Ch. 2
Statistical Learning

https://www.youtube.com/watch?v=cyCJMVNhfnI&t=0s

ISLR 강의 by Hastie & Tibshirani

- https://www.youtube.com/playlist?list=PLwyWqcwXkazRpdF7AUDb4GK3Nn8f9jUoK

강의 by Hamed Hasheminia

- https://www.youtube.com/playlist?list=PL06ytJZ4Ak1rXmlvxTyAdOEfiVEzH00IK

보세요

문제가 있나요?

- 영어?
  - 피할수 없어요
  - 이 분야 대부분 자료, 지식이 영어로 되었어요
  - 이런 자료를 이용 못하고 영어로 소통 못하면 3류에 그쳐요
  - 다른 나라에 가 일하려면 필요해요
- 이론은 필요없고 그냥 해법(cookbook) 만 알려줘요
  - 이 정도의 이론도 모르는 사람이 한 분석을 나라면 "콩으로 메주를 쑨다 해도 안 믿어요"
- 수학?
  - 교재의 수학은 고등학교 1학년 수준을 넘지 않아요. 그나마도 개념만 알면 되요. 고1 수학책 봐요.

Supervised Learning (지도 학습)

- 입력 X와 그에 따라 원하는 결과 Y가 훈련 데이터(Training Set)로 주어질 때,

  Y = $f(X)$: X, Y 로 부터 f 를 *잘~~ 추정함이 목표*

  - *(X, Y)* 는 관측된 데이터

- **X** : 입력(input), 특징(feature/attribute), 독립변수(independent variable), predictor, covariate, 또는 그냥 "변수" 라고도 하며 보통 vector 형태
- **Y** : 결과(response, output, outcome), 종속변수(dependent variable), target 등으로 불림
- 학습은 **X, Y**를 이용해 **함수(알고리즘, 모델)** f 를 추정(estimate)하는 것

Supervised Learning (지도학습) – cont.

- 학습을 통해 추정한 $\hat{f}$ 는 완벽하지 않아 $\hat{f}(X)$가 실제로 Y와 똑같이 되기 어려움. $\hat{f}(X)$를 $\hat{\mathbf{Y}}$ (Y-hat) 으로 표기. $\hat{Y} \rightarrow Y$ 되도록 함이 학습의 목표
- Y가 숫자인 학습형태를 **regression** 이라 함
- Y가 카테고리 범주형이면 **classification** 이라 함
- $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ x_i : input, y_i : target response 쌍으로 구성된 training set (학습을 시키기 위한 훈련 자료)를 통해 $\hat{f}$를 구하는 과정이 학습
- Training set 학습으로 구한 $\hat{f}$에 테스트 데이터 (test set)를 적용하여 $\hat{f}$의 성능을 판단. 이를 Test Performance.
- Test Set: 모델을 훈련할 때 쓰지 않은 데이터(Training set에 없는 데이터)
- 개별 (x_i, y_i)를 instance, observation 또는 example이라 함

Is it a Classification or a Regression?

- 모호할 때도 많다
- 많은 Classification이 사실은 내부적으로 Regression

  - 가령, target 클래스 레이블들이 확률분포...

What are ;

Multiclass(multinomial) Classification

Multilabel Classification

Multioutput Regression

Multioutput Classification

- Is it a Dog or Cat : Binary

  Ans : Dog

- Is it a Dog or Cat or Chicken : Multiclass

  Ans : Dog

- Are there Dog, Cat, and Rabbit : Multilabel

  Ans : Yes, Yes, No (wrong ans)

- Multioutput : for each output label
  do Multiclass operation

  Ans : Dog, Cat, Rabbit

머신러닝을 나누는 또 다른 방법 – 훈련 데이터를 어떻게 활용할까?

- Batch Learning
  - 갖고 있는 모든 training set 데이터를 한꺼번에 이용해 학습
  - 대부분 offline 상태에서 학습
  - takes time, and needs RAM to store the data

- Online (mini-batch) Learning
  - 데이터를 조금씩 순차적으로 이용해 학습
  - each learning step is fast & cheap and can be done on the fly
  - 초기 학습은 보통 offline 상태에서 (즉, not on the live system)

Parametric Model (Linear models, LDA, Naïve Bayes, 단순한 신경망)

- f(모델 형태)를 가정하고, 모델의 특성을 규정하는 parameter (매개변수, β_i) 들을 학습을 통해 추정하는 형태

$$Y = f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- 위의 경우, 모델 원형 f 를 feature들 $X_1, \ldots X_p$ 로 된 ***선형식***으로 가정

- linear parametric 모델은 parameter β_i ($i=0, \ldots p$) 로 규정됨

- 위의 linear parametric 모델에서 $\hat{f}$ 를 추정함은 $\beta_0, \ldots \beta_p$ 들의 추정치 $\hat{\beta}_0 \ldots \hat{\beta}_p$ 들을 구하는 것

- 따라서, parametric model의 학습은 training set를 이용해 $\hat{\beta}_i$ 들을 잘 구해 $\hat{f}$(X) = Ŷ의 값이 실제 Y에 최대한 가깝도록 함이 목표

Non-Parametric Model

- **f의 형태(구조)가 어떨 것이라는 것을 미리 가정하지 않음**
- (More) Data Driven
- Training set의 크기가 작을 때 문제가 될 수 있음
- KNN, Decision Tree, Boosting/Random Forest, Support Vector Machine, 복잡한 신경망

Remember,

All the models are wrong!

