인공지능을 위한 수학 I 9주차 강의

김영록

한국외국어대학교 대학원

2024년 5월 8일

차례

학습포인트

1. 평균과 기댓값이 같은 의미라는 것을 안다.
2. 분산과 공분산의 계산 방법을 이해할 수 있다.

https://en.wikipedia.org/wiki/Average

https://en.wikipedia.org/wiki/Mean

Example

가상의 온라인 쇼핑몰 Mamazon.com의 매출 데이터를 분석해 보자. 고객의 구매 데이터가 다음 표와 같을 때 다음 달인 7월의 매출을 추정하시오.

Mamazon.com 매출 데이터 - 2018년 상반기

고객명	1월	2월	3월	4월	5월	6월	소계
백소연	5,000원	5,000원	5,000원	5,000원	5,000원	5,000원	30,000원
이민준	10,000원	3,000원	1,000원	1,000원	15,000원	0원	30,000원
이용진	3,000원	7,000원	2,000원	8,000원	4,000원	6,000원	30,000원

1. 이 데이터를 보고 우선 생각할 수 있는 것은 과거 6개월간의 매출을 근거로 이후 한 달 동안의 매출이 어느 정도 나올지 기댓값을 구해보는 것이다.
2. 지난 6개월간의 매출을 모두 더해보면 총 90,000원이 되는데, 이것을 한 달 기준으로 환산하면 15,000원이다.
3. 별 다른 일 없이 이대로만 이어진다면 이후 한 달 동안에도 15,000원의 매출이 발생할 것이라 기대해 볼 수도 있다.
4. 이렇게 생각하는 방식이 우리가 알고 있는 평균에 대한 생각이다.
5. 평균은 수학적으로 확률에서 말하는 기댓값과 같은 의미이다.
6. 지난 6개월간의 매출 평균이 다음 달의 예상 매출액이 된다라고 하는 것을 확률의 관점에서 달리 표현하면 6개의 확률변수(각 달의 매출액)가 각각 같은 확률 $\left(\frac{1}{6}\right)$로 발생하므로, 다음 한 달 동안의 매출에 대한 기댓값은 각 월의 매출에 $\frac{1}{6}$을 곱한 것을 모두 더한 합계와 같다.

Theorem

n개의 확률변수가 각각 $x_1, x_2, \ldots, x_n$이라는 값을 가질 때 평균값 $\bar{x}$는 다음과 같다.

$$\bar{x} = \sum_{i=1}^{n} \frac{1}{n} \cdot x_k = \frac{1}{n} \sum_{i=1}^{n} x_k$$

1. 과연 앞의 방식대로 평균값만 구하면 다음 달의 매출을 올바르게 예상할 수 있는가?
2. 과거 6개월간의 월 매출을 자세히 보면 적게는 8,000원부터 많게는 24,000원까지 다양한 값을 가지고 있다.
3. 세 명의 고객으로부터 발생하는 매출은 각각 서로 다른 패턴으로 변화하는 것이 관찰된다.
4. 7월의 매출이 평균 매출과 같은 15,000원이 된다는 보장은 어디에도 없고, 평균값으로부터 얼마나 차이가 날지에 대하여 전혀 알 수 없다.

5. 그래서 단순히 평균값을 구하는 방법 이외에도 평균값과 데이터가 얼마나 차이가 나는지에 대해 데이터의 흩어진 정도를 표현할 방법도 생각해 볼 필요가 있다.
6. 우선 평균값으로부터의 차이, 즉 편차(deviation)에 주목하자.
7. 주어진 데이터에 의하면 각 고객으로부터 발생한 과거 6개월간의 매출액은 각각 30,000원씩으로, 평균 월 매출은 5,000원이 나온다. 편차는 각 월의 매출액에서 평균값을 빼면 구할 수 있다.

https://en.wikipedia.org/wiki/Standard_deviation

Mamazon.com 매출 데이터 - 2018년 상반기(편차)

고객명	평균매출	1월	2월	3월	4월	5월	6월	편차합
백소연	5,000원	0원	0원	0원	0원	0원	0원	0원
이민준	5,000원	5,000원	−2,000원	−4,000원	−4,000원	10,000원	−5,000원	0원
이용진	5,000원	−2,000원	2,000원	−3,000원	3,000원	−1,000원	1,000원	0원

8. 이처럼 편차의 관점에서 보면 매달 얼마만큼의 매출액이 고객별로 흩어져 있는지를 알 수 있다.
9. 이때, 편차의 합계를 구하여 보면 0이 되는데, 이 것은 편차가 평균값을 중심으로 계산되었기 때문에, (+) 방향으로 흩어진 매출의 차이와 (−) 방향으로 흩어진 매출의 차이가 상쇄되기 때문이다.
10. 이때, 분산(variance)이라는 개념이 필요하다.
    https://en.wikipedia.org/wiki/Variance
11. 편차는 (+) 방향과 (−) 방향의 양쪽에 모두 있기 때문에 합계를 구해 보면 0이 된다.
12. 데이터가 흩어진 정보를 얻어내려면 편차의 (+)와 (−) 같은 부호를 없애주어야 하는데, 편차를 제곱한 다음 합계를 구하고, 이 것을 다시 평균값으로 만든 것이 분산 σ^2이다.
13. 분산을 이대로 사용하면 제곱한 값이기 때문에 단위를 표현하기가 애매하다.

