9.3 베이즈 추정법

베이즈 추정법(Bayesian estimation)은 모수값이 가질 수 있는 모든 가능성의 분포를 계산하는 작업이다.

어떤 확률분포함수의 모수를 μ라고 할 때,

— 최대가능도 추정법 : μ = 상수

— 베이즈 추정법 : μ = 확률변수 (확률밀도함수를 가지는)

⇒ 어떤 값이 가능성 높고 낮은지 살펴보겠다는 의미

* 베이즈 추정법을 사용하는 의미/이유?

"추정된 모수값 하나만으로는 추정의 신뢰도나 신뢰구간을 구할 수 없기 때문."

ex) 인터넷 쇼핑몰 상품 A, B에 대한 사용자 반응

- 상품 A) 전체 리뷰 3개, 좋아요 2개, 싫어요 1개
- 상품 B) 전체 리뷰 100개, 좋아요 60개, 싫어요 40개

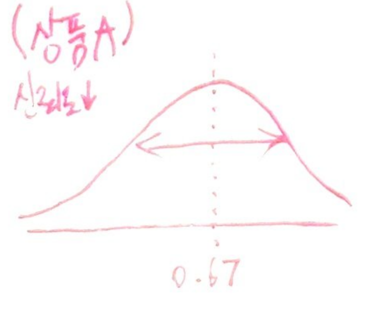
→ 각 상품을 사용했을 때 '좋아요' or '싫어요'가 나오는 사건은 베르누이분포 모형화 할 수 있다.

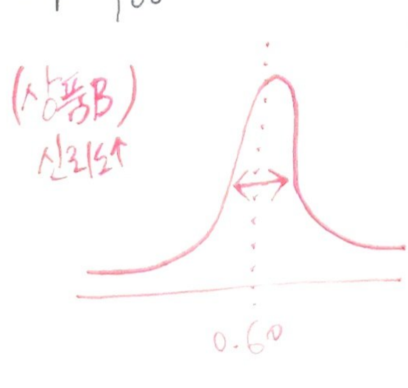
최대가능도 추정법에 따르면, 상품 A와 상품 B에서 '좋아요'가 나온 비율을 사용하여 베르누이 모수를 구하면 다음과 같이 상품 A의 모수가 높다.

- 상품 A의 모수 $\frac{2}{3} = 0.67$
- 상품 B의 모수 $\frac{60}{100} = 0.60$

} 상품 A가 상품 B보다 평가가 좋다고 확신할 수 있을까?

⇒ 베이즈 추정법 필요!





베이즈 추정법의 기본 원리

수학적으로 베이즈 추정법은 주어진 데이터 $\{x_1, \ldots, x_N\}$를 기반으로 모수 μ의 조건부 확률분포 $p(\mu | x_1, \ldots, x_N)$을 계산하는 작업이다.

조건부 확률분포를 구하려면 베이즈 정리를 사용한다.

$$p(\mu | x_1, \ldots, x_N) = \frac{p(x_1, \ldots, x_N | \mu) \cdot p(\mu)}{p(x_1, \ldots, x_N)} \propto p(x_1, \ldots, x_N | \mu) \cdot p(\mu)$$

- $\boxed{p(\mu)}$ = 모수의 사전 (prior) 분포

  베이즈 추정 이전에 알고 있던 모수 μ의 분포,
  모수 분포에 대해 아무런 지식이 없는 경우에는
  균일 (uniform) 분포 Beta (1, 1) 이나, 0을 중심으로 하는 정규분포
  $N(0, \sigma_0^2)$ 등의 무정보분포 (non-informative distribution)를
  사용할 수 있다. 무정보분포에 대해서는 다음 장에서 공부한다.

- $\boxed{p(\mu | x_1, \ldots, x_N)}$ 모수의 사후 (posterior) 분포

  데이터 $x_1, \ldots, x_N$이 주어진 상태에서의 μ에 대한 조건부 확률분포
  ⇒ 우리가 베이즈 추정법으로 구하고자 하는 분포.

- $\boxed{p(x_1, \ldots, x_N | \mu)}$는 가능도 (likelihood) 분포

  모수 μ가 특정한 값으로 주어졌을 때, 주어진 데이터 $\{x_1, \ldots, x_N\}$가 나올 수 있는 확률값을 나타낸다.

이 때 계산된 모수의 분포는 두 가지 방법으로 표현한다.

