Foundation of Machine Learning 9주차

정재헌, 우지수 / 2023.03.16

Computational Data Science LAB

CONTENTS

- $f_\beta - score$

$$F_\beta = (1+\beta^2) * \frac{precision \ * recall}{(\beta^2 * precision) + recall}$$

✓ Precision과 Recall의 조화평균 값

✓ β값이 1보다 크면 Recall의 비중이 커짐 (Ex. $\beta = 2$)

- Recall : 실제로 positive인 샘플 중에 모델이 positive라고 예측한 것
- Precision : positive라고 예측한 것 중에 실제로 positive 인 것

$F_2 = (1+4) * \frac{precision * recall}{(4*precision)+recall}$ → positive를 더 잘 찾아내는 모델 ($\frac{TP}{TP+FP}$)

✓ β값이 1보다 작으면 Precision 비중이 커짐 (Ex. $\beta = 0.25$)

$F_{0.5} = (1+0.25) * \frac{precision*recall}{(0.25*precision)+recall}$ → negative를 더 잘 찾아내는 모델 ($\frac{TP}{TP+FN}$)

00 Introduction

Background

✓ 관측치 (데이터) 를 잘 나타내는 모델을 찾아 예측을 하는 것이 목표

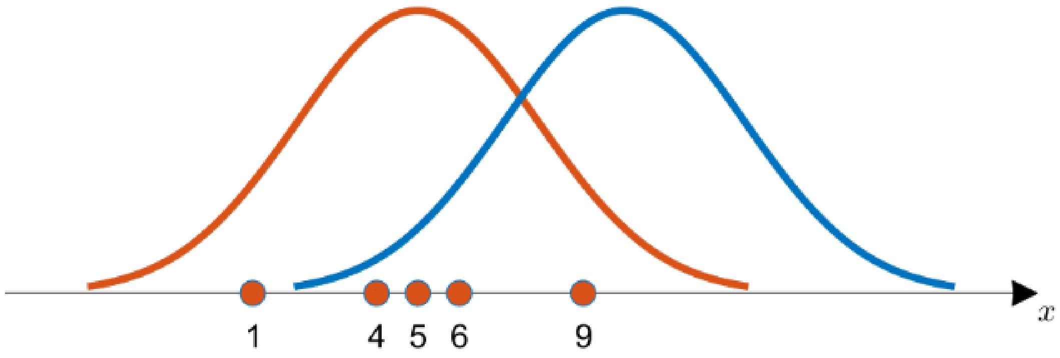
↓ 통계학적인 관점

✓ 모집단을 알기 위해 표본을 추출한 뒤, 이를 바탕으로 모집단의 분포를 잘 나타내는 모수 (파라미터) 를 추정

✓ 모집단에서 새로운 관측치 (데이터) 가 발생할 확률 예측 가능

Background

$$D = \{1, 4, 5, 6, 9\}$$



- ✓ 두 분포 중 위 표본 관측치는 어떤 분포로부터 추출되었을 확률이 더 높을까?

정규분포

- ✓ 우리는 관측치를 관찰함으로써 해당 관측치가 추출되었을 것으로 생각되는 모수 추정 가능 → HOW ?

Ex) 평균

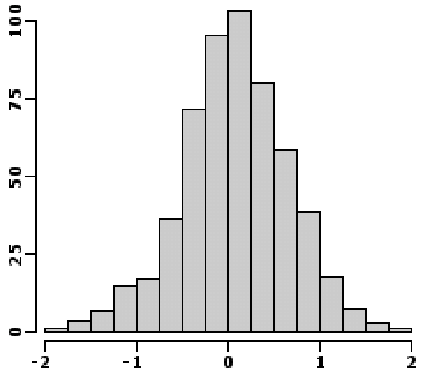
- Probability Distribution (확률분포)

  ✓ 확률변수가 취할 수 있는 값들과 그 값들이 나타날 확률의 분포

  어떤 확률적인 현상에서 발생하는 값들을 수학적으로 표현한 변수

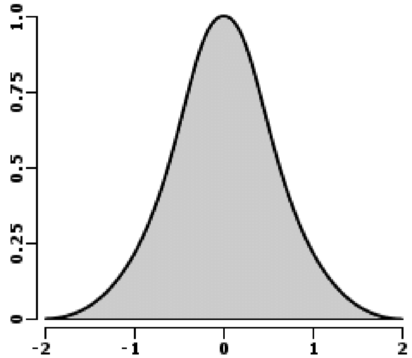
Type

1. Discrete Probability Distribution (이산확률분포)



- ✓ 확률변수가 취할 수 있는 값들이 이산적인 경우의 확률분포
- ✓ Ex) 베르누이 분포, 포아송 분포
- ✓ 확률질량함수 (Probability Mass Function, PMF) 를 사용하여 정의