14. 이 예제에서는 이민준씨의 상반기 매출에 대한 분산이 $\sigma^2 = 31,000,000$원^{2}이 되는데, 본래의 단위 의미를 도로 찾기 위하여 분산 σ^2의 제곱근인 σ를 사용할 수 있다.
15. 이때, 이러한 σ를 표준편차(standard deviation)라고 한다.

Theorem

n개의 확률변수가 각각 $x_1, x_2, \ldots, x_n$이라는 값을 가지고 평균값이 $\bar{x}$일 때 분산 σ^2은 다음과 같다.

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$$

그리고 표준편차 σ는 다음과 같다.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

16 이 방법으로 Mamazon.com의 매출 데이터에 대한 분산과 표준편차를 계산한 것이 다음의 표이다.

17 표준편차가 가장 큰 것은 이민준씨의 5,568원이고, 가장 작은 것은 백소연씨의 0원입니다. 백소연씨의 경우는 데이터가 전혀 흩어지지 않아 매출에 대한 변화가 전혀 없다는 것을 알 수 있다.

Mamazon.com 매출 데이터 - 2018년 상반기(분산, 표준편차)

고객명	평균매출	분산 σ^2	표준편차 σ
백소연	5,000원	0(원^2)	0원
이민준	5,000원	31,000,000(원^2)	5,568원
이용진	5,000원	4,666,667(원^2)	2,160원

18 이렇게 분산과 표준편차를 사용하면 데이터가 얼마나 흩어져 있는지, 얼마나 차이가 심한지를 알 수 있다.

19 기본적으로 평균과 분산, 표준편차는 데이터의 경향을 표현할 때 사용한다.

20. 참고로 표준편차 σ의 특성을 이용하면 7월의 예상 매출이 정규분포를 따를 때 약 68%의 확률로 5,000$\pm 1\sigma$가 된다라고 추정할 수 있다.
21. 그리고 이민준씨의 경우는 5월 한달 동안 15,000원이라는 큰 매출을 발생시켰지만, 사실 5,000$\pm 1\sigma$원 $= 10,568$원 을 넘어서는 매출이 일어날 확률은 약 16%에 불과했다는 것도 알 수 있다.
22. 한편 Mamazon.com이 더 많은 고객을 수용할 수 있는 온라인 쇼핑몰이고, 지금까지 보아온 고객 세 명의 매출 정보는 더 많은 데이터 중의 극히 일부라고 할 때, 다음과 같은 질문을 할 수 있다.
23. 과연 세 명의 고객 중에서 전체 매출의 월간 동향에 반응하며 트랜드에 민감한 구매 성향을 보이는 고객은 누구일까요?
24. Mamazon.com 전체의 월 매출이 다음 표와 같이 주어졌을 때, 세 명의 고객 각각이 월 매출과 어느 정도의 상관관계를 가지는지 조사하려면 공분산(covariance)이라는 개념이 필요하다.

Mamazon.com 매출 데이터 - 2018년 상반기 (월 매출)

고객명	1월	2월	3월	4월	5월	6월	소계
백소연	5,000원	5,000원	5,000원	5,000원	5,000원	5,000원	30,000원
이민준	10,000원	3,000원	1,000원	1,000원	15,000원	0원	30,000원
이용진	3,000원	7,000원	2,000원	8,000원	4,000원	6,000원	30,000원
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
월 매출	2천 5백만원	4천만원	2천만원	5천5백만원	3천5백만원	4천5백만원	2억2천만원

https://en.wikipedia.org/wiki/Covariance

Theorem

두 가지 데이터에 대한 n조의 확률변수 $(X, Y) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$이 있다고 가정한다. X의 평균이 μ_x이고 Y의 평균이 μ_y라고 할 때 공분산 Cov(X, Y)는 다음과 같다.

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu_x)(y_k - \mu_y)$$

25. 이 공식을 이용하려면 우선 두 가지 데이터를 결정해야 하는데, 이 예에서는 이민준씨의 매출과 월 매출로 시험해 보자.
26. 공식을 자세히 살펴보면 매월 두 가지의 데이터에 대하여, 각각의 편차를 구한 다음 서로 곱하고, 그리고 그것을 전체 개월 수만큼 합한 다음에 다시 개월 수로 나누어 평균으로 만들고 있다는 것을 알 수 있다.
27. 월 매출의 평균을 실제로 계산해 보면 2억2천만원÷6개월=36.66 백만원, 즉 약 3천7백만원이 나온다. 계산을 쉽게 하기 위하여 월 매출을 백만원 단위로 계산하여 공분산을 구하면 다음과 같다.

