첫 인쇄: 2023년 7월 5일

우리 부모님들께:

앨리슨과 마이클 제임스 키아라 나피와 에드워드 위튼 발레리와 패트릭 헤이스티 베라와 사미 팁시라니 존 테일러와 브렌다 테일러

그리고 우리 가족들에게:

마이클, 다니엘, 캐서린 테사, 테오, 오토, 아리 사만다, 티모시, 린다 찰리, 라이언, 줄리, 셰릴 리앤과 이소벨

머리말

통계적 학습은 복잡한 내용을 이해하기 위한 도구 세트를 의미합니다. 데이터 세트. 최근 몇 년 동안 규모와 규모가 눈에 띄게 증가했습니다. 과학과 산업의 거의 모든 영역에 걸친 데이터 수집 범위. 결과적으로 통계 학습은 다음과 같은 모든 사람에게 중요한 툴킷이 되었습니다. 데이터를 이해하고 싶어하며 오늘날의 직업이 점점 더 많아지면서 이는 통계 학습이 빠르게 중요한 툴킷이 되고 있음을 의미합니다. 모두를 위해. 통계 학습에 관한 최초의 책 중 하나인 The Elements of Statistical Learning(ESL, Hastie, Tibshirani 및 Friedman 저서)이 출판되었습니다. 2001년에, 2009년에 두 번째 판이 나왔습니다. ESL은 인기 있는 교과서가 되었습니다. 통계뿐 아니라 관련 분야에서도 마찬가지다. ESL을 선택하는 이유 중 하나 인기는 상대적으로 접근하기 쉬운 스타일입니다. 그러나 ESL은 수학 과학에 대한 고급 교육을 받은 개인에게 가장 적합합니다. R(ISLR) 애플리케이션을 사용한 통계 학습 소개 — 2013년에 처음 출판되었고, 2021년에 두 번째 판이 나왔습니다. 핵심 주제를 더 광범위하고 덜 기술적으로 다루어야 한다는 분명한 필요성 통계 학습 중. 선형 회귀 검토 외에도 ISLR 오늘날 가장 중요한 통계 및 기계 학습을 다룹니다. 리샘플링, 분류 및 회귀를 위한 희소 방법, 일반화된 추가 모델, 트리 기반 방법, 지원 벡터를 포함한 접근 방식 기계, 딥 러닝, 생존 분석, 클러스터링 및 다중 테스트. 2013년에 출판된 이후 ISLR은 전 세계 학부 및 대학원 수업의 중심이 되었을 뿐만 아니라 중요한 교육 과정이 되었습니다. 데이터 과학자를 위한 참고서. 성공의 열쇠 중 하나는 2장부터 시작하여 각 장에는 다음을 설명하는 R 랩이 포함되어 있습니다. 해당 장에서 살펴본 통계적 학습 방법을 구현하는 방법, 독자에게 귀중한 실무 경험을 제공합니다. 그러나 최근 몇 년 동안 Python은 데이터 과학에서 점점 더 인기 있는 언어가 되었으며 Pythonvii에 대한 수요도 증가하고 있습니다.

viii

머리말

ISLR에 대한 기반 대안. 그래서 이 책『통계 입문』ISLP(Learning, With Application in Python)에서는 동일한 자료를 다룹니다. ISLR과 같지만 Python으로 구현된 랩을 사용합니다. 새로운 공동 저자인 Jonathan Taylor가 추가되었습니다. 여러 실험실에서 사용 수행을 용이하게 하기 위해 작성한 ISLP Python 패키지 Python의 각 장에서 다루는 통계 학습 방법. 이것들 실습은 Python 초보자는 물론 숙련된 사용자에게도 유용합니다. ISLP(및 ISLR)의 의도는 다음 사항에 더 집중하는 것입니다. 방법을 적용하고 수학적 세부 사항은 적습니다. 통계학 고급 학부생이나 석사과정 학생에게 적합 또는 관련된 양적 분야, 또는 다른 분야의 개인을 위한 통계 학습 도구를 사용하여 데이터를 분석하고 싶습니다. 그것은

사용될 수 있습니다 두 학기 동안 진행되는 과정의 교과서로 사용됩니다. 소중한 논평을 해주신 독자 여러분께 감사드립니다. ISLR 초판: Pallavi Basu, Alexandra Chouldechova, Patrick Danaher, Will Fithian, Luella Fu, Sam Gross, Max Grazier G'Sell, Courtney Paulson, xinghaoq iiao, Elisa Sheng, Noah Simon, KE Press Ming tan, ξ NL U 탠 껍질. ISLR 제2판에 대한 유용한 의견을 주신 독자들에게 감사드립니다. 앨런 아그레스티, 이아인 카마이클, 이쿤 첸, 에린 크레이그, 데이지 딩, 루시 가오, 이스마엘 렘하드리, 브라이언 마틴, 안나 뉴펠드, 제프 팀스, 카스텐 Voelkmann, Steve Yadlowsky, James Zou. 우리는 정말 감사합니다. ISLR에 대한 도움을 준 Balasubramanian "Naras" Narasimhan에게 감사드립니다. 그리고 ISLP. 상당한 영향을 목격한 것은 우리에게 영광이자 특권이었습니다. ISLR이 통계적 학습을 실행하는 방식에 있어, 두 가지 모두 학업 환경 안팎에서. 우리는 이 새로운 Python 버전이 오늘과 내일의 응용통계학자 및 데이터를 계속 제공하겠습니다. 과학자들은 데이터 중심 세계에서 성공하는 데 필요한 도구를 제공합니다.

특히 미래에 대해 예측하는 것은 어렵습니다. -요기 베라

내용물

1 소개

통계 학습 개요 통계적 학습은 데이터를 이해하기 위한 광범위한 도구 세트를 의미합니다. 이것들 도구는 감독됨 또는 감독되지 않음으로 분류될 수 있습니다. 크게 말하면, 지도 통계 학습에는 하나 이상의 입력을 기반으로 출력을 예측하거나 추정하기 위한 통계 모델을 구축하는 작업이 포함됩니다. 문제 이러한 성격은 비즈니스, 의학, 천체 물리학 및 기타 다양한 분야에서 발생합니다. 공공 정책. 비지도 통계 학습에는 입력이 있지만 감독 출력 없음; 그럼에도 불구하고 우리는 그러한 데이터로부터 관계와 구조를 배울 수 있습니다. 일부 응용 프로그램에 대한 설명을 제공합니다. 통계 학습을 통해 우리는 세 가지 실제 데이터 세트에 대해 간략하게 논의합니다. 이 책에서 고려했다.

임금 데이터 이 애플리케이션(이 전체에서 임금 데이터 세트라고 함)에서 책), 우리는 그룹의 임금과 관련된 여러 가지 요소를 조사합니다. 미국 대서양 지역 출신의 남성. 특히 우리는 바란다 직원의 나이와 교육 사이의 연관성을 이해하기 위해 그의 임금에 대한 달력 연도도 마찬가지입니다. 예를 들어 왼손을 생각해보자. 그림 1.1의 패널은 데이터 세트에 있는 각 개인의 임금 대 연령을 표시합니다. 나이가 들수록 임금이 증가한다는 증거가 있지만, 대략 60세 이후에 다시 감소합니다. 파란색 선은 특정 연령의 평균 임금 추정치를 보면 이러한 추세가 더 명확해집니다. 직원의 나이가 주어지면 이 곡선을 사용하여 그의 임금을 예측할 수 있습니다. 하지만, 또한 그림 1.1에서 이 평균값과 관련된 상당한 양의 변동성이 있음이 분명하므로 연령만으로는 영향을 미치지 않을 가능성이 높습니다. 특정 남성의 임금을 정확하게 예측합니다.

