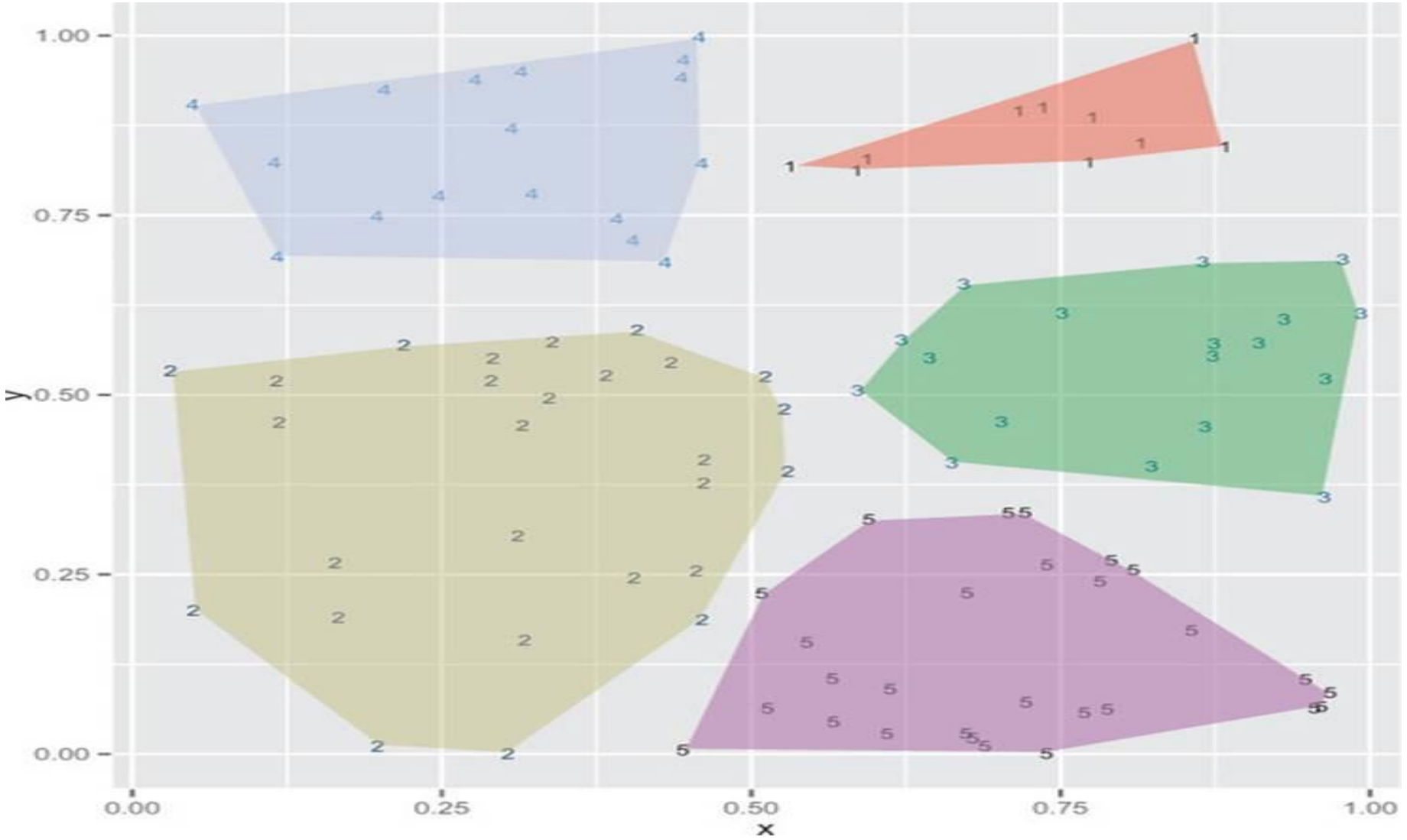
클러스터를 분류화나 스코어처럼 다루기

1. 거리 매트릭스는 클러스터 알고리즘의 성능을 점검하는데 효율적이지만 비즈니스 요구에 적합한 것은 아님

2. 프로젝트 스폰서나 고객과 클러스터에 대한 내용을 공유할 때는 클러스터에 관련한 작업을 '분류화'해라!