Mamazon.com 매출 데이터 – 2018년 상반기(분산, 표준편차, 공분산)

고객명	평균매출	분산 σ^2	표준편차 σ	공분산
백소연	5,000원	0(원^2)	0원	0
이민준	5,000원	31,000,000(원^2)	5,568원	−21,667
이용진	5,000원	4,666,667(원^2)	2,160원	24,167
평균 월 매출	약3천7백만원	138.89백만(원^2)	약1천2백만원	-

28. 공분산은 양수가 나오기도 하고 음수가 나오기도 한다.
29. 공분산이 양의 값을 가질 때, 두 가지 데이터는 양의 관계가 있다고 하고, 공분산이 음의 값을 가질 때, 두 가지 데이터는 음의 관계가 있다고 한다.
30. 양의 관계란 두 데이터 중 어느 한쪽이 증가할 때 다른 한쪽도 증가하는 관계라는 의미이고, 음의 관계란 두 데이터 중 어느 한쪽이 증가할 때 다른 한쪽은 감소하는 관계라는 의미이다.
31. 위의 표에 따르면 이민준씨는 음의 관계인 경향을 띄고 있어서, 전체 매출이 오를 때 이민준씨의 구매액은 줄어든다.
32. 반대로 이용진씨는 양의 관계인 경향을 띄고 있어서 전체 매출이 오를 때 이용진씨의 구매액도 따라서 늘어난다는 것을 알 수 있다.
33. 다만, 공분산의 절댓값이 크다고 해서 양의 관계나 음의 관계의 강도가 더 세다고 말할 수는 없다.
34. 양의 관계나 음의 관계의 강도는 다음 절에서 배울 상관계수로 비교할 수 있으며, 이 값은 표준편차와 공분산을 통해서 계산할 수 있다.

인공지능에서는 이렇게 활용한다.

1. 평균과 분산, 그리고 표준편차는 과거의 데이터로부터 어떤 특징이나 경향을 밝혀낼 수 있는 가장 기본적인 방법으로, 인공지능 모델을 만들기 전에 데이터의 특징을 파악할 때 사용한다.

연습문제 4-5

영업부에서 영업 실적이 좋은 사원들의 특징을 알고 싶어 합니다. 영업 사원 4명을 선발하고 다음과 같은 5개의 지표를 뽑아보았다.

1. A. 적성 검사 결과에서 산출된 영업 직무적합도 (10점 만점)
2. B. 상사의 평가 점수 (10점 만점)
3. C. 월 평균 잔업시간
4. D. 근속년수
5. E. 계약 건수와 계약 단가 등에서 산출된 영업 실적 점수 (높을수록 우수)

	A.직무적합도	B.상사평가	C.잔업시간	D.근속년수	E.영업실적점수
권성환	9.0	9.0	20	6	100
도경태	10.0	9.5	35	8	90
권민준	8.0	7.0	5	9	75
한익준	9.0	6.0	10	9	60

1. 영업 실적점수와 나머지 네 개 지표 사이의 공분산을 각각 구하시오.
2. 다음 중 바르게 설명한 것을 하나만 고르시오.
   (가) 직무적합도와 영업 실적점수는 양의 관계이기 때문에 적성 검사 준비를 더 잘하면 영업 실적점수도 올라갈 것이다.
   (나) 상사평가와 잔업시간 중에서 영업 실적점수와의 공분산이 더 큰 것은 잔업시간이다. 영업 실적점수를 더 올리고 싶다면 잔업을 더 하면 된다.
   (다) 직무적합도와 상사평가 중에서 영업 실적점수와의 공분산이 더 큰 것은 직무적합도이다. 그러므로 상사평가가 적성 검사보다도 실제 능력을 더 잘 반영하고 있다.
   (라) 근속년수와 영업 실적점수는 음의 관계이나 현재 데이터만으로는 왜 그런지 알 수 없다.

Answer:

(1) 공분산을 계산하기에 앞서 평균값부터 구하자. 영업 실적점수를 $\mathrm{p_i}$ 라 하고 근속년수를 $\mathrm{q_i}$ 라 할 때, 영업 실적점수의 평균 μ_p 와 근속년수의 평균 μ_q 를 구하면 다음과 같다. (이때, i는 1에서 4)

$$\mu_\mathrm{p} = \frac{1}{4}(100+90+75+60) = \frac{325}{4}$$
$$\mu_\mathrm{q} = \frac{1}{4}(6+8+9+9) = 8$$

이제 이러한 평균값을 이용해서 $\mathrm{Cov}(\mathrm{P},\mathrm{Q})$ 를 구한다.

$$\mathrm{Cov}(\mathrm{P},\mathrm{Q}) = \frac{1}{4}\sum_{\mathrm{i}=1}^{4}(\mathrm{p_i}-\mu_\mathrm{p})(\mathrm{q_i}-\mu_\mathrm{q})$$
$$= \frac{1}{4}\left\{\left(100-\frac{325}{4}\right)\cdot(6-8)+\left(90-\frac{325}{4}\right)\cdot(8-8)+\cdots\right\} = -\frac{65}{4}$$

같은 방식으로 영업 실적점수와 직무적합도, 영업 실적점수와 상사평가, 영업 실적점수와 잔업시간에 대한 공분산을 구할 수 있다.
정답:
직무적합도: $\frac{15}{4} = 3.75$, 상사평가: $\frac{645}{32} \fallingdotseq 20.16$,
잔업시간: $\frac{875}{8} \fallingdotseq 109.38$, 근속년수: $-\frac{65}{4} = -16.25$

(2) 정답: (라)
(가)는 뒷부분의 내용이 잘못되었다. 비록 양의 관계에 있다고 하더라도 적성 검사를 준비하는 것이 영업 실적점수를 올리는 데 직접적인 도움이 된다고 단정할 수는 없다. 공분산이나 뒤에 나올 상관계수로는 인과관계를 설명할 수 없다.
(나)는 공분산의 크기를 비교하는데 사용하고 있는 것이 잘못되었다.
(다)는 공분산의 크기를 비교하는데 사용하고 있으며, 심지어 대소관계도 잘못되었다.