20

40

60

80

나이

300 200 50 100

값

200 50 100

값

200 50 100

값

300

1. 소개

300

2

2003년

2006년 년도

2009년

1

2

3

4

5

교육 수준

그림 1.1. 남성에 대한 소득 조사 정보가 포함된 임금 데이터 미국 중부 대서양 지역 출신. 왼쪽: 임금에 따른 임금 나이. 평균적으로 임금은 약 60세까지 연령에 따라 증가합니다. 시점부터 하락하기 시작합니다. 중앙: 연도별 임금. 느린 것이 있다 그러나 2003년 사이 평균 임금은 약 $10,000씩 꾸준히 증가했습니다. 및 2009. 오른쪽: 교육의 함수로서 임금을 표시하는 상자 그림(1) 가장 낮은 수준(고등학교 졸업장 없음)을 나타내고 5는 가장 높은 수준(고급 대학원 학위). 평균적으로 임금은 교육 수준에 따라 증가합니다.

우리는 또한 각 직원의 교육 수준에 관한 정보와 임금을 받은 연도. 중앙 및 오른쪽 패널 임금을 연도와 교육의 함수로 표시한 그림 1.1의 이 두 요소가 모두

임금과 연관되어 있음을 나타냅니다. 임금 인상 대략 선형(또는 직선) 방식으로 약 $10,000만큼 2003년과 2009년 사이에 이러한 증가는 데이터의 변동성에 비해 매우 미미합니다. 또한 임금은 일반적으로 다음과 같은 개인의 경우 더 높습니다. 고등 교육 수준: 교육 수준이 가장 낮은 남성(1)은 교육 수준이 가장 높은 사람보다 임금이 훨씬 낮습니다. (5). 분명히 특정 남성의 임금에 대한 가장 정확한 예측은 다음과 같습니다. 그의 나이, 학력, 연도를 합산하여 얻은 것입니다. 3장에서는 이를 통해 임금을 예측하는 데 사용할 수 있는 선형 회귀에 대해 논의합니다. 데이터 세트. 이상적으로는 임금을 예측하는 방식으로 임금을 예측해야 합니다. 임금과 연령 사이의 비선형 관계. 7장에서는 이 문제를 해결하기 위한 접근 방식 클래스.

주식 시장 데이터 임금 데이터에는 연속적 또는 정량적 출력 값을 예측하는 작업이 포함됩니다. 이를 흔히 회귀 문제라고 합니다. 그러나 어떤 경우에는 그 대신 숫자가 아닌 값, 즉 범주형 값을 예측하고 싶을 수도 있습니다. 또는 질적인 결과물. 예를 들어, 4장에서는 주식 시장을 살펴봅니다. Standard & Poor's 500의 일일 움직임을 포함하는 데이터 세트 (S&P) 주가 지수는 2001년부터 2005년까지 5년 동안 측정되었습니다. 이를 Smarket 데이터로 사용합니다. 목표는 지수가 지난 5일간의 백분율을 사용하여 특정 날짜의 증가 또는 감소 인덱스의 변화. 여기서는 통계적 학습 문제가 포함되지 않습니다. 수치를 예측합니다. 대신에 주어진 여부를 예측하는 것이 포함됩니다.

1. 소개

위로

0

2

4

6 위로

-2

S&P의 백분율 변화 아래에

오늘의 방향

-4

4 2 0 -2

S&P의 백분율 변화 아래에

오늘의 방향

-4

4 2 0 -2 -4

S&P의 백분율 변화

3일 전

6

이틀 전

6

어제

3

아래에

위로

오늘의 방향

그림 1.2. 왼쪽: 전날 S&P 지수 변동률을 나타내는 상자 그림 시장이 상승하거나 하락한 일수에 대한 지수로, 스마트마켓 데이터. 중앙 및 오른쪽: 왼쪽 패널과 동일하지만 백분율이 변경됩니다. 2일전과 3일전의 내용이 표시됩니다.

그날의 주식 시장 성과는 상승 버킷 또는 하락 버킷에 속합니다. 버킷. 이를 분류 문제라고 합니다. 할 수 있는 모델 시장이 움직일 방향을 정확하게 예측하는 것은 매우 유용합니다! 그림 1.2의 왼쪽 패널에는 이전 그림의 두 개의 상자 그림이 표시됩니다. 일별 주가지수 변동률: 648일 동안의 변동률 시장은 다음날 상승했고, 602일 동안 한 번 상승했습니다. 시장이 줄어든 것. 두 플롯은 거의 동일해 보이며, 이는 어제의 움직임을 사용하는 간단한 전략이 없음을 시사합니다. S&P가 오늘의 수익률을 예측합니다. 오늘부터 2일 및 3일 전의 백분율 변화에 대한 상자 그림을 표시하는 나머지 패널도 마찬가지로 과거 수익률과 현재 수익률 사이의 연관성이 거의 없음을 나타냅니다. 물론, 이것은 패턴 부족이 예상됩니다. 연속적인 일일 수익률 간에 강한 상관관계가 있는 경우 간단한 거래 전략을 채택할 수 있습니다. 시장에서 이익을 창출하기 위해. 그럼에도 불구하고 4장에서는 다양한 통계 학습 방법을 사용하여 이러한 데이터를 수집합니다. 재미있게, 데이터에는 최소한 다음과 같은 약한 경향이 있다는 힌트가 있습니다. 5년 동안의 방향을 정확하게 예측하는 것이 가능합니다. 시장의 움직임은 약 60% 정도입니다(그림 1.3).

유전자 발현 데이터 이전 두 애플리케이션은 입력과 데이터가 모두 포함된 데이터 세트를 보여줍니다. 출력 변수. 그러나 또 다른 중요한 문제 종류는 다음과 같습니다. 입력 변수만 관찰하고 해당 변수가 없는 상황 산출. 예를 들어 마케팅 환경에서는 인구통계학적 정보가 있을 수 있습니다. 다수의 현재 또는 잠재 고객에 대한 정보. 우리는 원할 수도 있습니다 그룹화를 통해 어떤 유형의 고객이 서로 유사한지 파악 관찰된 특성에 따라 개인을 분류합니다. 이것은 다음과 같이 알려져 있습니다.

1. 소개

0.50 0.48 0.46

예측 확률

0.52

4

아래에

위로

오늘의 방향

그림 1.3. 우리는 2차 판별 분석 모델을 하위 집합에 적합합니다. 2001~2004년 기간에 해당하는 Smarket 데이터를 분석하고 예측합니다. 2005년 데이터를 이용하여 주식시장이 하락할 확률. 평균적으로, 예측된 감소 확률은 시장이 하락하는 날에 더 높습니다. 감소하다. 이러한 결과를 바탕으로 우리는 방향을 정확하게 예측할 수 있습니다. 60%의 시간 동안 시장에서 움직임.