(1) 모수적 (parametric) 방법

- 다른 확률분포를 사용하여 추정된 모수의 분포를 나타낸다.
- 모수 분포를 표현하는 확률분포함수의 모수를 하이퍼모수 (hyper-parameter) 라고 부른다.
- 모수적 방법을 사용한 베이즈 추정법은 결국 하이퍼모수값을 계산하는 작업이다.

(2) 비모수적 (non-parametric) 방법

- 모수의 분포와 동일한 분포를 가지는 실제 표본 집합을 생성하여 히스토그램이나 최빈값 등으로 분포를 표현한다.
- MCMC (Markov chain Monte Carlo)와 같은 몬테카를로 방법이 비모수적 방법이다.

여기에서는 "모수적 방법의 몇가지 간단한 예를" 살펴본다.

베르누이 분포의 모수 추정

베르누이 분포의 모수 μ를 베이즈방법으로 추정해보자.

가장 단순한 이산확률변수인 베르누이 분포의 모수는 0부터 1 사이의 값이 되므로, 사전 분포는 하이퍼모수 $a=b=1$인 베타분포라고 가정하자.

$$p(\mu) \propto \mu^{a-1}(1-\mu)^{b-1} \quad (a=1,\ b=1)$$

정보없을 때 기준 (정보 있다면 다른 값)

데이터는 모두 독립적인 베르누이 분포의 곱이므로, 가능도함수는 다음과 같다.

$$p(x_1, \ldots, x_N \mid \mu) = \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i}$$

베이즈 정리를 사용하면, 사후분포가 다음처럼 갱신된 하이퍼모수 a', b'를 가지는 또다른 베타분포가 된다.

$$p(\mu \mid x_1, \ldots, x_N) \propto p(x_1, \ldots, x_N \mid \mu)\, p(\mu)$$

$$= \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} \cdot \mu^{a-1}(1-\mu)^{b-1}$$

$$= \mu^{\left(\sum_{i=1}^{N} x_i\right)+a-1} (1-\mu)^{\left(\sum_{i=1}^{N}(1-x_i)\right)+b-1}$$

$$= \mu^{N_1+a-1}(1-\mu)^{N_0+b-1}$$

$$= \mu^{a'-1}(1-\mu)^{b'-1}$$

이렇게 사전분포와 사후분포가 모수값만 다르고 함수형태가 같은 확률밀도 함수로 표현될 수 있도록 해주는 사전분포를 켤레 사전확률분포 (conjugate prior)라고 한다.

갱신된 하이퍼모수값은 다음과 같다.

$$a' = N_1 + a$$
$$b' = N_0 + b$$

$\boxed{\text{예제}}$

동전을 10번 던져서 앞면 6번 (뒷면 4번)
사전분포가 하이퍼모수 $a=b=1$인 베타분포라면,
사후분포는 다음과 같은 하이퍼모수를 가지는 베타분포가 된다.

$$a' = 6+1 = 7$$
$$b' = 4+1 = 5$$

* 최대가능도 추정법으로 구한 베르누이 모수의 값은 0.6

베이즈 추정법의 장점

"순차적 (sequential) 계산이 가능"

ex) 50개의 데이터를 매일 수집할 경우

전날의 사후확률 = 다음날의 사전확률

(베이즈 추정법)
- 첫날 50개 데이터로 모수 추정
- 다음날 추가된 데이터 50개로 모수값을 더 정확하게 추정
- 계산량은 매일 동일하며 증가하지 않음

1일 prior $= p(\mu)$
posterior $= p(\mu \mid D_1)$

2일 prior $= p(\mu) \triangleq p(\mu \mid D_1)$
posterior $= p(\mu \mid D_1, D_2)$

