Introduction to Statistical Learning

Statistical Learning

수업 교재

- 수업 교재: **An Introduction to Statistical Learning with Applications in R (2nd edition), James, Witten, Hastie, and Tibshirani**, Springer (한국어 교재 존재)

Introduction

인공지능과 머신러닝?

- **인공지능(Artificial Intelligence, AI)**: 컴퓨터로 만들어진 지능으로 `사람에 의해 통상적으로 수행되는 지적인 산업을 자동화하는 것을 목표'(Chollet, 2018)로 하는 학문 영역
- **머신러닝(Machine learning, ML)**: 4차 산업혁명의 핵심인 **통계적 머신러닝(statistical machine learning)**과 **딥러닝(deep learning)**을 포괄하는 광범위한 개념
  - 데이터와 해답을 입력하고 이 둘의 관계를 규명하여 통계적 규칙을 출력하는 구조
  - 허용 가능한 오차가 존재, 규칙은 손실함수를 최소화함으로써 구현
  - 손실함수: 오차의 제곱합이나 음의 로그 우도함수(log likelihood) 등

통계적 머신러닝과 통계학의 차이 (1)

- 전통적인 통계학은 출력결과인 통계적 규칙에 대한 통계적 추론(inference), 모형의 타당성 검증, 모형에 포함된 모수 추정량과 **예측**에 대한 통계적 성질 규명 등에 중점을 둠
- 통계적 머신러닝에서는 모형을 설정하고 모수를 추정할 학습데이터(training data)와 검증을 위한 시험데이터(test data)를 미리 분리, 학습데이터로 모형을 설정하고 모수를 추정한 후, 시험데이터에 적용하여 학습데이터만큼의 성능이 있는지를 점검 (일반화 과정)
- 통계적 추론이 거의 불가능한 데이터 적응(adaptive) 머신러닝 기법, random forest, boosting, XGBoosting 등도 어렵지 않게 일반화 과정을 통해 모형의 성능과 타당성을 쉽게 점검 가능

통계적 머신러닝과 통계학의 차이 (2)

- 전통적인 통계학은 통계학을 다변량, 시계열, 범주형 등으로 자료의 특성에 따라 분류하고 세분화
- 통계적 머신러닝에서는 단순하게 목적변수의 관측여부에 따라 **지도학습(supervised learning)**과 **비지도학습(unsupervised learning)**으로 구분
- **지도학습**: 목적변수 y와 이를 설명해주는 설명변수 x가 모두 관측
  - Regression, logistic regression, decision tree, support vector machine 등
- **비지도학습**: 설명변수 x만 관측되고 y가 관측되지 않은 경우
  - Principal component analysis, clustering, matrix decomposition 등
- **준지도학습**: 관측치 중 일부는 설명변수와 반응변수 측정값을 모두 갖고 있고, 나머지 관측치에 대해서는 설명변수 측정값만 있고 반응변수 측정값은 없는 경우
- 강화학습: 주어진 환경에 의해 시스템의 성능을 향상시키는 머신러닝이며 게임이나 로봇공학에 주로 사용

회귀와 분류문제

- 양적 변수: 수치 값을 취하는 것으로 사람의 나이, 키 또는 수입, 집 값, 주식 가격 등이 해당됨
- 질적 변수: K개의 다른 클래스(classes) 또는 카테고리(category)중 하나를 값으로 가지며, 사람의 성별(남/여), 어떤 사람의 채무 지불 여부(연체 또는 연체 아님) 등을 포함
- 양적 반응변수를 가지는 문제는 **회귀(regression)** 문제라 하고, 질적 반응변수를 가지는 문제는 **분류(classification)** 문제라고 부름

통계적 머신러닝과 딥러닝의 비교

- 공통점: 통계적 이론과 방법이 일치하는 손실함수, 모수추정방법, 일반화 과정
- 차이점: 모형의 설계, 통계적 머신러닝에 비해 복잡한 최적화 과정

1. 데이터 크기: 중/소 크기 vs 빅데이터
2. 분석자료 형태: 2차원 텐서 vs 2차원 이상 텐서
3. 최적화에 사용되는 자료: (일반적으로) 전체 데이터 vs 배치데이터
4. 강점을 갖는 자료: 정형화된 자료 vs 비정형자료 (이미지, 소리, 언어 등)
5. 모형: 매우 많음 vs (기본적으로) 3개

- 딥러닝은 기본적으로 Multilayer perceptron (MLP), Convolutional neural network (CNN), Recurrent neural network (RNN)의 변형과 조합