클러스터링 문제. 이전 예와 달리 여기서는 시도하지 않습니다. 출력 변수를 예측합니다. 우리는 12장에서 통계적 학습 방법에 대해 논의합니다. 자연적인 출력 변수를 사용할 수 없는 문제의 경우. 우리는 고려한다 6,830개의 유전자 발현 측정으로 구성된 NCI60 데이터 세트 64개의 암세포주 각각에 대해. 특정 결과를 예측하는 대신 변수가 있는 경우 그룹이 있는지 여부를 확인하는 데 관심이 있습니다. 유전자 발현 측정을 기반으로 한 세포주 중 클러스터. 부분적으로는 수천 개가 있기 때문에 이것은 해결하기 어려운 질문입니다. 세포주당 유전자 발현 측정값이 많아 시각화가 어렵습니다. 데이터. 그림 1.4의 왼쪽 패널에서는 Z1과 Z2라는 두 숫자만 사용하여 64개의 세포주 각각을 나타냄으로써 이 문제를 해결합니다. 이것들 데이터의 처음 두 가지 주요 구성 요소는 다음을 요약합니다. 두 개의 숫자까지 각 세포주에 대한 6,830개의 발현 측정값 또는 치수. 이러한 차원 감소로 인해 다음과 같은 결과가 발생할 가능성이 높습니다. 일부 정보가 손실되었으므로 이제 데이터를 시각적으로 검사할 수 있습니다. 클러스터링의 증거를 위해. 클러스터 수를 결정하는 것은 종종 어려운 문제. 그러나 그림 1.4의 왼쪽 패널은 최소한 별도의 색상을 사용하여 표현한 네 가지 세포주 그룹. 이 특정 데이터 세트에서는 세포주가 일치하는 것으로 나타났습니다. 14가지 종류의 암. (단, 이 정보는 사용되지 않았습니다. 그림 1.4의 왼쪽 패널을 생성합니다.) 그림 1.4의 오른쪽 패널은 14가지 암 유형을 제외하고 왼쪽 패널과 동일합니다. 고유한 색상의 기호를 사용하여 표시됩니다. 세포가 있다는 분명한 증거가 있습니다. 동일한 암 유형을 가진 계통은 서로 가까이 위치하는 경향이 있습니다. 2차원 표현. 또한 왼쪽 패널을 제작하는데 암 정보를 사용하지 않았음에도 불구하고 클러스터링이 얻어졌다. 관찰된 실제 암 유형 중 일부와 일부 유사합니다. 오른쪽 패널에서. 이는 다음 사항에 대한 독립적인 검증을 제공합니다. 클러스터링 분석의 정확성.

Z2

-20

0

20

5

-60

-40

-20 -60

-40

Z2

0

20

1. 소개

-40

-20

0

20

40

60

-40

-20

Z1

0

20

40

60

Z1

그림 1.4. 왼쪽: NCI60 유전자 발현 데이터 세트의 표현 2차원 공간 Z1 및 Z2. 각 포인트는 64개 중 하나에 해당합니다. 세포주. 우리가 대표하는 세포주에는 네 가지 그룹이 있는 것으로 보입니다. 다른 색상을 사용합니다. 오른쪽: 우리가 표현한 것을 제외하면 왼쪽 패널과 동일 14가지 종류의 암은 각각 다른 색상의 기호를 사용합니다. 세포주 동일한 유형의 암에 해당하는 것은 2차원적으로 근처에 있는 경향이 있습니다. 공간.

통계 학습의 간략한 역사 통계 학습이라는 용어는 상당히 새로운 개념이지만, 통계 학습에 사용되는 많은 개념은 이 분야의 기초는 오래 전에 개발되었습니다. 19세기 초 최소제곱법이 개발되어 현재 선형 회귀라고 알려진 것의 초기 형태입니다. 접근 방식 천문학 문제에 처음으로 성공적으로 적용되었습니다. 선형 회귀 개인의 급여와 같은 정량적 가치를 예측하는 데 사용됩니다. 환자의 생존 여부 등 정성적 가치를 예측하기 위해 죽느냐, 죽느냐, 주식시장이 오르든 내리든 선형판별분석은 1936년에 제안되었다. 1940년대에는 다양한 저자들이 다음과 같이 말했다. 대안적인 접근법인 로지스틱 회귀를 제시합니다. 1970년대 초, 일반화 선형 모델이라는 용어는 전체 클래스를 설명하기 위해 개발되었습니다. 선형 및 로지스틱 회귀를 모두 포함하는 통계 학습 방법 특별한 경우로. 1970년대 말에는 데이터 학습을 위한 더 많은 기술이 등장했습니다. 이용 가능했습니다. 그러나 비선형 관계를 맞추는 것이 계산적으로 어려웠기 때문에 거의 선형 방법이었습니다. 시간. 1980년대에 이르러 컴퓨팅 기술은 마침내 충분히 향상되었습니다. 비선형 방법은 더 이상 계산적으로 불가능하지 않습니다. ~ 안에 1980년대 중반에는 분류 및

회귀 트리가 개발되었으며, 이후 곧 일반화된 덧셈 모델을 사용합니다. 신경망이 인기를 얻었습니다. 1980년대에는 지원 벡터 머신이 등장했고 1990년대에는 지원 벡터 머신이 등장했습니다. 그 이후로 통계 학습은 과학의 새로운 하위 분야로 등장했습니다. 감독 및 비지도 모델링 및 예측에 중점을 둔 통계입니다. 최근 몇 년간 통계 학습의 진전은 다음과 같이 두드러졌습니다. 강력하고 상대적으로 사용자 친화적인 소프트웨어의 가용성이 증가하고 있습니다. 대중적이고 무료로 사용 가능한 Python 시스템입니다. 이는 다음과 같은 가능성이 있습니다. 사용된 일련의 기술을 통해 해당 분야의 변화를 계속하고

통계학자와 컴퓨터 과학자가 필수 툴킷으로 개발했습니다. 훨씬 더 넓은 커뮤니티를 위해.

이 책 Hastie, Tibshirani 및 ESL(통계 학습 요소) Friedman은 2001년에 처음 출판되었습니다. 그 이후로 이 책은 통계적 기계 학습의 기초에 대한 중요한 참고 자료입니다. 그 성공은 많은 문제를 포괄적이고 자세하게 처리한 데서 비롯됩니다. 통계 학습에서 중요한 주제뿐만 아니라 (많은 상위 통계 교과서) 이 책은 광범위한 청중이 접근할 수 있습니다. 그러나 ESL 성공의 가장 큰 요인은 화제성이었습니다. 자연. 출판 당시 통계 분야에 대한 관심이 높았습니다. 학습이 폭발하기 시작했습니다. ESL은 최초의 접근 가능한 솔루션 중 하나를 제공했습니다. 주제에 대한 포괄적인 소개. ESL이 처음 출판된 이후 통계 학습 분야는 계속해서 발전해 왔습니다. 이 분야의 확장은 두 가지 형태를 취했습니다. 가장 명백한 성장에는 다양한 과학적 질문에 답하기 위한 새롭고 향상된 통계 학습 접근법의 개발이 포함되었습니다. 여러 분야에 걸쳐. 그러나 통계학습 분야는 또한 청중을 확대했습니다. 1990년대에는 컴퓨팅 능력이 향상되었습니다. 비통계학자로부터 이 분야에 대한 관심이 급증했습니다. 최첨단 통계 도구를 사용하여 데이터를 분석하고 싶어합니다. 불행하게도 이러한 접근 방식의 고도의 기술적 특성으로 인해 사용자는 커뮤니티는 주로 통계, 컴퓨터 분야의 전문가로만 제한되었습니다. 과학 및 관련 분야를 이해하고 훈련할 수 있는 훈련(및 시간)이 필요합니다. 그것들을 구현하십시오. 최근 몇 년 동안 새롭고 향상된 소프트웨어 패키지가 크게 증가했습니다. 많은 통계 학습 방법에 대한 구현 부담을 완화했습니다. 동시에 여러 곳에서 인지도가 높아지고 있습니다. 비즈니스에서 의료, 유전학, 사회 과학에 이르기까지 다양한 분야 그 이상으로, 통계적 학습은 중요한 실용적 기능을 갖춘 강력한 도구입니다. 응용 프로그램. 그 결과, 해당 분야는 주로 학술적인 분야에서 벗어나게 되었습니다. 엄청난 잠재 청중이 있는 주류 학문에 대한 관심. 이러한 추세는 엄청난 양의 가용성이 증가함에 따라 확실히 계속될 것입니다. 엄청난 양의 데이터와 이를 분석하는 소프트웨어. 통계 학습 입문(ISL)의 목적은 통계 학습을 학문에서 주류로 전환하는 것을 촉진하는 것입니다. 필드. ISL은 ESL을 대체하기 위한 것이 아닙니다. ESL은 고려된 접근 방식의 수와 그들이 탐구되는 깊이. 우리는 ESL을 중요하게 생각합니다 전문가를 위한 동반자(통계, 기계 분야 대학원 학위 보유) 학습 또는 관련 분야) 기술적인 세부 사항을 이해해야 하는 사람 통계적 학습 접근법 뒤에 숨어 있습니다. 그러나 사용자 커뮤니티에서는 통계 학습 기술은 다음과 같은 개인을 포함하도록 확장되었습니다. 더 넓은 범위의 관심과 배경. 그러므로 다음을 위한 장소가 있다. 덜 기술적이고 접근하기 쉬운 ESL 버전입니다.