⋮

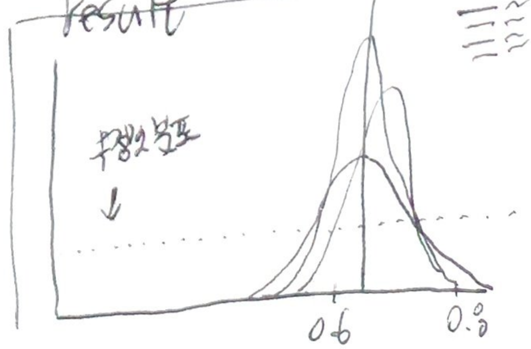
(최대가능도 추정법)
- 첫날 데이터 50개로 모수추정
- 다음날 데이터 100개로 모수추정
- 시간이 지날수록 계산량이 매일 증가함

python

```
a, b = 1, 1
print("초기 추정 : 모드 = 모름")
xx linsp xx = np.linspace(0, 1, 1000)
plt.plot(xx, sp.stats.beta(a, b).pdf(xx), ls=":", label="초기추정")
np.random.seed(0)
for i in range(3):
    x = sp.stats.bernoulli(mu0).rvs(50)
    N0, N1 = np.bincount(x, minlength=2)
    a, b = a + N1, b + N0
    plt.plot(xx, sp.stats.beta(a,b).pdf(xx), ls="-.",
             label="{}차 추정".format(i))
    print("{}차 추정: 모드 = {:4.2f}".format(i, (a-1)/(a+b-2)))
plt.vlines(x=0.65, ymin=0, ymax=12)
plt.ylim(0, 12)
plt.legend()
plt.title("베르누이분포의 모수를 베이즈추정법으로 추정한 결과")
plt.show()
```

loop 돌면서 모수 a, b 업데이트

result



카테고리분포의 모수추정

"클래스 갯수가 K개인 카테고리분포의 모수 μ벡터를 베이즈 추정법으로 추정"

- 카테고리 분포의 모수의 각 원소는 모두 0부터 1 사이의 값을 가지므로,

사전 분포는 하이퍼모수 $a_k = 1$인 디리클리 분포라고 가정한다.

$$p(\mu) \propto \prod_{k=1}^{K} \mu_k^{a_k - 1} \quad (a_k = 1, \text{ for all } k)$$

데이터는 모두 독립적인 카테고리 분포의 곱이므로, 가능도 함수는 다음처럼 다항분포다.

$$p(x_1, \ldots, x_N \mid \mu) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{i,k}}$$

베이즈 정리로 사후 분포를 구하면 다음과 같이 갱신된 하이퍼모수 a_i'를 가지는

디리클리 분포가 된다.

$$p(\mu \mid x_1, \ldots, x_N) \propto p(x_1, \ldots, x_N \mid \mu)\, p(\mu)$$

$$= \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{i,k}} \cdot \prod_{k=1}^{K} \mu_k^{a_k - 1}$$

$$= \prod_{k=1}^{K} \mu^{\sum_{i=1}^{N} x_{i,k} + a_k - 1}$$

$$= \prod_{k=1}^{K} \mu^{N_k + a_k - 1}$$

$$= \prod_{k=1}^{K} \mu^{a_k' - 1}$$

이 경우에도 마찬가지로 디리클레분포는 켤레사전분포임을 알 수 있다.
갱신된 하이퍼모수값은 다음과 같다.

$$\alpha_k' = N_k + \alpha_k$$

⑪ 정규분포의 기댓값 모수 추정

이번에는 정규분포의 기댓값 모수를 베이지안 방법으로 추정한다. (분산 σ^2은 알고있다고 가정)
모수의 사전분포로는 기댓값은 $-\infty$부터 ∞까지 모든 수가 가능하므로, 정규분포를 사용한다.

$$p(\mu) = N(\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma^2}\right)$$

데이터는 모두 독립적인 정규분포의 곱이므로, 가능도 함수는 다음과 같다.

$$p(x_1, \dots, x_N \mid \mu) = \prod_{i=1}^{N} N(x_i \mid \mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$p(\mu \mid x_1, \dots, x_N) \propto p(x_1, \dots, x_N \mid \mu)\, p(\mu)$$

$$\propto \exp\left(-\frac{(\mu-\mu_0')^2}{2\sigma_0'^2}\right)$$

베이즈 정리를 사용하여 사후 분포를 구하면, 다음과 같이 갱신된 하이퍼모수를 갖는 정규분포가 된다. N이 작으면 기존 모수를 믿겠다

N이 크면 2번째 항을 믿겠다

$$\mu_0' = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\frac{\sum x_i}{N}$$

(precision)

$$\frac{1}{\sigma_0'^2} = \underbrace{\frac{1}{\sigma_0^2}}_{\text{옛날에 갖고 있던 믿음의 정도}} + \frac{N}{\sigma^2}$$

N (데이터 갯수)에 비례해서 믿음 (precision)이 커진다.