칼럼: 표준편차와 표준점수 1

1. 표준편차 σ는 고등학교나 대학교에서 성적을 분석할 때 사용하는 표준점수와 관련이 깊다.
2. 시험 점수라는 것은 출제된 문제의 경향이나 수험자의 학습 수준과 같이 다양한 변수로부터 영향받기 때문에 상황에 따라 점수에 대한 의미는 달라지기 마련이다.
3. 실제로 시험 점수 자체에 절대적인 의미를 부여하기가 곤란한 상황이 발생할 수 있는데, 예를 들어 100점 만점의 시험에서 수험자들 사이에 우열을 가려야 한다고 가정하자.
4. 이때 평균 점수가 60점인 시험에서 80점을 받은 수험자와 평균 점수가 30 점인 시험에서 60점을 받은 수험자가 있다면, 둘 중 누가 더 우수한 수험자일까요? 적어도 이때만큼은 시험 점수 자체가 좋은 평가 기준이 될 수 없다.
5. 서로 다른 시험의 난이도나 서로 다른 수험자들이라 하더라도 서로 비교가 가능한 평가 지표가 필요하게 되는데, 이때 사용하는 것이 표준점수이다.

칼럼: 표준편차와 표준점수 2

6. 어떤 수험자의 평균 점수가 μ이고, 표준편차는 σ, 그리고 이번 시험 점수가 $\mathsf{x_i}$라고 할 때 이 수험자의 표준점수 $\mathsf{X_i}$는 다음과 같이 구할 수 있다.

$$\mathsf{X_i} = \frac{10(\mathsf{x_i} - \mu)}{\sigma} + 50$$

7. 이 식에 의하면 수험자가 받은 점수가 평균일 때 표준 점수는 50이 된다.

$$\mathsf{X_i} = \frac{10 \cdot 0}{\sigma} + 50 = 50$$

8. 그리고 편차가 $+\sigma$가 되면 표준 점수는 60이 되고, $+2\sigma$가 되면 표준 점수는 70이 된다.

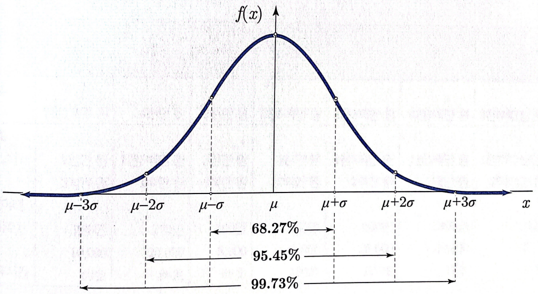
$$\mathsf{X_i} = \frac{10 \cdot \sigma}{\sigma} + 50 = 60, \ \mathsf{X_i} = \frac{10 \cdot 2\sigma}{\sigma} + 50 = 70$$

9. 표준 점수는 편차 정보를 활용해서 점수를 환산하도록 만들어져 있다.

칼럼: 표준편차와 표준점수 3

10. 평균이 μ, 분산이 σ^2일 때 정규분포는 다음과 같은 식으로 표현된다.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



▲ 그림 4.5.1 정규분포 그래프

칼럼: 표준편차와 표준점수 4

11. 정규분포에서는 $\mu - 1\sigma$에서 $\mu + 1\sigma$ 사이의 구간에 모집단(population)의 약 68.27%가 모여 있고, $\mu - 2\sigma$에서 $\mu + 2\sigma$ 사이의 구간에 약 95.45%가, $\mu - 3\sigma$에서 $\mu + 3\sigma$ 사이의 구간에 약 99.73%가 모여 있다.
12. 앞서 표준점수가 60일 때를 생각하면 이는 편차가 $+\sigma$이므로 $\mu + 1\sigma$의 구간에 해당하며 전체 정규분포에서 오른쪽 그래프의 상위 부분만 계산하면 $\frac{100\% - 68.27\%}{2} = \frac{31.73\%}{2} = 15.865\%$이 된다.
13. 그래서 표준점수가 60이라는 의미는 수험자의 점수 분포가 정규분포에 가깝다고 가정할 때 전체 수험자들 중 상위 15.865%에 해당한다라고 해석할 수 있다.
14. 같은 방식으로 표준점수가 70일 때의 의미($\mu + 2\sigma$), 표준점수가 80일 때의 의미($\mu + 3\sigma$)도 알 수 있다.
15. 이 계산 방법대로라면 표준점수 80이라는 의미가 전체 모집단 중에서 상위 0.135%에 해당하는 셈이니 이 표준점수가 얼마나 받기 어려운 것인지 알 수 있다.