2. Continuous Probability Distribution (연속확률분포)



- ✓ 확률변수가 취할 수 있는 값들이 연속적인 경우의 확률분포
- ✓ Ex) 정규분포, 지수분포, 감마분포
- ✓ 확률밀도함수 (Probability Density Function, PDF) 를 사용하여 정의

- Probability Distribution (확률분포)

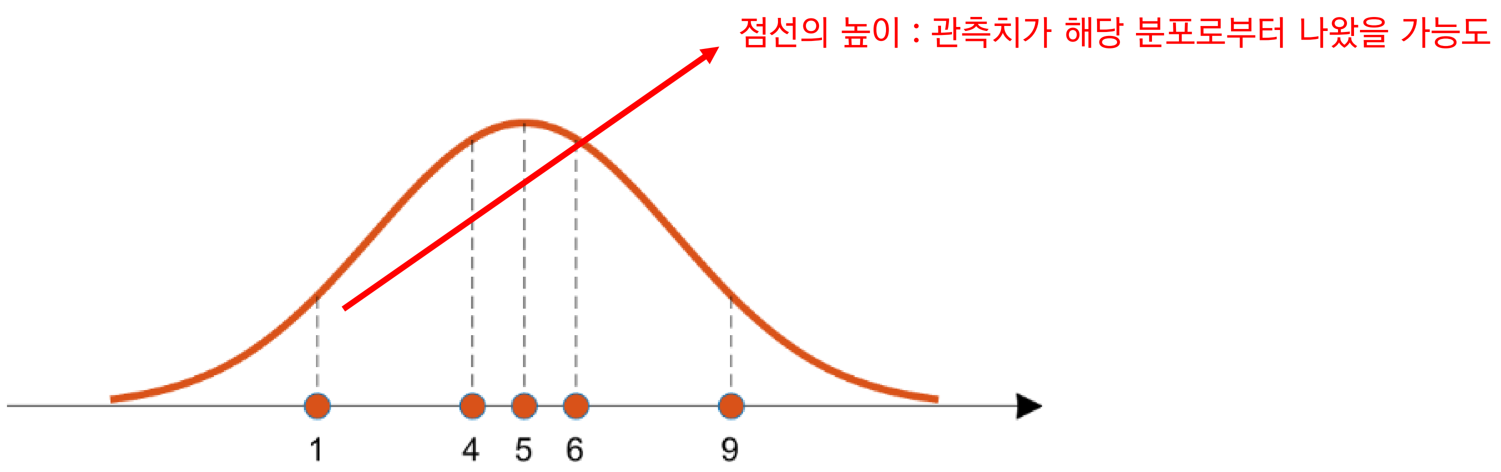
✓ 확률변수가 취할 수 있는 값들과 그 값들이 나타날 확률의 분포

어떤 확률적인 현상에서 발생하는 값들을 수학적으로 표현한 변수

✓ 관측치 (데이터) 에 대한 함수
✓ 우리가 가지고 있는 표본 관측치 (데이터) 를 바탕으로 모집단의 확률분포를 파악하고 예측 수행 가능
✓ 이때, 모집단의 분포를 잘 나타내는 모수를 추정하는 과정이 필요

02 | Likelihood Function
Definition

- Likelihood (가능도)



Definition

- Likelihood function (가능도 함수)

✓ $D = \{y_1, \cdots, y_n\}$: 표본 (관측치)

✓ $p(y)$: 표본 (관측치) 의 확률분포

✓ θ : 모수 (파라미터)

무한 곱에 log함수를 취함으로써 summation로 만들어 미분 계산 용이

✓ $p(D;\hat{\theta}) = \underline{\prod_{i=1}^{n} p(y_i;\hat{\theta})}$ → $$logp(D;\hat{\theta}) = \sum_{i=1}^{n} logp(y_i;\theta)$$

