Foundation of Machine Learning 10주차

정재헌, 우지수 / 2023.03.22

Computational Data Science LAB

CONTENTS

00 Introduction
Background

- Bayesian Probability

Frequentist Probability vs Bayesian Probability

✓ 동전의 앞면이 나올 확률은 우리가 직접 던져 봄으로써 구할 수 있음.

✓ 시행횟수를 반복하여 빈도수를 측정

✓ 화산이 폭발할 확률은 직접 실험해볼 수 없음.

✓ 하지만 이러한 사건과 관련 있는 여러 확률을 이용해 새롭게 일어날 사건을 추정 가능

- Bayesian Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

✓ 관측치가 주어졌을 때의 조건부 확률을 구하는 공식

✓ 관측치가 주어지기 전의 확률 값이 관측치가 주어지면서 어떻게 변하는지 계산 가능

- Bayesian Theorem

[증명]

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \rightarrow P(A \cap B) = P(A|B)P(B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \rightarrow P(A \cap B) = P(B|A)P(A)$$

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

가능도

$$\boxed{P(A|B)} = \frac{\boxed{P(B|A)}\boxed{P(A)}}{\boxed{P(B)}}$$

사후확률

사전확률

정규화 상수 = 모든 가능한 사건에 대한 확률의 합이 1이 되도록 정규화 해야 하는 확률 기본 원리

- $P(A)$: 사전확률, 사건 B가 발생하기 전의 사건 A의 확률
- $P(B)$: 정규화 상수
- $P(B|A)$: 가능도, 사건 A가 발생한 경우 사건 B의 확률
- $P(A|B)$: 사후확률, 사건 B가 발생한 후 갱신된 사건 A의 확률

- Bayesian Theorem

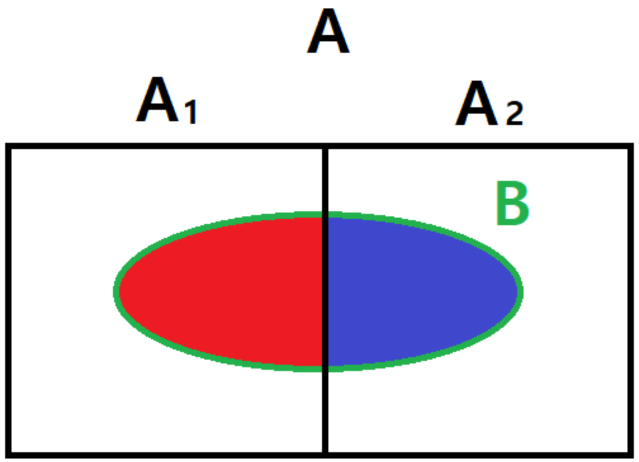
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

✓ 사전확률과 사후확률 사이의 관계를 나타내는 정리
✓ 사건 B가 발생함으로써 사건 A의 확률이 어떻게 변화하는지를 표현한 정리
✓ 새로운 정보가 기존의 추론에 어떻게 영향을 미치는지 나타내는 정리

✓ 관측치가 주어지기 전에 이미 어느정도 확률 값을 알고 있을 때, 이를 새로 수집한 관측치를 바탕으로 예측

Ex) 데이터를 매일 추가적으로 얻는 상황에서도 매일 전체 데이터를 대상으로 새로운 분석 작업을 할 필요없이 어제 분석결과에 오늘 들어온 데이터를 합쳐 업데이트만 하면 됨.