1. 소개

7

수년에 걸쳐 이러한 주제를 가르치면서 우리는 이러한 주제가 다음과 같다는 사실을 발견했습니다. 비즈니스와 같이 서로 다른 분야의 석사 및 박사 과정 학생들의 관심 분야 행정, 생물학, 컴퓨터 과학뿐만 아니라 양적 지향의 상위 학부 학부생에게도 제공됩니다. 이 다양한 것이 중요합니다 그룹이 모델, 직관, 강점을 이해하고 다양한 접근법의 약점. 하지만 이 청중에게는 많은 최적화 알고리즘 및 이론적 속성과 같은 통계적 학습 방법 뒤에 있는 기술적 세부 사항은 주요 관심사가 아닙니다. 우리는 이 학생들이 이러한 것들에 대한 깊은 이해가 필요하지 않다고 믿습니다. 다양한 방법론에 대해 잘 알고 있는 사용자가 되기 위한 측면 통계를 사용하여 선택한 분야에 기여하기 위해 학습 도구. ISL은 다음 네 가지 전제를 기반으로 합니다. 1. 많은 통계적 학습 방법은 광범위한 분야에서 관련성이 있고 유용합니다. 통계학을 넘어 다양한 학문적, 비학술적 학문 분야를 포괄합니다. 우리는 많은 현대 통계 학습 절차가 다음과 같이 널리 이용 가능하고 사용되어야 하며 앞으로도 그렇게 될 것이라고 믿습니다. 현재 선형 회귀와 같은 고전적인 방법의 경우와 같습니다. 결과적으로 가능한 모든 것을 고려하려고 하기보다는 (불가능한 작업) 접근 방식을 제시하는 데 집중했습니다. 우리가 믿는 방법은 가장 널리 적용 가능하다고 생각합니다. 2. 통계적 학습을 일련의 블랙박스로 보아서는 안 됩니다. 아니요 단일 접근 방식은 가능한 모든 애플리케이션에서 잘 작동합니다. 상자 안의 모든 톱니바퀴나 상호작용을 이해하지 못한 채 그 톱니바퀴 사이에서 가장 좋은 상자를 선택하는 것은 불가능합니다. 따라서 우리는 우리가 고려하는 각 방법 뒤에 있는 모델, 직관, 가정 및 장단점을 주의 깊게 설명하려고 시도했습니다. 3. 각 톱니바퀴가 어떤 작업을 수행하는지 아는 것도 중요하지만, 내부에 기계를 구성하는 기술이 필요하지 않습니다. 상자! 이에 관련 기술적 사항에 대한 논의를 최소화하였습니다. 피팅 절차와 이론적 특성에 대해 설명합니다. 우리는 독자는 기본적인 수학적 개념에 익숙하지만 우리는 알고 있습니다. 수학 과학 분야의 대학원 학위를 취득하지 마십시오. 예를 들어, 우리는 행렬 대수학의 사용을 거의 완전히 피했습니다. 자세한 설명 없이도 책 전체를 이해할 수 있습니다. 행렬과 벡터에 대한 지식. 4. 우리는 독자가 통계 학습 방법을 실제 문제에 적용하는 데 관심이 있다고 가정합니다. 또한 이를 원활하게 하기 위해 논의된 기술에 동기를 부여하기 위해 한 섹션을 할당했습니다. 각 장 내에서 컴퓨터실까지. 각 연구실에서 우리는 독자를 안내합니다. 해당 장에서 고려한 방법을 현실적으로 적용함으로써. 우리 코스에서 이 자료를 가르쳤을 때 우리는 수업 시간의 약 1/3을 다음 작업에 할당했습니다. 우리는 이것이 매우 유용하다는 것을 알았습니다. 많은 처음에는 겁을 먹은 덜 계산 지향적인 학생 실험실에서는 분기 동안 상황을 파악했거나 학기. 이 책은 원래 출간되었습니다(2013년, 2판 2021).

8

1. 소개

R 언어로 작성된 컴퓨터 실습실이 있습니다. 그 이후로 통계 학습에서 중요한 기술을 Python으로 구현하려는 수요가 증가하고 있습니다. 결과적으로 이 버전에는 Python 연구실. 사용할 수 있는 Python 패키지의 수가 빠르게 증가하고 있으며, 시작 부분에서 가져오기를 검토한 결과 각 연구실에서 독자들은 우리가 신중하게 선택하고 사용했다는 것을 알게 될 것입니다. 가장 적절합니다. 우리는

또한 몇 가지 추가 코드를 제공했으며 우리 패키지 ISLP의 기능. 그러나 ISL의 실습은 독립적이므로 독자가 다른 실습을 사용하려는 경우 건너뛸 수 있습니다. 소프트웨어 패키지 또는 논의된 방법을 적용하고 싶지 않습니다. 현실 세계의 문제.

이 책은 누가 읽어야 하는가? 이 책은 데이터 모델링 및 예측을 위해 최신 통계 방법을 사용하는 데 관심이 있는 모든 사람을 대상으로 합니다. 이 그룹에는 다음이 포함됩니다. 과학자, 엔지니어, 데이터 분석가, 데이터 과학자, 퀀트뿐만 아니라 그 이하도 포함됩니다. 사회 과학이나 비즈니스와 같은 비계량적 분야의 학위를 보유한 기술 개인입니다. 우리는 독자가 최소한 통계학의 한 초등학교 과정. 선형 회귀의 배경도 다음과 같습니다. 필수는 아니지만 유용합니다. 선형의 핵심 개념을 검토하기 때문입니다. 이 책의 수학적 수준은 보통 수준이지만, 행렬 연산에 대한 자세한 지식이 필요하지 않습니다. 이 책 Python에 대한 소개를 제공합니다. 프로그래밍에 대한 이전 노출 MATLAB이나 R과 같은 언어가 유용하지만 필수는 아닙니다. 이 교과서의 초판은 석사과정과 석사과정을 가르치는 데 사용되었습니다. 경영학, 경제학, 컴퓨터 과학, 생물학, 지구 과학, 심리학, 기타 물리 및 사회 과학 분야의 박사 과정 학생입니다. 또한 이미 학부생을 가르치는 데에도 사용되었습니다. 선형회귀에 대한 강의를 들었습니다. ESL이 기본 교과서로 사용되는 보다 수학적으로 엄격한 과정의 맥락에서 ISL 계산적인 측면을 가르치기 위한 보충 교재로 사용될 수 있습니다. 다양한 접근 방식 중.

표기법과 단순 행렬 대수학 교과서 표기법을 선택하는 것은 항상 어려운 작업입니다. 대부분의 경우 부분에서는 ESL과 동일한 표기 규칙을 채택합니다. n을 사용하여 샘플의 고유한 데이터 포인트 또는 관측치 수를 나타냅니다. p를 변수의 수로 지정하겠습니다. 예측에 사용할 수 있습니다. 예를 들어, 임금 데이터 세트는 3,000명에 대한 11개의 변수로 구성되어 있으므로 n = 3,000개의 관측값이 있고 p = 11개 변수(예: 연도, 연령, 인종 등) 전체적으로 참고하세요 이 책에서는 변수 이름을 컬러 글꼴로 표시합니다: 변수 이름. 일부 예에서 p는 수천 또는 심지어 수백만과 같이 상당히 클 수 있습니다. 이러한 상황은 예를 들어 다음과 같은 경우에 자주 발생합니다. 현대 생물학적 데이터 또는 웹 기반 광고 데이터 분석.