6. 특성변수의 정규화 및 표준화: 선택 vs 필수
7. 해석여부: (대체로) 쉬움 vs 어렵거나 불가능

머신러닝의 분석 절차

0. 문제의 정의
1. 자료의 수집: 공개된 데이터, 비공개 데이터, 수집 데이터, 시뮬레이션 데이터
2. 자료의 사전정리(preprocessing): 누락된 데이터 제거?, 이상치, 자료 변환
3. 학습: 지도, 비지도, 강화학습? …
4. 평가
5. 예측

머신러닝의 분석 절차: 자료의 사전정리

- 숫자 자료뿐만 아니라 색깔, 위치, 형태 등의 이미지를 실수로 구성된 텐서자료로 전환
- 자료에 포함된 결측치를 대체 또는 제외
- 변수의 특성 파악, 분석의 조건과 모형선택의 기준 등 파악 (기초통계, 산포도, 히스토그램, 상관관계 등을 이용하여 특성변수 선택, 특이치 유무, 특성변수 및 출력변수 분포의 형태 등)
- 정규화나 표준화

머신러닝의 분석 절차: 학습 및 검증

- 학습데이터와 시험데이터로 표본을 분할 (전통적인 통계학의 표본추출기법)
- 학습데이터: 분석모형의 선택과 모수추정
- 최적화: 손실함수를 정의하고 이를 최소화하는 모수를 추정
- 주의점: 모형의 성능이 학습데이터에만 최대화될 수 있음
- 학습된 모형을 검증데이터(학습데이터의 일부)에 적합시켜, 모형의 성능이 학습데이터에서 구현된 성능과 비슷한 수준인지 점검하고 초모수를 조율

머신러닝의 분석 절차: 평가

- 시험데이터: 학습데이터에서 추정된 모형의 일반화성능을 측정, 과대적합/과소적합 여부 판단
- 과대적합: 일반적으로 시험데이터에서의 손실함수값이 학습데이터의 손실함수값보다 큼
- 해결방안: 자료의 크기를 증가시켜 과대적합을 해결, 크기를 증가시키기 어려울 경우, 모수에 제한을 두는 **규제화(regularization)**나 **앙상블학습(ensemble learning)** 활용
- 규제화: 모형이 너무 복잡하여 과대적합이 발생했을 때 모형을 단순화하기 위해 사용하는 방법
- 앙상블학습: 통계적 기법을 통해 학습데이터의 크기를 인위적으로 증가시키는 방법 (bagging, random forest, boosting, XGBoost 등)

머신러닝에 활용 가능한 데이터셋

- **Kaggle**: 데이터과학자와 머신러닝 분석자들을 위한 가장 큰 커뮤니티
- **US Government Open Data**: 농업, 기상, 금융 등 수천 개의 데이터를 제공
- **Amazon Web Service Data**: NASA NEX 데이터, 독일 뱅크 공공자료 등 제공
- **Google Dataset Search**: 수많은 공공자료 제공
- **UCI ML Repository**: 머신러닝과 딥러닝 실습을 위한 잘 알려진 데이터 저장소
- **공공데이터포털**
- **기상청 기상자료개방포털**

`Wage` 자료 (연속적 출력값, 회귀문제)

- 미국의 대서양 지역에 거주하는 한 그룹의 남성들에 대한 임금(`wage`)과 관련된 여러 가지 요소들에 대해 살펴보는 자료
- 고용인의 `age`(나이)와 `education`(교육), 그리고 임금을 받은 `year`(연도) 사이의 관련성을 이해하는 것이 목표

```
       year age           maritl     race       education             region
231655 2006  18 1. Never Married 1. White    1. < HS Grad 2. Middle Atlantic
86582  2004  24 1. Never Married 1. White 4. College Grad 2. Middle Atlantic
161300 2003  45       2. Married 1. White 3. Some College 2. Middle Atlantic
155159 2003  43       2. Married 3. Asian 4. College Grad 2. Middle Atlantic
11443  2005  50      4. Divorced 1. White      2. HS Grad 2. Middle Atlantic
376662 2008  54       2. Married 1. White 4. College Grad 2. Middle Atlantic
              jobclass          health health_ins  logwage      wage
231655   1. Industrial      1. <=Good      2. No 4.318063  75.04315
86582  2. Information 2. >=Very Good      2. No 4.255273  70.47602
161300   1. Industrial      1. <=Good     1. Yes 4.875061 130.98218
155159 2. Information 2. >=Very Good     1. Yes 5.041393 154.68529
11443  2. Information      1. <=Good     1. Yes 4.318063  75.04315
376662 2. Information 2. >=Very Good     1. Yes 4.845098 127.11574
```