- Bayesian Theorem

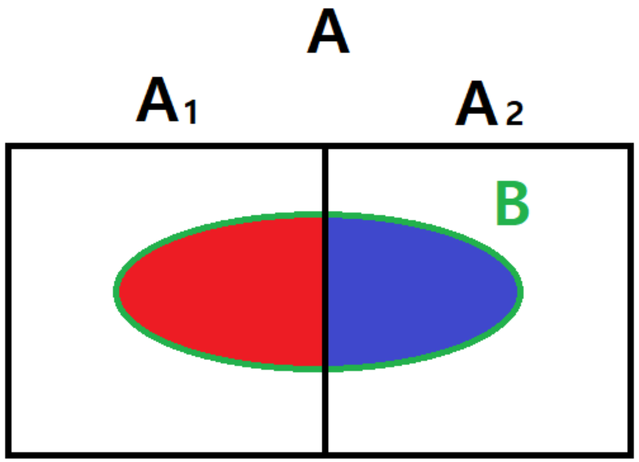


✓ 조건 1) A_1과 A_2가 서로 배타적 (교집합이 없음.)

✓ 조건 2) A_1, A_2가 서로 완전 (합집합이 표본공간)

Example

- Bayesian Theorem



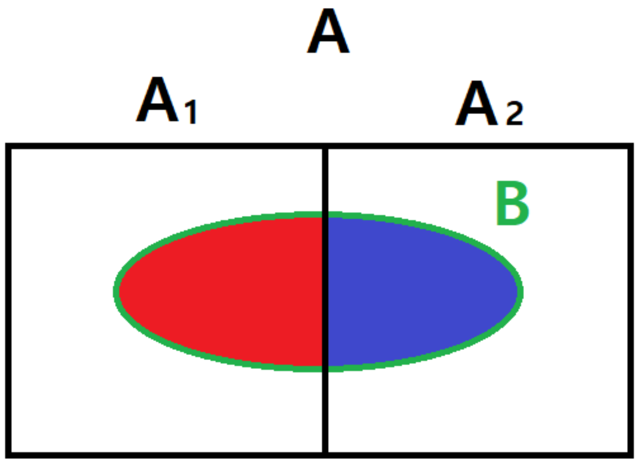
각각을 조건부 확률로 표현

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1 \cap B)}{\underline{P(A_1 \cap B) + P(A_2 \cap B)}} = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)}$$

전체를 A로 쪼갤 수 있으므로 B도 A를 활용하여 쪼개짐.

배반 사건이므로 확률의 합으로 표현함.

- Bayesian Theorem



$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1 \cap B)}{P(A_1 \cap B) + P(A_2 \cap B)} = \frac{P(B|A_1)P(A_1)}{\underline{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)}}$$

i값이 바뀌어도 항상 같은 값이므로 분자의 값만 비교

✓ Multi-Class Classification 문제에서 Bayesian Theorem이 어떻게 사용되는지를 보여주는 수식

✓ 여러 배타적이고 완전한 사건 중에서 가장 확률이 높은 하나의 사건을 고르는 문제

- Bayesian Methods

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

✓ Bayesian Theorem을 기반으로 한 통계적 분석 방법론

✓ 사전 확률 분포를 사용하여 사후 확률 분포를 추론하는 방법

✓ 새로운 데이터가 추가되면 이를 토대로 사전 확률 분포를 갱신하여 사후 확률 분포를 계산한 후, 최종적으로 모수 추론

- Bayesian Methods

$$\underbrace{\boxed{P(\theta|\mathcal{D})}}_{\text{사후 확률 분포}} = \frac{\overbrace{\boxed{P(\mathcal{D}|\theta)}}^{\text{우도}}\overbrace{\boxed{P(\theta)}}^{\text{사전 확률 분포}}}{P(\mathcal{D})}$$

- θ = 모델의 파라미터
- $\mathcal{D}$ = 주어진 관측치

- Prior Distribution (사전 확률 분포)
  - ✓ 파라미터 θ의 사전 확률 분포

  Ex) 어떠한 질병의 발병률을 알기 전에 이미 해당 질병 검사에서
  양성을 보인 사람들의 비율 등과 같은 사전 정보가 존재

- Posterior Distribution (사후 확률 분포)
  - ✓ 관측치 $\mathcal{D}$가 주어졌을 때, 파라미터 θ의 사후 확률 분포
- Likelihood (우도)
  - ✓ 관측치 $\mathcal{D}$가 주어졌을 때 파라미터 θ가 얼마나 적합한지를 나타내는 함수