우리 모델은 그 무슨 수학 원리를, 우주 법칙을 말하는 것이 아니고

Most models perform lousy if you got insufficient data

* 모델의 강력함(flexibility/capacity) 에 따른 prediction - inference, 일반화(generalize)의 중요함, training 성능 – test 성능, bias – variance, underfit – overfit, training 시간 들을 2장에서 얘기합니다. 머신러닝의 중요한 개념이라 정확히 알아야 합니다. 하지만, 이 개념을 **교조적으로 이해하면 안됩니다**.

Vector-Matrix 방정식 복습 – ISLR 2.1.2 에서의 표현식

n : 입력 data points (instance)들의 개수. 이것들을 이용해 모델 f 를 훈련

X : 위 입력 데이터들로 이루어진 predictor 집합.

x_i : X 내의 개별 입력 데이터 포인트 중 i번째 data point 값. (i=1, … n)

x_{ij} : i번째 입력 데이터 포인트의 j번째 predictor 값. X의 data point가 여러 개의 feature 들로 되어 있을 때 (보통 X는 여러 feature들로 됨). 이 책에서는 feature들의 개수를 p라 일반적으로 놓음 (즉, j = 1, . . . p)

* ISLR에서 i번째 data point(instance)를 가르킬때는 x_i, (i=1, … n)

i번째 feature set 를 가르킬때는 X_j 로 표시함, (j=1, … p)

* 어떤 책에서는 i번째 data point(instance)를 가르킬때 $x^{(i)}$ 으로 표기. 이 경우 3번째 instance의 8번 feature를 $x^{(3)}_8$ 로 표현

Y : X에 대응되는 response들의 집합

y_i : Y 내의 개별 입력 데이터 포인트 중 i번째 data point 값. (i=1, … n). 즉, 입력이 x_i 일 때의 response.

y_{ij} : i번째 response data point의 j번째 response값 (Response가 행렬로 표현). 이 경우는 개별 response data point가 여러 개의 response 값들로 이루어진 vector 형태일 때 적용된다 (Multioutput). ***ISLR에서는 이 경우를 다루지 않음***

Note

ISLR (Introduction to Statistical Learning with R) 과
ESL (The Elements of Statistical Learning) 책에서

$$\hat{Y} = X^T \hat{\beta}, \qquad (2.2)$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

두 형태의 표현식이 보인다. 이 경우 윗 식의 X^T는 $X^T = (X_1, X_2, \ldots, X_p)$ 로써 p개의 predictor로 구성된 하나의 observation을 뜻하고 row vector임을 나타냄. 이 X^T가 만약 i번째 observation이라면 ISLR/ESL은 x_i^T 로도 표현.

- Exception : 어떤 경우 X_j가 n개 observation들의 j번째 predictor들로 표현되기도 한다. 이 경우 X_j는 vector (of length n)가 되겠고 $\boldsymbol{X}^T$는 matrix가 됨. Predictor X_j가 scalar인지 또는 n observation 개수 만큼의 vector_of_length n 인지는 문맥을 통해 알 수 있다. 또 경우에 따라서 대문자 X는 vector를, 소문자 x는 scalar를 나타내기도 함.

반면 아랫 식의 bold **X**는 matrix(행렬)를 나타냄. 만약 predictor data set이 n개의 observation, 각 observation이 p개의 predictor들로 되어있으면 **X**는 "n x p" matrix.

기본적으로 ISLR/ESL은 벡터가 Column Vector라 가정. 따라서 어떤 벡터 x가 row vector일 시 그것을 x^T 로 표시.

Linear System의 Vector-Matrix 방정식 의미:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p. \qquad (2.4)$$

Linear model로 가정하고,

⇩ fit to given (X, Y) training set

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

ie. if Y can be approximated as a linear combination of $X_1, X_2, \ldots, X_p$

⇩ expanding…

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_p x_{1p} \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \cdots + \beta_p x_{2p} \\
y_3 &= \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \cdots + \beta_p x_{3p} \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_p x_{np}
\end{aligned}$$

* β_i : 우리가 구하려는 패러미터

⇩

$$\begin{pmatrix} Y \\ y_1 \\ y_2 \\ y_3 \\ \\ \\ \\ y_n \end{pmatrix} = \beta_0 + \beta_1 \begin{pmatrix} X_1 \\ x_{11} \\ x_{21} \\ x_{31} \\ \\ \\ \\ x_{n1} \end{pmatrix} + \beta_2 \begin{pmatrix} X_2 \\ x_{12} \\ x_{22} \\ x_{32} \\ \\ \\ \\ x_{n2} \end{pmatrix} + \beta_3 \begin{pmatrix} X_3 \\ x_{13} \\ x_{23} \\ x_{33} \\ \\ \\ \\ x_{n3} \end{pmatrix} + \; . \; . + \beta_p \begin{pmatrix} X_p \\ x_{1p} \\ x_{2p} \\ x_{3p} \\ \\ \\ \\ x_{np} \end{pmatrix}$$

$$\underset{Y}{Y - \beta_0} = \underset{X}{\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & & x_{2p} \\ x_{31} & x_{32} & x_{33} & & x_{3p} \\ \vdots & & \ddots & & \\ x_{n1} & x_{n2} & x_{n3} & & x_{np} \end{pmatrix}} \underset{\beta}{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \\ \\ \\ \beta_p \end{pmatrix}}$$

① *(X, Y)는 Training set 데이터*

② *RSS를 최소화하는 $\hat{\beta}$들을 구하자 (linear regression의 parameter를 구하자)*

③ *그리고는 새로운 X가 주어지면 $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ 하여 $\hat{Y}$ 추정*