일반적으로 xij는 j번째 변수의 값을 나타냅니다. i번째 관측치, 여기서 i = 1, 2, . . . , n 및 j = 1, 2, . . . , p. 이 전반에 걸쳐 책에서 i는 샘플 또는 관찰(1부터 n까지)을 색인화하는 데 사용됩니다. j는 변수(1부터 p까지)를 인덱싱하는 데 사용됩니다. X가 (i, j)번째 요소가 xij 인 n × p 행렬입니다. 즉, ⊠ ⊠ x11 x12 . . . x1p ⊠ x21 x22 . . . x2p ⊠ ⊠ ⊠ X=⊠ . .. .. ⊠ . .. ⊠ .. . . . ⊠ xn1 xn2 . . . xnp

행렬에 익숙하지 않은 독자에게는 X를 다음과 같이 시각화하는 것이 유용합니다. n행과 p열로 구성된 숫자 스프레드시트입니다. 때때로 우리는 X의 행에 관심을 갖게 될 것입니다. x1 , x2 , . . . , xn . 여기서 xi는 p 변수를 포함하는 길이 p의 벡터입니다. i번째 관찰에 대한 측정값입니다. 즉, ⊠ ⊠ xi1 ⊠xi2 ⊠ ⊠ ⊠ xi = ⊠ . ⊠ . (1.1) ⊠ .. ⊠ 칩

(벡터는 기본적으로 열로 표시됩니다.) 예를 들어, 임금 데이터에서 xi는 연도, 나이, 인종 및 기타 항목으로 구성된 길이가 11인 벡터입니다. i 번째 개인의 가치. 다른 때에는 우리가 관심을 가질 것입니다. X의 열에 x1 , x2 , . . . , XP. 각각은 다음의

벡터입니다. 길이 n. 즉, ☒ ☒ x1j ☒ x2j ☒ ☒ ☒ xj = ☒ . ☒ . ☒ .. ☒ xnj

예를 들어, 임금 데이터의 경우 x1에는 연도에 대한 n = 3,000 값이 포함됩니다. 이 표기법을 사용하면 행렬 X는 다음과 같이 쓸 수 있습니다. ’ (X = x1 x2 · · · xp , 또는

☒ ☒ x1 ☒xT2 ☒ ☒ ☒ X = ☒ . ☒. ☒ .. ☒ 리틀 테네시

T 표기법은 행렬 또는 벡터의 전치를 나타냅니다. 예를 들어, ☒ ☒ x11 x21 . . . xn1 ☒x12 x22 . . . xn2 ☒ ☒ ☒ XT = ☒ . .. .. ☒ , ☒ .. . . ☒ x1p x2p . . . xnp

~하는 동안

’ xTi = x11

x2

...

(칩

yi를 사용하여 변수의 i번째 관측값을 나타냅니다. 임금 등을 예측하고 싶습니다. 따라서 우리는 모든 n의 집합을 씁니다. 벡터 형식의 관측값 ☒ ☒ y1 ☒ y2 ☒ ☒ ☒ y = ☒ . ☒. ☒ .. ☒ ~에

그런 다음 관측된 데이터는 {(x1 , y1), (x2 , y2), . . . , (xn, yn)}, 여기서 각 xi는 길이가 p인 벡터입니다. (p = 1이면 xi는 단순히 스칼라입니다.) 이 텍스트에서 길이 n의 벡터는 항상 소문자로 표시됩니다. 용감한; 예를 들어 ☒ ☒ a1 ☒ a2 ☒ ☒ ☒ a = ☒ . ☒. ☒ .. ☒ 안

그러나 길이 n이 아닌 벡터(예: 길이의 특징 벡터) p((1.1)에서와 같이)는 소문자 일반 글꼴로 표시됩니다. 에이. 스칼라는 또한 소문자 일반 글꼴로 표시됩니다. 에이. 드문 경우이지만 소문자 일반 글꼴에 대한 이 두 가지 용도는 모호성을 야기하므로 명확하게 설명하겠습니다. 어떤 용도로 사용되는지. 행렬은 굵은 대문자를 사용하여 표시됩니다. A와 같습니다. 임의 변수는 대문자 일반 글꼴을 사용하여 표시됩니다. 에이, 크기에 관계없이. 때때로 우리는 특정 객체의 차원을 나타내기를 원할 것입니다. 객체가 스칼라임을 나타내기 위해 a ∈ R 표기법을 사용합니다. 길이가 k인 벡터임을 나타내려면 ∈ Rk(또는 다음과 같은 경우에는 ∈ Rn)를 사용합니다. 길이는 n입니다.) 다음을 사용하여 객체가 r × s 행렬임을 나타냅니다. A ∈ Rr×s . 우리는 가능할 때마다 행렬 대수학을 사용하지 않았습니다. 그러나 몇몇 경우에는 완전히 피하기에는 너무 번거롭습니다. 이들에서는 드문 경우지만 곱셈의 개념을 이해하는 것이 중요합니다. 두 개의 행렬. A ∈ Rr×d 및 B ∈ Rd×s라고 가정합니다. 그러면 그 제품 A와 B의 AB를 표시합니다. AB의 (i, j)번째 요소는 다음과 같이 계산됩니다. A 의 i번째 행의 각 요소를 곱함) 해당 요소로 디 B의 j번째 열. 즉, (AB)ij = k=1 aik bkj 입니다. 예를 들어, 고려하다 + + 1 2 5 6 A= 그리고 B = . 3 4 7 8 그 다음에

AB =

•

1 2 3 4

+*

5 6 7 8

•

=

•

1×5+2×7 3×5+4×7

• -

• 1×6+2×8 19 22 = . 3×6+4×8 43 50

이 작업은 r × s 행렬을 생성합니다. 에만 가능합니다 A의 열 수가 A의 열 수와 같으면 AB를 계산합니다. B의 행.

1. 소개

11

이 책의 구성 2장에서는 통계 학습의 기본 용어와 개념을 소개합니다. 이 장에서는 또한 K-최근접 이웃 분류기인 많은 문제에 놀라울 정도로 잘 작동하는 매우 간단한 방법입니다. 3장과 4장에서는 회귀 및 분류를 위한 고전적인 선형 방법을 다룹니다. 특히 3장에서는 모든 회귀 방법의 기본 출발점인 선형 회귀를 검토합니다. 4장에서는 다음 중 두 가지를 논의한다. 가장 중요한 고전 분류 방법, 로지스틱 회귀 및 선형 판별 분석. 모든 통계 학습 상황의 핵심 문제는 선택과 관련됩니다. 특정 애플리케이션에 가장 적합한 방법입니다. 따라서 5장에서는 교차 검증과 부트스트랩을 소개합니다. 가장 좋은 방법을 선택하기 위해 다양한 방법의 정확성. 통계 학습에 관한 최근 연구의 대부분은 다음 사항에 집중되어 있습니다. 비선형 방법. 그러나 선형 방법은 종종 장점이 있습니다. 해석 가능성 측면에서 비선형 경쟁자이며 때로는 정확성. 따라서 6장에서는 다양한 선형 방법을 고려합니다. 표준 선형 회귀에 비해 잠재적인 개선을 제공하는 클래식 및 최신 버전입니다. 여기에는 단계적 선택, 능선 회귀, 주요 구성 요소 회귀 및 올가미. 나머지 장은 비선형 통계의 세계로 이동합니다. 학습. 우리는 먼저 7장에서 단일 입력 변수 문제에 대해 잘 작동하는 여러 비선형 방법을 소개합니다. 그러면 우리는 이러한 방법을 사용하여 비선형 추가 모델을 적합하게 하는 방법을 보여줍니다. 하나 이상의 입력이 있습니다. 8장에서는 트리 기반을 조사합니다. 배깅, 부스팅, 랜덤 포레스트 등의 방법을 사용합니다. 지원 벡터 기계, 선형 및 비선형을 모두 수행하기 위한 일련의 접근 방식 분류에 대한 자세한 내용은 9장에서 논의합니다. 우리는 비선형 회귀 및 분류에 대한 접근 방식으로 많은 주목을 받은 딥러닝을 다룹니다. 최근 몇 년간 주목을 받은 부분은 10장입니다. 11장에서는 생존을 탐구합니다. 분석은 다음과 같은 환경에 특화된 회귀 접근 방식입니다. 출력 변수는 검열됩니다. 즉, 완전히 관찰되지 않습니다. 12장에서는 감독되지 않는 환경을 고려합니다. 입력 변수는 있지만 출력 변수는 없습니다. 특히 주성분 분석, K-평균 클러스터링, 계층적 클러스터링을 소개합니다. 마지막으로 13장에서는 다중 가설 검정이라는