학습포인트

1. 표준편차와 공분산으로부터 상관계수를 구할 수 있다.
2. 상관계수를 이용하면 관계의 강도를 비교할 수 있다.

https://en.wikipedia.org/wiki/Correlation_coefficient

Example

다음 표에 표시된 쇼핑몰 Mamazon.com의 다양한 데이터를 참고하여 월 매출과 관련이 깊은 지표를 찾아내시오. (이때, 월 매출 데이터는 유효숫자 2자리로, 평균값은 유효숫자 3자리로 표현함)

Mamazon.com 운영 데이터 - 2018년 상반기

데이터의 종류	1월	2월	3월	4월	5월	6월	평균
수입 월 매출	25백만원	40백만원	20백만원	55백만원	35백만원	45백만원	36.7백만원
지출 상품구입비 광고비	20백만원 2백만원	15백만원 1백만원	30백만원 4백만원	10백만원 3백만원	15백만원 2백만원	15백만원 2백만원	17.5백만원 3.33백만원
계측데이터 PV(조회수) 결제수 평균체류시간	1.8백만원 10,000 60초	2.7백만원 20,000 88초	1.6백만원 8,000 68초	6.2백만원 40,000 180초	3.2백만원 28,000 120초	3.9백만원 30,000 77초	3.23백만원 22,700 100초

1. 두 가지 데이터의 상관관계는 공분산을 구해보면 알 수 있다.
2. 월 매출은 그달에 지출한 광고비나 그달의 상품 조회수를 뜻하는 PV(page view)와 관련이 있다.
3. 월 매출 R, 광고비는 A, PV는 P라 두고, 공분산 $\mathrm{Cov}(R, A)$ 와 $\mathrm{Cov}(R, P)$ 를 계산해보자.
4. 공분산을 계산할 때는 단위를 신경쓰지 않아도 되기 때문에 숫자에만 주목하자.

Mamazon.com의 월 매출(R), 광고비(A), PV(P)의 편차 - 2018년 상반기

	1월편차	2월편차	3월편차	4월편차	5월편차	6월편차	표준편차
R	-11.7	3.3	-16.7	18.3	-1.7	8.3	11.8
A	-0.33	-1.33	1.67	0.67	-0.33	-0.33	0.943
P	-143	-53	-163	297	-3	67	154

$$\begin{aligned}\mathrm{Cov}(R, A) &= \frac{1}{6}\left((-11.7)\cdot(-0.33) + 3.3\cdot(-1.33) + \cdots + 8.3\cdot(-0.33)\right)\\ &= -3.056\\ \mathrm{Cov}(R, P) &= \frac{1}{6}\left((-11.7)\cdot(-143) + 3.3\cdot(-53) + \cdots + 8.3\cdot 67\right)\\ &= 1703\end{aligned}$$

1. 월 매출과 광고비는 음의 상관관계가, 월 매출과 PV는 양의 상관관계가 있다는 것을 알았다.

2. 이 상관관계들은 얼마나 강한 관계일까요?
3. 우선 공분산의 값만 보면 $\mathrm{Cov}(R, P)$의 쪽이 큰데, 사실 계산 과정에서 P의 값 자체가 다른 데이터에 비해 월등히 크기 때문에 $\mathrm{Cov}(R, P)$가 크게 나오는 것은 당연하다.
4. 단위만 보더라도 금액끼리 계산한 $\mathrm{Cov}(R, A)$와 금액과 PV 사이에 계산한 $\mathrm{Cov}(R, P)$를 단순 비교하는 것은 큰 의미가 없다.
5. 그래서 도입하는 것이 상관계수(correlation coefficient)이다.

Definition

확률변수 X와 Y의 분산이 양수이고 각각의 표준편차가 σ_X, σ_Y, 공분산이 σ_{XY}라고 할 때의 상관계수는 다음과 같다. (이때, $-1 \le \rho \le 1$)

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

6. 상관계수 ρ는 공분산을 각각의 표준편차로 나누어 단위를 없애버린 값으로, 단위가 없는 무차원수(dimensionless number)이다.
7. 또한, 상관계수를 계산할 때 공분산을 표준편차의 곱으로 나누게 되는데, 이 과정에서 ρ는 -1에서 1의 사이의 값을 가지게 되고 이러한 조작을 정규화라고 부른다.
8. 지금까지는 값이 제 각각이어서 비교할 방법이 없었던 공분산도 상관계수로 변환함에 따라 상관관계의 강약을 비교할 수 있게 되었다.
9. 실제로 상관계수 ρ_{RA}와 ρ_{RP}를 계산해 보자.

$$\rho_{\mathrm{RA}} = \frac{\mathrm{Cov}(\mathrm{R},\mathrm{A})}{\sigma_{\mathrm{R}}\sigma_{\mathrm{A}}} = \frac{-3.056}{11.8 \times 0.943} = -0.2746 \tag{1}$$

$$\rho_{\mathrm{RP}} = \frac{\mathrm{Cov}(\mathrm{R},\mathrm{P})}{\sigma_{\mathrm{R}}\sigma_{\mathrm{P}}} = \frac{1703}{11.8 \times 154} = 0.9372 \tag{2}$$

10. 상관계수는 $+1$에 가까울수록 양의 관계가 강하고, -1에 가까울수록 음의 관계가 강하다.
11. 상관계수가 0에 가까울수록 상관관계가 약하다고 보는데, 일반적으로 상관계수의 절댓값이 0.7보다 클 때 상관관계가 강하다고 말한다.
12. 실제로 상관계수 ρ_{RA}와 ρ_{RP}를 계산한 결과를 보면 ρ_{RP}, 즉 월 매출과 PV의 상관계수가 ρ_{RA}보다 크게 나왔다.
13. 결과적으로 월 매출은 PV와 양의 강한 관계에 있다는 것을 알 수 있다.