- Bayesian Methods

[동전 던지기]

$P(Heads|\theta) = \theta$

$\theta \in \Theta = (0,1)$ → θ에 대한 사전 확률 분포인 $P(\theta)$ 필요

$$\theta \sim Beta(\alpha, \beta)$$

$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Bayesian Methods

[동전 던지기]

$P(Heads|\theta) = \theta$

$\theta \in \Theta = (0,1)$ ➡ θ에 대한 사전 확률 분포인 $P(\theta)$ 필요

$$\theta \sim Beta(h, t)$$

$$P(\theta) \propto \theta^{h-1}(1-\theta)^{t-1}$$

$$P(\mathcal{D}|\theta) = \theta^{n_h}(1-\theta)^{n_t}$$

- Bayesian Methods

[동전 던지기]

$P(Heads|\theta) = \theta$

$\theta \in \Theta = (0,1)$ → θ에 대한 사전 확률 분포인 $P(\theta)$ 필요 → 베이즈 정리를 통해 사후 확률 분포 추정

$$P(\theta|\mathcal{D}) \propto \theta^{h-1}(1-\theta)^{t-1} \times \theta^{n_h}(1-\theta)^{n_t}$$

$$= \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}$$

$$\theta|\mathcal{D} \sim Beta(h+n_h, t+n_t)$$

사전 확률 분포 = 사후 확률 분포

- Conjugate Prior Distribution (켤레 분포)

  ✓ 사후 확률 분포가 사전 확률 분포와 같은 분포 계열에 속할 때의 사전 확률 분포
  ✓ 사전 확률 분포의 파라미터만 업데이트 하는 방식으로 사후 확률 분포를 계산할 수 있게 되어 계산이 간편해짐.

$$\theta \sim Beta(h, t)$$

$$\boxed{P(\theta) \propto \theta^{h-1}(1-\theta)^{t-1}}$$

사전 확률 분포 = 베타 분포 (0과 1 사이의 값)

$$P(\mathcal{D}|\theta) = \theta^{n_h}(1-\theta)^{n_t}$$

$$P(\theta|\mathcal{D}) \propto \theta^{h-1}(1-\theta)^{t-1} \times \underline{\theta^{n_h}(1-\theta)^{n_t}}$$

베르누이 분포 (0과 1 사이의 값)

$$= \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}$$

사후 확률 분포 = 베타 분포

$$\theta|\mathcal{D} \sim Beta(h + n_h, t + n_t)$$

03 | Conjugate Prior Distribution
Definition

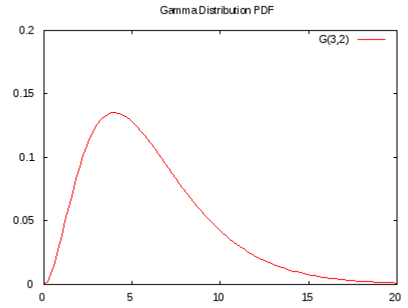
- Conjugate Prior Distribution (켤레 분포)

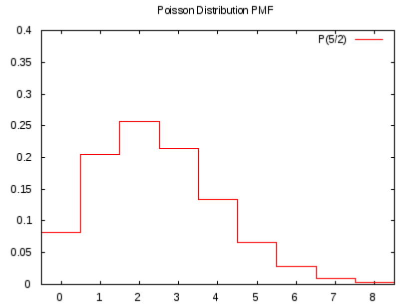
✓ 사전 확률 분포와 가능도 함수의 곱이 다시 해당 분포와 같은 분포를 따르는 것