매우 중요한 주제를 다룹니다. 각 장의 마지막에는 하나 이상의 Python 실습 섹션이 제공됩니다. 여기서는 해당 장에서 논의된 다양한 방법을 적용하여 체계적으로 작업합니다. 이 실험실에서는 강점과 다양한 접근 방식의 약점을 파악하고 유용한 참고 자료도 제공합니다. 다양한 메소드를 구현하는 데 필요한 구문에 대해 설명합니다. 독자는 아마도 자신의 속도에 맞춰 실습을 진행하도록 선택하거나 실습이 교실 환경의 일부로서 그룹 세션의 초점. 각각 내에서 Python 연구실에서 수행했을 때 얻은 결과를 발표합니다. 이 책을 집필할 당시의 연구실. 그러나 새로운 버전의 Python은 지속적으로 출시되며 시간이 지남에 따라 실험실에서 호출된 패키지는 업데이트됩니다. 따라서 앞으로는 다음과 같은 결과가 나올 가능성이 있다.

12

1. 소개

이름

자동 자전거 공유 보스턴 뇌암 대상 카시트 대학 신용 거래 기본 축적 타자 칸 NCI60 뉴욕증권거래소 오제이 포트폴리오 출판 헤로인 미국 체포 값 주간

설명

자동차의 연비, 마력, 기타 정보를 제공합니다. 워싱턴 DC의 자전거 공유 프로그램 시간당 사용량. 보스턴 인구 조사 지역에 대한 주택 가치 및 기타 정보. 뇌암 진단을 받은 환자의 생존 시간. 캐러밴 보험을 제공받은 개인에 대한 정보입니다. 400개 매장의 카시트 판매 정보입니다. 미국 대학의 인구통계학적 특성, 등록금 등. 400명의 고객에 대한 신용카드 부채에 대한 정보입니다. 신용카드 회사의 고객 기본 기록입니다. 50개월간 헤지펀드 매니저 2,000명의 수익률. 야구선수의 기록과 연봉. 4가지 암 유형에 대한 유전자 발현 측정. 64개 암 세포주에 대한 유전자 발현 측정. 뉴욕 증권 거래소의 수익률, 변동성 및 거래량. 시트러스힐, 미닛메이드 오렌지주스 판매정보입니다. 포트폴리오 배분에 사용하기 위한 금융 자산의 과거 가치. 244건의 임상시험이 출판될 시간입니다. 5년 동안 S&P 500의 일일 백분율 수익률입니다. 미국 50개 주 주민 10만 명당 범죄 통계입니다. 미국 중부 대서양 지역 남성을 대상으로 한 소득 조사 데이터입니다. 21년 동안 주간 주식시장 수익률은 1,089회였습니다.

표 1.1. 이 실습과 연습을 수행하는 데 필요한 데이터 세트 목록 교과서. 다음을 제외한 모든 데이터 세트는 ISLP 패키지에서 사용할 수 있습니다. USArrests는 기본 R 배포판의 일부이지만 Python에서 액세스할 수 있습니다.

랩 섹션이 더 이상 얻은 결과와 정확하게 일치하지 않을 수 있습니다. 실습을 수행하는 독자가 작성합니다. 필요에 따라 업데이트를 게시할 예정입니다. 책 웹사이트의 연구실. 우리는 더 많은 내용이 포함된 섹션이나 연습을 나타내는 기호 도전적인 개념. 읽지 않는 독자는 쉽게 건너뛸 수 있습니다. 자료를 깊이 파고들고 싶거나 수학적 지식이 부족한 사람 배경.

실험실 및 연습에 사용되는 데이터 세트 이 교과서에서는 마케팅, 금융, 생물학 및 기타 분야의 응용 프로그램을 사용하여 통계 학습 방법을 설명합니다. ISLP 패키지 작업을 수행하는 데 필요한 여러 데이터 세트가 포함되어 있습니다. 이 책과 관련된 실습 및 연습. 다른 데이터 세트 중 하나는 다음의 일부입니다. 기본 R

분포(USArrests 데이터) 및 이에 액세스하는 방법을 보여줍니다. 섹션 12.5.1의 Python에서. 표 1.1에는 데이터 요약이 포함되어 있습니다. 실습과 연습을 수행하는 데 필요한 세트입니다. 이 데이터 세트 중 몇 가지 2장에서 사용할 수 있도록 도서 웹사이트에서 텍스트 파일로도 제공됩니다.

1. 소개

13

도서 웹사이트 이 책의 웹사이트는 다음과 같습니다. www.statlearning.com 여기에는 관련된 Python 패키지를 포함하여 다양한 리소스가 포함되어 있습니다. 이 책과 몇 가지 추가 데이터 세트를 사용하세요.

감사의 말 이 책의 플롯 중 일부는 ESL에서 가져온 것입니다. 그림 6.7, 8.3, 그리고 12.14. 다른 모든 플롯은 ISL의 R 버전에 대해 생성되었습니다. 그림 13.10의 경우 Python 소프트웨어 지원으로 인해 다릅니다. 줄거리.

2 통계적 학습

2.1

통계 학습이란 무엇입니까?

통계 학습에 대한 연구에 동기를 부여하기 위해 우리는 간단한 것부터 시작합니다. 예. 우리가 클라이언트에 의해 고용된 통계 컨설턴트라고 가정해 보겠습니다. 특정 제품의 광고와 판매 사이의 연관성을 조사합니다. 제품. 광고 데이터 세트는 해당 제품의 판매로 구성됩니다. 200개 시장에서 제품에 대한 광고 예산과 함께 각 시장에는 TV, 라디오, 신문이라는 세 가지 미디어가 있습니다. 데이터는 그림 2.1에 표시됩니다. 우리 고객은 불가능합니다. 제품 판매를 직접적으로 늘릴 수 있습니다. 반면에 그들은 통제할 수 있다. 3개 매체의 광고비. 그러므로 만일 우리가 광고와 판매 사이에 연관성이 있음을 확인한 다음 우리는 고객에게 광고 예산을 조정하도록 지시할 수 있으며, 이를 통해 간접적으로 매출 증가. 즉, 우리의 목표는 정확한 모델을 개발하는 것입니다. 이는 세 가지 미디어 예산을 기반으로 매출을 예측하는 데 사용할 수 있습니다. 이 설정에서는 광고 예산이 입력 변수가 되고 매출이 발생합니다. 입력 출력 변수입니다. 입력 변수는 일반적으로 변수를 사용하여 표시됩니다. 기호 X와 이를 구별하기 위한 아래 첨자가 있습니다. 따라서 X1은 TV 출력일 수 있습니다. 예산, 라디오 예산 X2, 신문 예산 X3입니다. 입력 변수 예측변수, 독립변수, 특징, 예언자 또는 때로는 단지 변수일 수도 있습니다. 출력 변수(이 경우 판매)는 다음과 같습니다. 독립적인 종종 반응변수 또는 종속변수라고 불리며, 일반적으로 변수로 표시됩니다. Y 기호를 사용합니다. 이 책 전체에서 우리는 이러한 용어를 모두 사용할 것입니다. 교대로. 변하기 쉬운 보다 일반적으로, 정량적 응답 Y 및 p 응답을 관찰한다고 가정합니다. 다양한 예측 변수 X1 , X2 , . . . , XP . 우리는 어떤 의존성이 있다고 가정합니다. Y와 X의 관계 = (X1, X2,⋯,Xp), 변수로 쓸 수 있음 아주 일반적인 형태로 Y = f(X) + “. (2.1)