각 표본 관측치가 해당 분포로부터 나왔을 가능도 (높이) 를 계산하여 곱함

= 관측치 추출은 독립적으로 연달아 일어나기 때문

02 | Likelihood Function

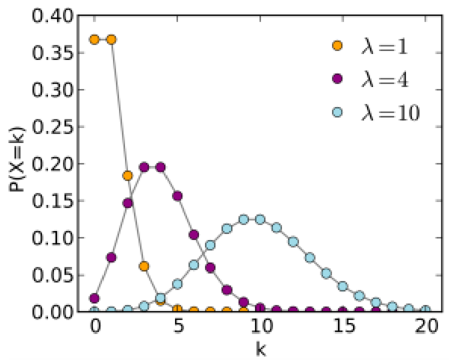
Definition

- Likelihood function (가능도 함수)

  ✓ 모수 (파라미터) 에 대한 함수

  ✓ 확률분포의 모수 (파라미터) 가 어떤 값일 때 가장 적합한지 결정 가능

Ex)

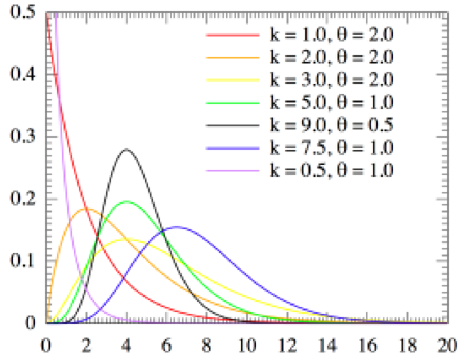


[포아송 분포]

$$p(k;\lambda) = \lambda^k e^{-\lambda}/(k!)$$

[베타 분포]

$$p(y;\alpha,\beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}$$



[감마 분포]

$$p(y;k,\theta) = \frac{1}{\gamma(k)\theta^k}x^{k-1}e^{-y/\theta}$$

- Maximum Likelihood Estimation (최대우도 추정)

$$\hat{\theta} = \underset{\theta\in\Theta}{argmax}\, logp(D, \hat{\theta})$$
$$= \underset{\theta\in\Theta}{argmax} \sum_{i=1}^{n} logp(y_i; \theta)$$

최적화

$\xrightarrow{\text{미분}}$ Closed form for the MLE

✓ 표본 관측치 (데이터) 를 가장 잘 설명하는 확률분포와 모수를 찾는 방법

How to find ?

- Maximum Likelihood Estimation (최대우도 추정)

✓ Ex) 포아송 분포의 MLE 구하기

- $p(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

- $\log[p(k;\lambda) = \log\left[\frac{\lambda^k e^{-\lambda}}{k!}\right] = klog\lambda - \lambda - \log(k!)$

- $logp(D,\lambda) = \sum_{i=1}^{k}[k_i log\lambda - \lambda - \log(k_i!)]$

- $0 = \frac{\partial}{\partial\lambda}[logp(D,\lambda)] = \sum_{i=1}^{n}\left[\frac{k_i}{\lambda} - 1\right] \Rightarrow \lambda = \frac{1}{n}\sum_{i=1}^{n} k_i$ → 포아송 분포의 추정 모수 $\hat{\lambda}$는 관측치 k의 평균

Compare to Linear Probabilistic Models

- GLM vs LPM

LPM	GLM
✓ 종속변수와 독립변수 사이의 관계를 설명하기 위해 선형함수 사용	✓ 종속변수와 독립변수 사이의 관계를 설명하기 위해 Transfer Function을 사용 ex) 베르누이 분포일 경우 로지스틱 함수를 Transfer Function으로 사용
✓ 종속변수가 연속형 일 경우에만 적용되며, 정규분포를 따른다고 가정	✓ 종속변수가 이산형, 연속형 일 수도 있음 ex) 베르누이 분포, 포아송 분포, 감마 분포 등

1. Bernoulli Regression (베르누이 회귀)

✓ Binary Classification (이진분류) 문제에서 사용
✓ 종속변수 Y가 이진형 (0또는 1) 일 때, 독립변수 X와의 관계를 모델링
✓ 각 x에 대해 $y = \{0,1\}$ 값을 가질 확률을 예측 → $p(y) =$ 베르누이 분포

$$\begin{array}{ccccc} x & \rightarrow & w^T x & \rightarrow & f(w^T x) = \theta = p(y=1|x) \\ \in R^d & & \in R & & \in [0,1] \end{array}$$