인공지능에서는 이렇게 활용한다.

1. 사람이 직관적으로 분석하기 어려울 만큼의 대량 데이터가 있다면, 컴퓨터로 하여금 무수히 많은 파라미터를 조합하고, 그들의 상관계수를 계산하면서, 상관관계가 강한 조합을 찾아내게 만들 수 있다. 이런 과정을 거치면 사람이 미처 발견하지 못했던 숨은 관계나 데이터의 특징을 찾을 수 있어 데이터를 보다 유용하게 활용할 수 있게 된다.

연습문제 4-6

앞서 살펴본 표 Mamazon.com 운영 데이터 – 2018년 상반기의 데이터를 활용하여 다음 물음에 답하시오.

1. 2018 상반기의 데이터에서 광고비 항목을 이번 달의 광고비가 아니라 지난 달의 광고비로 데이터를 변경하려 한다. 이때의 월 매출과 전월 광고비의 상관계수를 구하시오. (단, 2017년 12월의 광고비는 1백만원)
2. 다음 중 바르게 설명한 것을 하나만 고르시오.
   - (가) 광고비와 PV는 음의 상관관계이기 때문에 광고비를 늘리면 PV가 줄어들 수 있다.
   - (나) 전월의 광고비와 월 매출은 상관계수가 약 0.84인 양의 상관관계에 있기 때문에 광고비를 늘리면 다음 달의 매출이 늘어날 확률이 약 84%이다.
   - (다) 평균체류시간이 긴 달은 PV도 많아지는 경향이 있다.
   - (라) PV와 결제수, PV와 평균체류시간 중 상관관계의 강도가 센 쪽은 PV와 평균체류시간이다.

(1)

	1월	2월	3월	4월	5월	6월
해당 달의 광고비	2백만원	1백만원	4백만원	3백만원	2백만원	2백만원
이전 달의 광고비	1백만원	2백만원	1백만원	4백만원	3백만원	2백만원

Mamazon.com 운영 데이터 - 2018 상반기

데이터의 종류	1월	2월	3월	4월	5월	6월	평균
지출 이전 달의 광고비	1백만원	2백만원	1백만원	4백만원	3백만원	2백만원	2.17백만원

Mamazon.com의 월 매출(R), 광고비(A)의 편차 - 2018년 상반기

	1월편차	2월편차	3월편차	4월편차	5월편차	6월편차	표준편차
R	-11.7	3.3	-16.7	18.3	-1.7	8.3	11.8
A	-1.17	-0.17	-1.17	1.83	0.83	-0.17	1.07

월 매출과 광고비 간의 상관계수를 구하면 다음과 같다.

$$\mathrm{Cov}(\mathrm{R}, \mathrm{A}) = \frac{1}{6}((-11.7)\cdot(-1.17) + \cdots + 8.3\cdot(-0.17)) = 10.56$$

$$\rho_{\mathrm{RA}} = \frac{\mathrm{Cov}(\mathrm{R}, \mathrm{A})}{\sigma_{\mathrm{R}}\sigma_{\mathrm{A}}} = \frac{10.56}{11.8 \times 1.07} = 0.836$$

2. 정답: (다)
   (가)는 음의 상관관계가 아니라 양의 상관관계이다.
   (나)에서 상관계수는 대소 비교가 가능한 -1에서 $+1$의 값을 가지며, 그 자체가 확률을 의미하지는 않는다.
   (라)의 상관계수의 강도가 센 쪽은 PV와 결제수 쪽이다.
   PV와 결제수의 상관계수: 약 0.955
   PV와 평균체류시간의 상관계수: 약 0.884

학습포인트

1. 최대가능도추정의 개념을 이해할 수 있다.

1. 통상 우리가 알고 있기에는 동전을 던졌을 때 앞이 나올 확률은 $\frac{1}{2}$ 이고, 주사위를 던졌을 때 각 면의 숫자가 나올 확률은 $\frac{1}{6}$이다.
2. 하지만 이 확률들은 논리적으로 그럴 것이라 상상한 값이고, 현실 세계의 동전이나 주사위가 실제로 이 확률의 지배을 받고 있는지에 대해서는 알 수 있는 방법이 없습니다.
3. 결국 우리는 어떤 사건의 확률을 알기 위하여 몇 번이고 시행을 반복하면서, 그 과정에서 얻은 관측 결과를 통하여 추정을 해 보는 것 이외에는 확률을 구해볼 별 다른 방법이 없다.
4. 이번 절에서는 통계적인 추정 방법인 최대가능도추정(maximum likelihood estimation)에 대하여 알아보자.