15

0

50
100
200
300
25 20 5
10
15
매상
20 15
매상
5
10
15 5
10
매상
20
25
2. 통계적 학습
25
16
0
10
TV
20
30
40
50
0
20
라디오
40

60

80

100

신문

그림 2.1. 광고 데이터 세트. 플롯에는 판매량이 천 단위로 표시됩니다. TV, 라디오, 신문 예산의 함수로 단위를 수천 단위로 표시합니다. 200개 시장에 대한 달러입니다. 각 플롯에는 간단한 최소 제곱이 표시됩니다. 3장에 설명된 대로 해당 변수에 대한 매출의 적합성. 즉, 각 파란색 선은 TV, 라디오를 사용하여 매출을 예측하는 데 사용할 수 있는 간단한 모델을 나타냅니다. 그리고 신문.

여기서 f는 X1의 고정되었지만 알려지지 않은 함수입니다. . . , Xp 및 “는 무작위입니다. X와 독립적이고 평균 0을 갖는 오류항입니다. 이 공식에서 오류항 f는 X가 Y에 대해 제공하는 체계적인 정보를 나타냅니다. 80 70 60 50

소득

20

30

40

50 20

30

40

소득

60

70

80

체계적인

10

12

14

16

18

교육 기간

20

22

10

12

14

16

18

20

22

교육 기간

그림 2.2. 소득 데이터 세트. 왼쪽: 빨간색 점은 관측된 값입니다. 30명의 개인에 대한 소득(천 달러)과 교육 기간입니다. 오른쪽: 파란색 곡선은 소득과 소득 간의 실제 기본 관계를 나타냅니다. 일반적으로 알려지지 않은 교육 년수(그러나 이 경우에는 알려져 있음) 데이터가 시뮬레이션되었기 때문입니다.) 검은색 선은 관련된 오류를 나타냅니다. 관찰할 때마다. 일부 오류는 긍정적입니다(관찰 결과가 거짓인 경우). 파란색 곡선 위) 일부는 음수입니다(관측치가 파란색 곡선 아래에 있는 경우). 곡선). 전반적으로 이러한 오류는 대략 0을 의미합니다.

또 다른 예로 그림 2.2의 왼쪽 패널을 살펴보겠습니다.

소득 데이터 세트에서 30명의 개인에 대한 소득 대 교육 기간. 플롯은 수년간의 데이터를 사용하여 소득을 예측할 수 있음을 시사합니다. 교육. 그러나 입력 변수를 연결하는 함수 f는

출력 변수는 일반적으로 알 수 없습니다. 이 상황에서는 추정해야합니다 f 관찰된 점을 기반으로 합니다. 소득은 시뮬레이션된 데이터 세트이므로 f는 알려져 있으며 그림 2.2의 오른쪽 패널에 파란색 곡선으로 표시됩니다. 수직선은 오류 용어 “를 나타냅니다. 우리는 30개의 관측치가 파란색 곡선 위에 있고 일부는 그 아래에 있습니다. 전반적으로, 오류는 대략 0을 의미합니다. 일반적으로 함수 f에는 둘 이상의 입력 변수가 포함될 수 있습니다. 그림 2.3에서는 교육 기간의 함수로 소득을 표시하고 선임 순위. 여기서 f는 추정되어야 하는 2차원 표면입니다. 관찰된 데이터를 기반으로 합니다. 본질적으로, 통계적 학습은 추정을 위한 일련의 접근법을 의미합니다. 에프. 이 장에서는 발생하는 주요 이론적 개념 중 일부를 개괄적으로 설명합니다. f 추정 및 얻은 추정을 평가하기 위한 도구.

2.1.1

왜 f를 추정하는가?

f를 추정하려는 두 가지 주요 이유는 다음과 같습니다. 그리고 추론. 우리는 각각을 차례로 논의합니다. 예측 많은 상황에서 입력 X 세트는 쉽게 사용할 수 있지만 출력은 Y는 쉽게 구할 수 없습니다. 이 설정에서는 오차항의 평균이 0으로, 우리는 다음을 사용하여 Y를 예측할 수 있습니다. ⊠ = f⊠(X),

(2.2)

여기서 f⊠는 f에 대한 추정값을 나타내고, ⊠는 Y에 대한 결과 예측을 나타냅니다. 이 설정에서 f⊠는 종종 블랙박스로 취급됩니다. 일반적으로 f⊠의 정확한 형태에는 관심이 없습니다. Y 에 대한 정확한 예측을 산출합니다. 예를 들어 X1, . . . , Xp는 환자의 특성입니다. 실험실에서 쉽게 측정할 수 있는 혈액샘플, Y는 변수 특정 약물에 대한 심각한 부작용에 대한 환자의 위험을 인코딩합니다. 의약품. X를 사용하여 Y를 예측하는 것은 자연스러운 일입니다. 부작용의 위험이 높은 환자에게 문제의 약물을 투여하는 경우 반응, 즉 Y 추정치가 높은 환자입니다. Y에 대한 예측으로서 ⊠의 정확도는 두 가지 수량에 따라 달라집니다. 이를 환원 가능한 오류와 환원 불가능한 오류라고 부르겠습니다. 일반적으로, 축소할 수 있는 f⊠는 f에 대한 완벽한 추정이 아니며, 이러한 부정확성으로 인해 오류가 발생합니다. 일부 오류. 이 오류는 환원 불가능한 오류를 잠재적으로 개선할 수 있기 때문에 축소 가능합니다. 오류에 가장 적합한 통계 학습 기법을 사용하여 f⊠의 정확도 추정 f . 그러나 완벽한 추정이 가능하더라도 f 이므로 추정된 응답은 ⊠ = f(X) 형식을 취하므로 예측은 다음과 같습니다. 여전히 오류가 있을 수 있습니다! 이는 Y도 다음의 함수이기 때문입니다. “는 정의에 따라 X를 사용하여 예측할 수 없습니다. 따라서 변동성은”와 관련된 것도 우리 예측의 정확성에 영향을 미칩니다. 이것은 알려져 있습니다. 환원 불가능한 오류로, f를 아무리 잘 추정하더라도 “로 인한 오류를 줄일 수 없습니다. 환원 불가능한 오류가 0보다 큰 이유는 무엇입니까? 수량”에는 Y를 예측하는 데 유용한 측정되지 않은 변수가 포함될 수 있습니다.

RS

~의

Se ~에 또는

ity

이자형 수입

이인칭 대명사 에이

에드

UC

우리는 가지고 있었다

~에

그림 2.3. 도표는 교육 기간에 따른 소득을 표시합니다. 소득 데이터 세트의 연공서열. 파란색 표면은 실제를 나타냅니다. 소득과 교육 기간 및 연공서열 사이의 기본 관계, 이는 데이터가 시뮬레이션되었기 때문에 알려진 것입니다. 빨간색 점은 관찰된 것을 나타냅니다. 30명에 대한 이 수량의 값입니다.

측정하면 f는 예측에 사용할 수 없습니다. 수량은 “일 수 있습니다. 측정할 수 없는 변동도 포함되어 있습니다. 예를 들어, 불리한 상황의 위험 반응은 특정 날짜에 특정 환자마다 다를 수 있습니다. 약물 자체의 제조 변화 또는 환자의 일반적인 느낌 그날의 안녕. 주어진 추정값 f⊠와 예측변수 세트 X를 고려하면 다음과 같습니다.