Linear Predictor   Transfer Function

1. Bernoulli Regression (베르누이 회귀)

$$\underbrace{p_w(D)}_{\text{베르누이 분포}} = \prod_{i=1}^{n} p_w(y_i|x_i) = \prod_{i=1}^{n} [\underbrace{f(w^T x_i)}_{\text{로지스틱 함수}}]^{y_i} [1 - f(w^T x_i)]^{1-y_i}$$

$$logp_w(D) = \sum_{i=1}^{n} y_i logf(w^T x_i) + (1 - y_i)\log[1 - f(w^T x_i)]$$

- ✓ 로지스틱 함수 : 0과 1 사이의 값을 출력하며, 종속변수의 값이 1일 확률 도출
- ✓ 모델링을 위해 독립변수와 Transfer Function을 결정한 후, 최대 우도 추정 (MLE) 을 사용하여 모델을 학습
- ✓ MLE를 통해 관측된 데이터를 가장 잘 설명하는 파라미터 (w) 를 찾을 수 있음.

1. Bernoulli Regression (베르누이 회귀)

$$p_w(D) = \prod_{i=1}^{n} p_w(y_i|x_i) = \prod_{i=1}^{n}[f(w^T x_i)]^{y_i}[1-f(w^T x_i)]^{1-y_i}$$

$$logp_w(D) = \sum_{i=1}^{n} y_i logf(w^T x_i) + (1-y_i)\log[1-f(w^T x_i)]$$

$$J(w) = -[\sum_{i=1}^{n} y_i logf(w^T x_i) + (1-y_i)\log[1-f(w^T x_i)]]$$

Maimizing $logp_w(D)$ 문제를

Minimizing log-likelihood 목적함수로 변경

미분

2. Possion Regression (포아송 회귀)

- ✓ 종속변수 Y가 양의 정수 일 때, 독립변수 X와의 관계를 모델링
- ✓ 각 x에 대해 $y = \{0, \infty\}$ 값을 가질 확률을 예측 → $p(y)$ = 포아송 분포

$$x \rightarrow w^T x \rightarrow f(w^T x) = \lambda$$

$$\in R^d \qquad \underline{\in R} \qquad \underline{\in [0, \infty]}$$

Linear Predictor   Transfer Function

2. Possion Regression (포아송 회귀)

지수 함수

$$\underbrace{p_w(D)}_{\text{포아송 분포}} = \prod_{i=1}^{n} p_w(y_i|x_i) = \prod_{i=1}^{n} \frac{e^{-f(w^T x_i)} \overbrace{f(w^T x_i)^{y_i}}^{\text{지수 함수}}}{y_i!}$$

포아송 분포

$$logp_w(D) = \sum_{i=1}^{n} [y_i logw^T x_i - f(w^T x_i) - \log(y_i!)]$$

- ✓ 지수 함수 : 0과 ∞ 사이의 값을 출력하며, 종속변수의 값이 양의 정수일 확률 도출
- ✓ 모델링을 위해 독립변수와 Transfer Function을 결정한 후, 최대 우도 추정 (MLE) 을 사용하여 모델을 학습
- ✓ MLE를 통해 관측된 데이터를 가장 잘 설명하는 파라미터 (w) 를 찾을 수 있음.

3. Conditional Gaussian Regression (조건부 가우시안 회귀)

- ✓ 종속변수 Y가 실수 (연속형) 일 때, 독립변수 X와의 관계를 모델링
- ✓ 각 x에 대해 $y = R$ 값을 가질 확률을 예측 → $p(y) =$ 가우시안 분포

x	$\rightarrow$	$w^T x$	$\rightarrow$	$f(w^T x) = \mu$
$\in R^d$		$\in R$		$\in R$
		Linear Predictor		Transfer Function

3. Conditional Gaussian Regression (조건부 가우시안 회귀)

선형 함수

$$\underbrace{p_w(D)}_{\text{가우시안 분포}} = \prod_{i=1}^{n} p_w(y_i|x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y_i - \overline{w^T x_i})^2}{2\sigma^2}\right)$$

$$log p_w(D) = \sum_{i=1}^{n} log\left[\frac{1}{\sigma\sqrt{2\pi}}\right] + \sum_{i=1}^{n}\left(\frac{-(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

- ✓ 모델링을 위해 독립변수와 Transfer Function을 결정한 후, 최대 우도 추정 (MLE) 을 사용하여 모델을 학습
- ✓ MLE를 통해 관측된 데이터를 가장 잘 설명하는 파라미터 (w) 를 찾을 수 있음.

Q&A

감사합니다.