3. 각각의 클러스터 라벨에 클러스터에 결과변수를 할당하라 ->
분류기나 스코어링 모델 평가를 하는 방법을 클러스터의 값을 평가하는 것으로 하라

그림 5.10 클러스터 예제



5.3 모델 검증하기

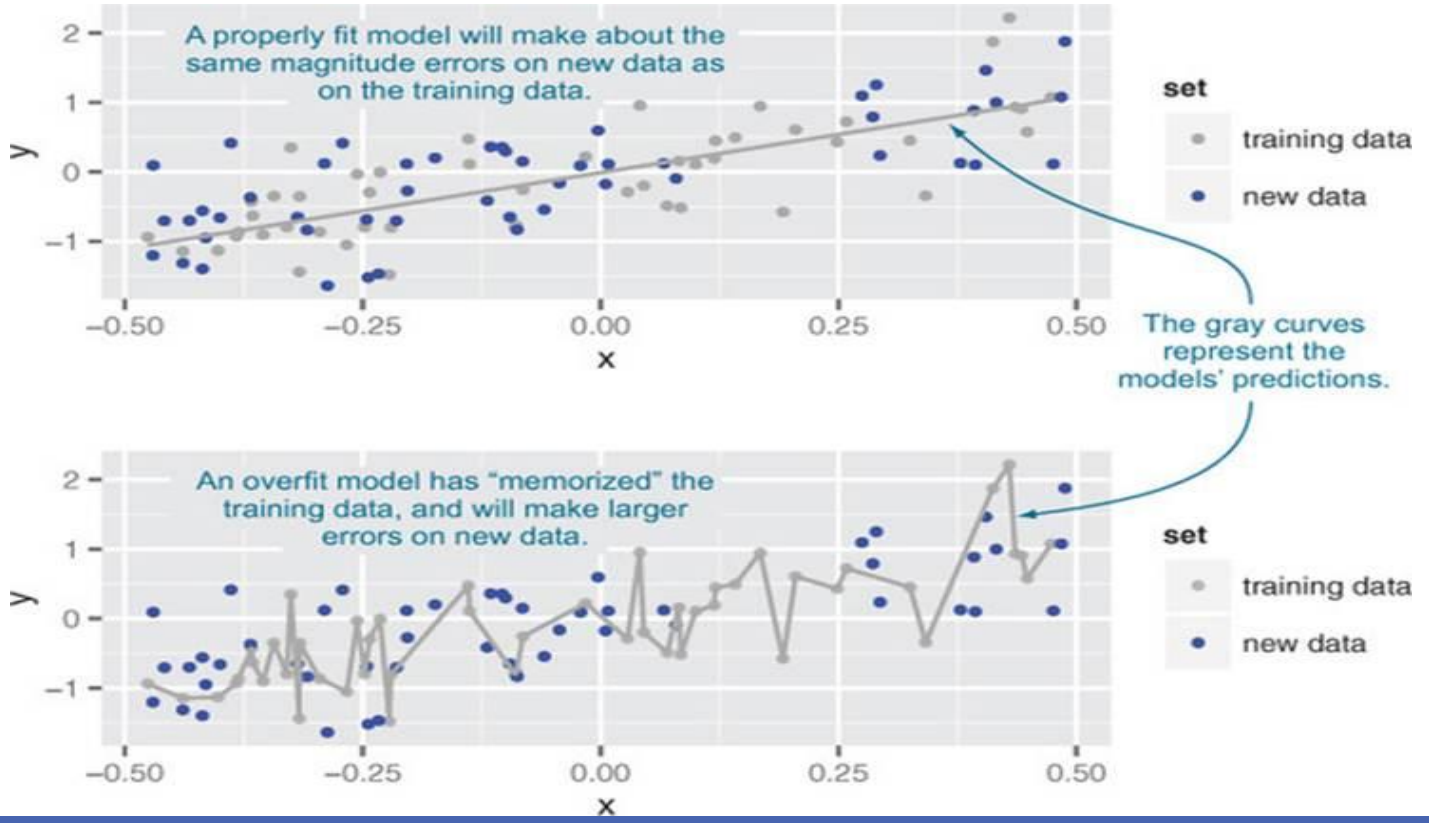
새로운 데이터 기반으로 모델을 테스트 하는 것을 '모델 검증'이라 한다.

5.3.1 일반적인 모델 문제 확인하기

표5.7 참조

과적합

- 일반화의 오류



5.3.3 모델 품질 보증

1. 홀드 아웃 데이터 기반 테스트
2. K-fold 교차 검증
3. 유의성 테스트
4. 신뢰구간
5. 통계적 용어 사용

5.4 요약

1. 항상 데이터를 탐색을 먼저하고 몇가지 측정 목표를 설계하기 전에 모델링을 시작하지마라.

2. 모델의 영향과(다양한 메트릭스 위의 성능), 건전성(과적합 같은 것을 야기하는 것이 아닌 정확한 모델 같은 것)으로 모델 테스트 항목을 나눠라.

3. 최종 테스트를 위한 모델링 작업을 위해 데이터의 일부분을 남겨둬라. 또한 트레이닝 데이터를 트레이닝과 보정을 위해 세부 분류하고 다양한 모델링 인자 값으로 최적의 값을 측정하라.

4. 마음 속에 다양한 모델 메트릭스를 두고 비즈니스 목표를 위한 최고의 모델을 골라라.

감사합니다.

Q&A