✓ 사전 확률 분포와 가능도 함수가 비슷한 형태를 가지기 때문

Ex 1) 감마 분포

- 사전 확률 분포 : 감마분포
- 가능도 함수 : 포아송 분포

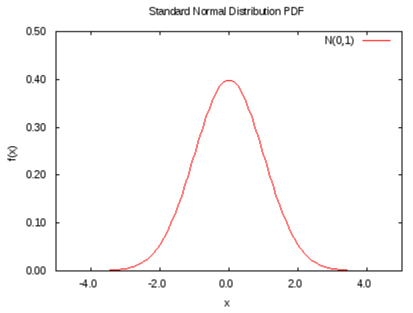




Ex 2) 정규 분포

사전 확률 분포 : 정규 분포

가능도 함수 : 정규 분포



Definition

- Maximum a Posteriori Estimate

$$argmax\ P(\theta|\mathcal{D}) = argmax\ P(\mathcal{D}|\theta) \times P(\theta)$$

✓ 사후 확률 분포에서 가장 높은 값을 가지는 모수(파라미터 θ)을 선택하는 방법

✓ MLE와 달리 모수에 대한 불확실성을 나타내는 확률 분포인 사전 확률 분포를 활용

✓ 모수에 대한 사전 지식을 고려하여 가능도와 사전 확률 분포의 곱의 최대값을 갖는 모수(θ)를 추정

- Bayesian Decision Theory

✓ 불확실성이 존재하는 환경에서 Bayesian Methods를 활용하여 최적의 의사결정을 내리는 방법론

1. Loss Function (의사결정에 따라 발생할 수 있는 비용) 정하기

✓ Posterior Risk : 주어진 관측치($\mathcal{D}$)와 의사결정(a)에 따른 평균적인 손실을 의미

$$r(a) = \mathbb{E}[\ell(\theta, a) \mid \mathcal{D}] = \int \underbrace{\ell(\theta, a)}_{\text{Loss Function}} \overbrace{P(\theta \mid \mathcal{D})}^{\text{사후 확률 분포}} d\theta$$

2. Posterior Risk를 최소화하는 의사결정 찾기

$$r(a^*) = \min_{a \in \mathcal{A}} r(a)$$

.

- Bayesian Decision Theory

Ex) Square Loss ($\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$)

$$r(a) = \mathbb{E}[\ell(\theta, a)\,|\mathcal{D}] = \int \ell(\theta, a)P(\theta|\mathcal{D})d\theta$$

$$r(\hat{\theta}) = \int (\hat{\theta} - \theta)^2 P(\theta|\mathcal{D})d\theta$$

$$\frac{dr(\hat{\theta})}{d\hat{\theta}} = -2\int \theta P(\theta|\mathcal{D})d\theta + 2\hat{\theta}$$

$$\boxed{\hat{\theta} = \int \theta P(\theta|\mathcal{D})d\theta = \mathbb{E}[\theta|\mathcal{D}]}$$

Square Loss를 최소화하는 최적의 의사결정 θ는
사후 평균 값

Bayesian Methods Application 2

- Bayesian Regression

✓ Bayesian Methods를 활용하여 회귀 분석을 수행하는 방법론

[회귀분석 vs 베이지안 회귀 분석]

