본론 : 추정 등급 구하기

고등학교 수업에서는 위의 베이즈 정리를 이산적인 상황에서밖에 다루지 않지만, 베이즈 정리는 연속적인 분포에서도 사용할 수 있다. 여기서는 사전 확률 대신에 **사전 분포**라는 개념이, 사후 확률 대신에 **사후 분포**라는 개념을 논한다. 그리고 이를 확률로 접근하는 게 아니라 **가능도**로 접근한다. 가능도에 대한 얘기도 이후에 하려고 한다. 여기서는, 그냥 확률밀도함수의 y값 정도로만 생각해도 될 것이다.

또한, 몇 가지 가정이 필요하다.

1. 성적의 분포는 정규 분포를 따른다.

2. 성적의 분포의 표준 편차를 알고 있다.

1번은 쉽게 수긍할 수 있는 가정이지만, 2번은 그렇지 않다. 그래서, 본문에서는, 예전에 치룬 비슷한 시험(난이도, 치루는 학생, 과목 등이)에서의 성적의 표준 편차를 이번 시험 성적의 표준 편차로 채택하기로 했다. 물론, 그것이 이번 시험의 표준 편차와 완전히 같지는 않겠지만, 어느 정도 비슷할 것이라고 예측되기 때문이다.

정규분포의 기대값 모수는 `-∞ ~ ∞`의 모든 수가 가능하기 때문에

성적 분포를 추정하기 위해 사전 분포를 선택한다. "사전 분포는 정규 분포여야 하며", 독자가 생각했을 때 그럴 듯한 평균(μ_0)과 그 평균의 확신도에 준하는 표준 편차(σ_0)을 적용한다. 주의할 것은, 이 μ_0와 σ_0 값은 모수 μ의 분포 추정을 위해 존재하는 값들이라는 사실이다. 즉, 여기서 논하는 μ_0는 최초의 평균 추정값이고, σ_0는 해당 추정값에 대해 어느 정도의 확신도가 있는지에 대한 얘기다.

본문에서 소개하는 베이즈 추정을 이용하면, 성적의 평균, μ를 추정할 수 있다. 이렇게 찾은 μ 값을 정규 분포에 다시 적용해서, 등급을 추산하면 될 것이다.

먼저, 가능도의 관점에서 다음이 성립한다.

$$\underbrace{P(\mu|X)}_{\text{사후분포}} \propto \underbrace{P(X|\mu)}_{\text{가능도}} \underbrace{P(\mu=\mu_0)}_{\text{사전분포}}$$

이때, X는 성적이라는 확률 변수이다. 이제부터 식을 전개할 텐데, 관계를 더 명확히 보기 위하여 관계를 한 번 더 서술한다.

$$① X \sim N(\mu, \sigma^2)$$
$$② \mu \sim N(\mu_0, \sigma_0^2)$$

이제 위 관계에서 오른쪽 식의 값을 계산 해 보자.

$$P(X|\mu)P(\mu=\mu_0) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

여기서, $\frac{1}{\sqrt{2\pi}\sigma}$와 $\frac{1}{\sqrt{2\pi}\sigma_0}$은 상수이기 때문에, 일단은 무시한다. 나중에, 확률 분포의 총 넓이 합이 1이 되게만 상수는 조정해 주면 된다. 이제, 지수 함수 부분을 정리하면,

$$e^{-\frac{(x-\mu)^2}{2\sigma^2}-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}} = e^{-\frac{\sigma_0^2(x-\mu)^2+\sigma^2(\mu-\mu_0)^2}{2\sigma^2\sigma_0^2}}$$

여기서 지수에 올라간 부분을 자세히 보자. 이를 μ로 정리하면,

$$-\frac{(\sigma_0^2+\sigma^2)\mu^2-2(\sigma_0^2x+\sigma^2\mu_0^2)\mu+\sigma_0^2x^2+\sigma^2\mu_0^2}{2\sigma^2\sigma_0^2}$$

이때, 분수의 뒤에 붙는 상수의 경우에도, 상수일 뿐이므로, 나중에 조정해 줄 수 있다. 다시 말해, 상수를 임의로 수정할 수 있으므로 아래의 완전제곱식으로 바꾸어 줄 수 있다.

$$-\frac{\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}(\mu-\frac{\sigma_0^2+\sigma^2\mu_0}{\sigma_0^2+\sigma^2})^2$$

이것을 다시 지수로 올리면,

$$e^{-\frac{\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}(\mu-\frac{\sigma_0^2+\sigma^2\mu_0}{\sigma_0^2+\sigma^2})^2}$$

위 식에서 $\frac{1}{\frac{\sigma_0^2+\sigma^2}{\sigma_0^2\sigma^2}} = \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}$을 분산으로, $\frac{\sigma_0^2x+\sigma^2\mu_0}{\sigma_0^2+\sigma^2}$을 평균으로 취급하면, 정규 분포의 꼴로 해석할 수 있다. 정규 분포의 확률 밀도 식의 형태를 적용하면, 원하는 결과를 얻는다.

$$P(\mu|X) = \frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}}} e^{-\frac{(\mu-\frac{\sigma_0^2 x+\sigma^2\mu_0}{\sigma_0^2+\sigma^2})^2}{2\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}}}$$

보기 불편하지만, 이 식은 다음을 의미한다. 사후 분포의 μ에 대해서,

$$\mu \sim N(\frac{\sigma_0^2 x + \sigma^2\mu_0}{\sigma_0^2+\sigma^2}, \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2})$$

이게 의미하는 것이 무엇인가? 자신의 점수나 다른 사람의 점수를 알게 되었을 때, 추정하는 모수의 평균과 표준 편차가, 위처럼 변한다는 말이 된다. 나중에, 더 이상의 업데이트가 필요없을 때, 이 분포의 평균을 모집단의 평균에 대한 추정값으로 사용할 수 있다.

결론

위의 방식을 따르면, 모집단의 평균을 추정할 수 있고, 원래 알고 있던 모집단의 표준 편차를 활용하여 자신의 점수가 상위 몇 %에 위치할지 계산할 수 있을 것이다. 물론, 이는 오차가 꽤나 크게 나오겠지만, 어디까지나 **추정**일 뿐이다. 정확한 결과는, 성적을 집계하는 기관에서 발표하기를 기대할 수밖에는 없다.