- `wage`는 나이에 따라 증가하지만 대략 60세 이후에는 다시 줄어듬
- `year`와 `education` 또한 `wage`와 관련이 있음
- 주어진 남성의 `wage`에 대한 가장 정확한 예측을 하려면 `age`, `year`와 `education`를 함께 고려해야 함

주식시장 자료 (범주형 출력값, 분류문제)

- `Smarket` 자료는 2001년과 2005년 사이 5년에 걸친 S&P 주가지수의 일일 변동량을 포함하는 자료
- 주어진 날짜의 주식시장이 상승(Up) 또는 하강(Down)하는지를 예측하는 문제

```
  Year   Lag1   Lag2   Lag3   Lag4   Lag5 Volume  Today Direction
1 2001  0.381 -0.192 -2.624 -1.055  5.010 1.1913  0.959        Up
2 2001  0.959  0.381 -0.192 -2.624 -1.055 1.2965  1.032        Up
3 2001  1.032  0.959  0.381 -0.192 -2.624 1.4112 -0.623      Down
4 2001 -0.623  1.032  0.959  0.381 -0.192 1.2760  0.614        Up
5 2001  0.614 -0.623  1.032  0.959  0.381 1.2057  0.213        Up
6 2001  0.213  0.614 -0.623  1.032  0.959 1.3491  1.392        Up
```

- 왼쪽부터 차례로 백분율로 나타낸 전날의 주가지수 변동, 다음날 주식시장이 증가한 647일, 다음날 주식시장이 감소한 602일에 대한 도표임
- 두 도표는 거의 동일하게 보이는데, 이것은 어제의 S&P 지수 움직임을 이용하여 오늘의 수익율을 예측할 수 있는 간단한 방법이 없음을 시사
- 흥미롭게도, 이 데이터에는 강하지는 않지만 어떤 추세에 대한 힌트가 있으며, 적어도 이 5년의 기간에 대해서는 대략 60% 정도 시장의 움직임의 방향을 정확하게 예측할 수 있다고 함

유전자 발현 자료 (클러스터링)

- `NCI60` 자료는 64개의 암 세포주(cell lines) 각각에 대한 6,830개의 유전자 발현 측정치임
- 세포주별로 수천 개의 유전자 발현 측정치가 있어 데이터를 시각화하기 어렵기 때문에, 특정 출력변수를 예측하는 대신 세포주들 사이에 그룹 또는 클러스터들이 있는지 결정하는 것이 관심을 둠
- 아래 그림은 64개 세포주 각각을 단지 두 개의 주성분(principal components) Z_1과 Z_2를 사용하여 표현함으로써 일부 정보의 손실을 초래하지만, 데이터를 시각적으로 살펴볼 수 있게 함

Data scientist가 되기 위해 필요한 지식

- AI의 뿌리 중의 하나는 통계학 (Quora Digest, 2019/09/02)

Data scientist가 되기 위해 필요한 지식: 표본의 조건

- 표본: 관측치, 여러 개의 특성변수와 목적변수로 구성, 같은 분포에서 독립적으로(identically and independent) 관측되어야 하는 것이 가장 중요
- 주식가격: 과거의 가격에 영향을 받아 독립적일 수 없음. 추가적으로 정상성(stationary) 여부에 따라 자료의 변환이 필요
- 전체 데이터를 학습데이터, 검증데이터, 시험데이터로 분류
- 표본추출: 단순임의표본추출(simple random sampling)과 목적변수가 범주형인 경우 범주에 따른 층화추출(stratified random sampling)의 개념과 방법

Data scientist가 되기 위해 필요한 지식: 대수의 법칙(law of large number)

- 대수의 법칙: 표본으로 만들어진 평균은 모평균으로 수렴
- 빅데이터의 경우, 통계학에서 사용하는 기대치를 간단하게 표본평균으로 바꿔 써도 된다는 이론적 근거이자 앙상블학습의 이론적 배경

Data scientist가 되기 위해 필요한 지식: Bayesian 개념

- 판별분석, Naive Bayesian 등은 아주 간단한 베이지안 분석방법
- 규제화는 통계학에서의 검증과 같은 역할, 과대적합 문제를 해결하는 중요한 도구
- 규제화의 해석은 Bayesian의 개념적 이해를 바탕으로 함

Data scientist가 되기 위해 필요한 지식: 임의성(randomness)