예측 ☒ = f☒(X). f☒와 X가 모두 고정되어 있다고 잠시 가정해 보겠습니다. 유일한 가변성은"에서 비롯됩니다. 그러면 다음을 쉽게 보여줄 수 있습니다. E(Y ☒ ☒)2

=

E[f(X) + " ☒ f☒(X)]2 [f (X) ☒ f☒(X)]2 + Var(“) , , -. / , -. / 환원 가능

(2.3)

줄일 수 없는

여기서 E(Y ☒ ☒)2는 제곱의 평균 또는 기대값을 나타냅니다. 예상되는 Y 의 예측 값과 실제 값 사이의 차이 및 Var(“) 표현 값 오류 용어"와 관련된 분산을 전송합니다. 변화 이 책의 초점은 다음을 목표로 f를 추정하는 기술에 있습니다. 감소 가능한 오류를 최소화합니다. 다음을 명심하는 것이 중요합니다. 환원 불가능한 오류는 항상 정확도의 상한선을 제공합니다. Y에 대한 예측입니다. 이 경계는 실제로는 거의 항상 알려져 있지 않습니다. 추론 우리는 종종 Y와 Y 사이의 연관성을 이해하는 데 관심이 있습니다. X1 , . . . , XP . 이 상황에서 우리는 f를 추정하고 싶지만 목표는 그렇지 않습니다. 반드시 Y에 대한 예측을 해야 합니다. 이제 f☒는 검정으로 취급될 수 없습니다. 상자의 정확한 형태를 알아야 하기 때문입니다. 이 설정에서는 다음 중 하나가 가능합니다. 다음 질문에 답변하고 싶습니다.

2.1 통계 학습이란 무엇입니까?

19

• 어떤 예측변수가 반응과 연관되어 있습니까? 종종 그런 경우가 있습니다 사용 가능한 예측 변수 중 극히 일부만이 실질적으로 Y 와 연관되어 있습니다. 몇 가지 중요한 예측 변수를 식별합니다. 가능한 변수의 큰 집합은 다음에 따라 매우 유용할 수 있습니다. 응용 프로그램. • 반응과 각 예측 변수 간의 관계는 무엇입니까? 일부 예측 변수는 다음과 같은 의미에서 Y 와 양의 관계를 가질 수 있습니다. 예측 변수의 더 큰 값은 더 큰 값과 연관되어 있습니다. Y. 다른 예측 변수는 반대 관계를 가질 수 있습니다. 따라 f의 복잡도에 따라 응답과 a 사이의 관계 주어진 예측변수는 다른 예측변수의 값에 따라 달라질 수도 있습니다. • Y와 각 예측 변수 간의 관계를 선형 방정식을 사용하여 적절하게 요약할 수 있습니까? 아니면 관계가 더 복잡합니까? 역사적으로 f를 추정하는 대부분의 방법은 선형을 취했습니다. 형태. 어떤 상황에서는 그러한 가정이 합리적이거나 심지어 바람직할 수도 있습니다. 그러나 종종 실제 관계는 더 복잡합니다. 선형 모델이 정확한 표현을 제공하지 못하는 경우 입력 변수와 출력 변수 사이의 관계. 이 책에서 우리는 예측에 해당하는 여러 가지 예를 볼 것입니다. 설정, 추론 설정 또는 이 둘의 조합입니다. 예를 들어, 다음과 같은 업무 수행에 관심이 있는 회사를 생각해 보십시오. 직접 마케팅 캠페인. 목표는 다음과 같은 개인을 식별하는 것입니다. 각 개인에 대해 측정된 인구통계학적 변수를 관찰한 결과, 메일링에 긍정적으로 반응할 가능성이 높습니다. 이 경우 인구통계학적 변수는 예측 변수 역할을 하며 마케팅 캠페인에 대한 반응(긍정적 또는 부정적)은 결과 역할을 합니다. 회사는 각 개별 예측 변수와 반응 간의 관계에 대한 깊은 이해를 얻는 데 관심이 없습니다. 대신 회사는 단순히 예측변수를 사용하여 반응을 정확하게 예측하려고 합니다. 이것 예측을 위한 모델링의 예입니다. 대조적으로, 그림 2.1에 설명된 광고 데이터를 고려하십시오.

하나 다음과 같은 질문에 답하는 데 관심이 있을 수 있습니다. - 판매와 관련된 미디어는 무엇입니까? - 매출이 가장 많이 증가하는 미디어는 무엇입니까? 또는 - 특정 증가에 따른 매출 증가의 크기는 얼마나 됩니까? TV 광고에서? 이 상황은 추론 패러다임에 속합니다. 또 다른 예는 다음과 같습니다. 고객이 구매할 수 있는 제품의 브랜드 모델링 가격, 매장 위치, 할인 수준, 경쟁 가격, 등등. 이런 상황에서 우리는 정말로 다음 사항에 가장 관심을 가질 수 있습니다. 각 변수와 구매 확률 간의 연관성. 예를 들어, 제품 가격이 판매와 어느 정도 연관되어 있습니까? 이것은 추론을 위한 모델링의 예. 마지막으로 예측과 추론을 위해 일부 모델링을 수행할 수 있습니다. 예를 들어, 부동산 환경에서는 가치를 연관시키려고 할 수 있습니다.

20

2. 통계적 학습

범죄율, 구역 설정, 강으로부터의 거리, 공기와 같은 입력에 대한 주택 수 품질, 학교, 지역 사회의 소득 수준, 주택 규모 등. ~ 안에 이 경우 각 개별 입력 변수와 주택 가격 간의 연관성에 관심이 있을 수 있습니다. 강이 보이는 집이라면 가치가 있을까요? 이것은 추론 문제입니다. 또는 단순히 특정 값을 예측하는 데 관심이 있을 수도 있습니다. 그 특성을 고려한 집: 이 집은 과소평가된 것인가, 아니면 과대평가된 것인가? 이것은 예측 문제. 우리의 궁극적인 목표가 예측인지, 추론인지, 아니면 예측인지에 따라 f를 추정하기 위한 두 가지 다른 방법을 조합하는 것이 적절할 수 있습니다. 예를 들어 선형 모델은 상대적으로 단순하고 선형적인 모델을 허용합니다. 해석 가능한 추론이지만 일부만큼 정확한 예측을 산출하지 못할 수도 있습니다. 다른 접근법. 대조적으로, 매우 비선형적인 접근법 중 일부는 이 책의 후반부에서 논의할 내용은 잠재적으로 다음과 같은 이점을 제공할 수 있습니다. Y에 대해 매우 정확한 예측을 할 수 있지만 이는 더 적은 비용으로 발생합니다. 추론이 더 어려운 해석 가능한 모델입니다.

2.1.2

f를 어떻게 추정하나요?

이 책 전체에서 우리는 다양한 선형 및 비선형 접근 방식을 탐구합니다. f를 추정하기 위해. 그러나 이러한 방법은 일반적으로 특정 특성을 공유합니다. 우리는 이 공유 특성에 대한 개요를 제공합니다. 부분. 우리는 항상 n개의 서로 다른 집합을 관찰했다고 가정합니다. 데이터 포인트. 예를 들어 그림 2.2에서는 n = 30개의 데이터 포인트를 관찰했습니다. 이러한 관찰을 훈련 데이터라고 합니다. 왜냐하면 우리가 이것을 사용할 것이기 때문입니다. 훈련 f를 추정하는 방법을 훈련하거나 가르치기 위한 관찰입니다. xij 데이터를 보자 관측치 i에 대한 j번째 예측 변수 또는 입력의 값을 나타냅니다. 여기서 나는 = 1, 2, . . . , n 및 j = 1, 2, . . . , p. 이에 따라 yi가 i번째 관측값에 대한 응답 변수입니다. 그런 다음 훈련 데이터는 다음과 같이 구성됩니다.