회귀 분석 : 관측치를 이용하여 모델의 파라미터 값을 추정

베이지안 회귀 분석 : 사전 확률 분포를 이용하여 파라미터 값의 분포를 설정하고 관측치를 이용하여 사후 확률 분포를 계산함으로써 추정

이때, 불확실성도 함께 고려하므로 회귀분석보다 더욱 유연하고 신뢰성 있는 모델링 가능

- Bayesian Regression

Ex) Gaussian Linear Regression

1. 모델의 파라미터에 대한 사전 확률 분포를 정의

$w \sim \mathcal{N}(0, \Sigma_0)$ = 정규분포로 가정

2. 관측 값($\mathcal{D}$)을 통해 파라미터의 사후 확률 분포를 계산

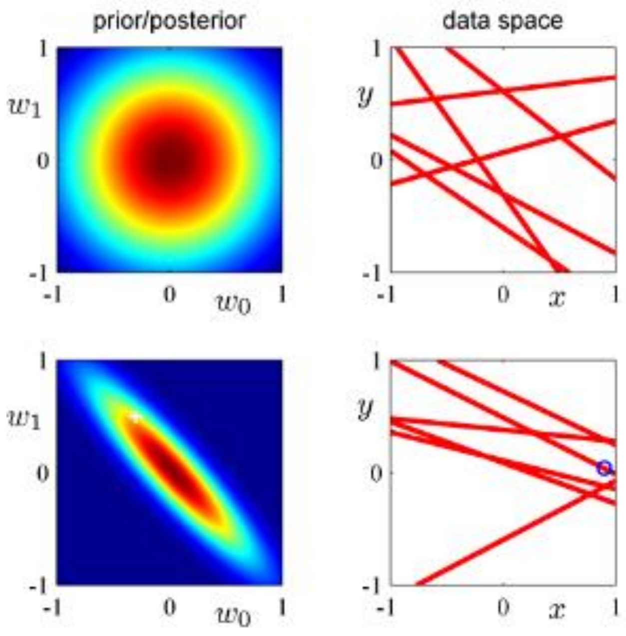
$p(w|\mathcal{D}) \propto L_{\mathcal{D}}(w)p(w)$

사후 확률 분포에서 가장 높은 확률을 가지는 파라미터 값을 선택하여 추정 값으로 사용 (MAP)

$$= \underbrace{\prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)\right]}_{\text{Likelihood}} \times \underbrace{|2\pi\Sigma_0|^{-1/2}\exp\left(-\frac{1}{2}w^T \sum w_0^{-1}\right)}_{\text{Prior Distribution}}$$

- Bayesian Regression

Ex) Gaussian Linear Regression (실제 사후 확률 분포를 이용하여 얻어지는 결과)



✓ 사전 확률 분포로부터 랜덤 생성된 것 (평균 0, 공분산 1)

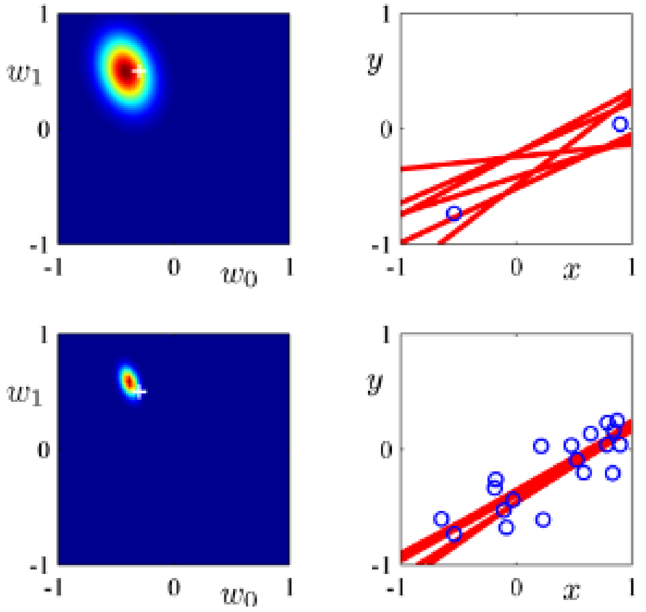
✓ w_0, w_1가 가우시안 분포에 의해 랜덤하게 생성되어 무작위적으로 직선 형성

✓ 가능도 함수를 통해 w의 확률을 결정한 후 사후 확률 분포 구하기

✓ 이때, 사후 확률 분포와 사전 확률 분포의 분포가 가우시안 분포로 같으므로
새로운 관측치가 업데이트 되더라도 사후 확률 분포를 사전 확률 분포로 활용 가능

- Bayesian Regression

Ex) Gaussian Linear Regression (실제 사후 확률 분포를 이용하여 얻어지는 결과)



✓ 관측치가 추가될 수록 사후 확률 분포는 갱신되어 특정 값으로 근사화

모수의 추정치

Q&A

감사합니다.