- 통계모형에서의 임의성은 오차항으로 구현, i.i.d. (identically and independently distributed)하다는 가정
- 모든 통계모형은 다음과 같이 구성됨
  목적변수 = 특정변수의 선형 또는 비선형 함수 + 오차항
- 목적변수에서 특성변수의 함수인 통계모형을 빼면 오차항이 되므로 통계모형은 가정에 부합되는 특성변수의 함수를 찾고 추정하는 것으로 요약 가능
- 가정에 부합되는 오차항이 존재한다는 것은 모형이 잘 선택되었고 모수도 잘 추정되었다는 의미
- 따라서 다른 데이터에 적용을 하더라도 오차항의 임의성 때문에 모형의 성능이 유지됨
- 임의성의 강화를 위해 머신러닝에서는 규제화, dropout 등의 기법을 사용

통계학습의 간단한 역사

- 19세기 초반: 르장드르(Legendre)와 가우스(Gauss)가 최소제곱법(least squares)에 대한 논문을 발표하며 후에 선형회귀로 알려진 형태를 최초 구현
- 1936년: 피셔(Fisher)가 질적 값들을 예측하기 위한 선형판별분석(linear discriminant analysis, LDA) 제안
- 1940년대: 다양한 저자들이 LDA의 대안적 방법은 로지스틱 회귀(logistic regression)를 제안
- 1970년대: 넬더(Nelder)와 웨더번(Wedderburn)은 전체 통계학습방법들에 대해 일반화된 선형모형(generalized linear model)을 만들었으며, 선형회귀와 로지스틱 회귀는 특수한 경우로 이에 포함됨
- 1980년대: 브라이먼(Breiman), 프리드먼(Friedman), 올쉔(Olshen), 스톤(Stone)은 분류 및 회귀나무(classifcation and regression tree)를 도입하였고, 모델 선택을 위한 교차검증을 포함하여 처음으로 상세하고 실질적인 구현의 유용성에 대해 보여줌
- 1986년: 해스티(Hastie)와 티브시라니(Tibshirani)가 일반화된 선형모형의 비선형적 확장에 대해 일반화가법모형(generalized additive model)이란 용어를 만듬

Data sets on the textbook

Notation and data matrix

- Wage data set, `data(wage)` in `R` along with `library(ISLR2)`
- n (number of observations): $n = 3000$
- p (number of variables): $p = 11$
- Variable name: `year`, `age`, `race`, and more

Notation and data matrix

- $\mathbf{X} = \{x_{ij}\}, i = 1, \ldots, n; j = 1, \ldots, p$: $(n \times p)$ 행렬
- $$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$
- x_{ij}: 행렬 $\mathbf{X}$의 (i, j)번째 원소
- x_i: i번째 행벡터(row vector)

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

Notation and data matrix

- $\mathbf{x}_j$: j번째 열벡터(column vector)

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{pmatrix}$$

- ^{T} : 행렬 또는 벡터의 전치(transpose)를 나타냄

$$x_i^T = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}, \quad \mathbf{X}^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}$$

Notation and data matrix

- 객체의 차원
  - 스칼라: $a \in \mathbb{R}$
  - 길이가 n인 벡터: $\mathbf{a} \in \mathbb{R}^n$
  - 길이가 n이 아닌 $(k \neq n)$ 벡터: $a \in \mathbb{R}^k$
  - 객체가 $r \times s$인 행렬: $\mathbf{A} \in \mathbb{R}^{r \times s}$
- 행렬의 곱셈: 행렬 $\mathbf{A}$와 $\mathbf{B}$의 곱은 $\mathbf{AB}$로, $(\mathbf{AB})_{ij} = \sum_{i=1}^{d} a_{ik} b_{kj}$
- 행렬의 곱셈 예제: $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$이면,

$$\mathbf{AB} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \times 5 + 2 \times 7 & 1 \times 6 + 2 \times 8 \\ 3 \times 5 + 4 \times 7 & 3 \times 6 + 4 \times 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}.$$

Notation and data matrix

- y_i: i번째 목적변수 관측치 (예측대상)

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

- $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$: 관측된 데이터셋의 집합 표현
  - x_i: 길이 p인 벡터
  - y_i: 스칼라(scalar)
- (X, Y): 확률벡터로 관찰된 데이터셋을 대표함, 특정 확률 분포를 따른다고 가정하면, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$은 해당 분포에서 i.i.d.로 추출한 표본으로 간주
- Note: X가 p차원이면 $X = (X_1, \ldots, X_n)^T$로 표현

Reference

- James, G., Witten, D., Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*, 2nd edition. Springer. Chapter 1.
- 박유성. *파이썬을 이용한 통계적 머신러닝*, 제2판. 자유아카데미. 1장.

Error ×