5. 최대가능도추정은 다른 표현으로 최대우도추정이라고도 하는데, 이때의 우라는 글자의 뜻을 보면 매우이고, 그럴듯하다라는 의미를 가지고 있다.
6. 즉, 최대가능도추정이란 가장 그럴듯하게 (값을) 추정한다는 의미로, 영어로는 가능도를 likelihood라고 표현한다. A star like a diamond를 해석하면 다이아몬드와 같은 별이라고 할 때의~와 같은의 like가 변형된 것이다.
7. 최대가능도를 추정한다는 말은 곧, 파라미터 θ에 대한 가능도함수 $\mathrm{L}(\theta)$를 최대화할 수 있는 θ값을 구하는 것을 의미한다.
8. 최댓값을 가지는 지점은 1계 미분을 했을 때 $\frac{d\mathrm{L}(\theta)}{d\theta} = 0$이 되는 지점이고, 이때의 θ를 구하면 된다.

https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

Theorem

최대가능도추정이란 어떤 파라미터 θ의 값을 추정하는 방법이며, θ에 대한 가능도함수 $\mathrm{L}(\theta)$를 최대로 만드는 θ를 찾으면 된다. 따라서 이때의 θ에 대한 추정값은 다음 방정식을 만족한다.

$$\frac{d\mathrm{L}(\theta)}{d\theta} = 0$$

Example

1. 주사위를 던졌을 때 숫자 1이 나올 확률은 $\frac{1}{6}$이라고 알고 있지만, 꼭 그렇다고 단언할 수는 없기 때문에 일단 그 확률을 θ라고 하자.
2. 주사위를 많이 던지다 보면 자연스럽게 확률을 알게 될 것이라 생각하며 100번을 던졌는데, 그 중에서 숫자 1이 나온 것은 모두 20번이었다.

Example

3. 확률은 잘 모르지만 어쨌거나 100번을 던졌을 때 1이 20번 나왔다라는 관찰 결과가 나온 것으로부터 최대가능도추정은 시작된다.
4. 최댓값을 가지는 지점은 1계 미분을 했을 때 $\frac{dL(\theta)}{d\theta} = 0$이 되는 지점이고, 이때의 θ를 구하면 된다.
5. 100번 중에서 1이 20번이 나오는 경우의 수는 ${}_{100}C_{20}$이고, 이러한 관찰 결과가 발생할 확률을 가능도함수 $L(\theta)$라고 할 때, 다음과 같은 식이 성립한다.

$$L(\theta) = {}_{100}C_{20}\theta^{20}(1-\theta)^{80}$$

6. 일반적인 이산확률분포의 식은 확률의 곱으로 표현되는 일이 많다 보니 미분 자체가 어려운 일이 비일비재하다. 이러한 어려움을 피하는 방법으로는 가능도함수에 자연로그를 붙여 주어 로그가능도함수 $\log_e L(\theta)$를 만들면 된다.

로그가능도함수로 바뀐 식이라 할지라도, 이 식을 최대로 만드는 θ가 가능도함수 $\mathsf{L}(\theta)$도 최대로 만들기 때문에 답을 구하는데에는 전혀 문제가 없다.

Theorem

가능도함수 $\mathsf{L}(\theta)$를 최대로 하는 θ는 로그가능도함수 $\log_e \mathsf{L}(\theta)$에 대하여 다음 방정식을 만족한다.

$$\frac{d}{d\theta} \log_e \mathsf{L}(\theta) = 0$$

1. 왜 멀쩡한 식에 로그를 적용하는지 궁금할 수 있다.
2. 다음 식을 보면 알 수 있듯이 로그를 사용하면 곱셈을 덧셈으로 바꿀 수 있기 때문에 고차 방정식을 단숨에 1차방정식으로 만들 수 있어 수식을 다루는 난이도를 낮추는 효과가 있다.

3.

$$\begin{aligned}\log_e L(\theta) &= \log_e \left({}_{100}C_{20}\theta^{20}(1-\theta)^{80} \right) \\ &= \log_e {}_{100}C_{20} + 20\log_e \theta + 80\log_e(1-\theta)\end{aligned}$$

4. 로그를 적용함으로써 100차 방정식이 1차 방정식으로 모양이 바뀌어 미분을 한결 더 쉽게 할 수 있게 되었다. 이제 이 식을 미분해 보자.

$$\frac{d}{d\theta}\log_e L(\theta) = 0 + \frac{20}{\theta} - \frac{80}{1-\theta} = 0$$

5. 이 식을 풀면 $\theta = 0.2$가 된다. 결국 어떤 주사위를 던졌을 때 숫자 1이 나올 확률로 가장 그럴듯한 것은 0.2이다라는 결론을 얻을 수 있었다.
6. 한편, 정규분포와 같은 연속확률분포에서는 파라미터가 여러 개인 경우도 있다. 이런 경우는 각각의 파라미터에 대하여 편미분을 하면 된다.

Theorem

가능도함수 $\mathsf{L}(\theta_1, \theta_2, \cdots, \theta_\mathsf{m})$ 을 최대로 하는 $\theta_1, \theta_2, \cdots, \theta_\mathsf{m}$ 은 다음 방정식을 만족한다.

$$\frac{\partial}{\partial \theta_1} \log_e L(\theta_1, \theta_2, \cdots, \theta_m) = 0$$
$$\frac{\partial}{\partial \theta_2} \log_e L(\theta_1, \theta_2, \cdots, \theta_m) = 0$$
$$\cdots$$
$$\frac{\partial}{\partial \theta_m} \log_e L(\theta_1, \theta_2, \cdots, \theta_m) = 0$$

1. 이산확률분포와 연속확률분포, 둘 다 가능도함수로 사용할 수 있고, 파라미터가 여러 개라 하더라도 문제가 되지는 않는다.
2. 오히려 실제로 문제가 되는 것은 수집한 데이터(사건의 관찰 결과)를 확률분포가 얼마나 적절히 잘 표현하고 있는가라는 점이다.

인공지능에서는 이렇게 활용한다.

1. 최대가능도추정은 이미 확보한 데이터를 사용해서 미처 발견하지 못한 확률 모델의 파라미터를 추정할 때 사용하는 통계 기법이다. 실제로 과거의 데이터로부터 미래를 예측할 때 이러한 방법을 많이 사용한다.

연습문제 4-7

사격에서 300발을 쏘았다. 한 발 쏠 때마다 탄착점이 표적의 중심에 가까운 순으로 10점에서 0점까지의 점수가 주어진다.

(1) 300발 중에서 10점은 20번 나왔다. 이때, 10점을 얻을 확률 θ의 최대가능도를 추정하시오.

(2) 추가로 300발을 더 쏘았다. 10점이 나온 횟수는 600발 중에서 48번이었다. 이때, 10점을 얻을 확률 θ의 최대가능도를 추정하시오.

1. 300발 중에서 10점은 20번 나왔다. 이때, 10점을 얻을 확률을 θ라 할 때, 이 사건의 관찰 결과를 반영한 가능도함수 $\mathsf{L}(\theta)$은 다음과 같다.

$$\mathsf{L}(\theta) = {}_{300}\mathsf{C}_{20}\theta^{20}(1-\theta)^{280}$$

이제 미분을 쉽게 하기 위하여 로그가능도함수로 만든다.

$$\log_{\mathsf{e}} \mathsf{L}(\theta) = \log_{\mathsf{e}} \left({}_{300}\mathsf{C}_{20}\theta^{20}(1-\theta)^{280}\right)$$

이 식의 양변을 θ로 미분한다.

$$\text{좌변} : \frac{\mathsf{d}}{\mathsf{d}\theta}\log_{\mathsf{e}} \mathsf{L}(\theta) = 0$$

$$\text{우변} : \frac{\mathsf{d}}{\mathsf{d}\theta}\left(\log_{\mathsf{e}}\left({}_{300}\mathsf{C}_{20}\right) + 20\log_{\mathsf{e}}\theta + 280\log_{\mathsf{e}}(1-\theta)\right) = 0 + \frac{20}{\theta} - \frac{280}{1-\theta}$$

좌변 $=$ 우변으로 수식을 풀면 다음과 같다.

$$0 = 0 + \frac{20}{\theta} - \frac{280}{1-\theta} \Rightarrow \frac{20}{\theta} = \frac{280}{1-\theta} \Rightarrow 280\theta = 20(1-\theta) \Rightarrow 300\theta = 20 \Rightarrow \theta = \frac{1}{15}$$

2. 600발 중에서 10점은 48번 나왔다. 이때, 10점을 얻을 확률을 θ라 할 때, 이 사건의 관찰 결과를 반영한 가능도함수 $\mathsf{L}(\theta)$은 다음과 같다.

$$\mathsf{L}(\theta) = {}_{600}\mathsf{C}_{20}\theta^{48}(1-\theta)^{552}$$

이제 미분을 쉽게 하기 위하여 로그가능도함수로 만든다.

$$\log_e \mathsf{L}(\theta) = \log_e \left({}_{600}\mathsf{C}_{20}\theta^{48}(1-\theta)^{552}\right)$$

이 식의 양변을 θ로 미분한다.

$$\text{좌변}: \frac{d}{d\theta}\log_e \mathsf{L}(\theta) = 0$$

$$\text{우변}: \frac{d}{d\theta}\left(\log_e\left({}_{600}\mathsf{C}_{48}\right) + 48\log_e\theta + 552\log_e(1-\theta)\right) = 0 + \frac{48}{\theta} - \frac{552}{1-\theta}$$

좌변 $=$ 우변으로 수식을 풀면 다음과 같다.

$$0 = 0 + \frac{48}{\theta} - \frac{552}{1-\theta} \Rightarrow \frac{48}{\theta} = \frac{552}{1-\theta} \Rightarrow 552\theta = 48(1-\theta) \Rightarrow 600\theta = 48 \Rightarrow \theta = \frac{2}{25}$$

수학적으로 정확한 최대가능도추정법 vs 실용적이지만 의심스러운 베이즈 추정법

1. 최대가능도추정법의 접근 방식은 진정한 확률 모델은 존재하며 관찰된 데이터는 그러한 모델을 충실히 따르고 있다.
2. 따라서 시행을 반복하면서 결과를 평균을 내다 보면 진정한 확률 모델이 보이기 시작할 것이다.
3. 다만, 시행을 무한히 할 수는 없기 때문에 지금 당장 얻을 수 있는 데이터로부터 가장 그럴듯한 확률을 이끌어 낼 수 밖에 없다.
4. 즉, 관찰되는 데이터를 믿을 수 밖에 없다와 같은 생각을 바탕으로 하고 있다.
5. 그래서 관찰 결과가 어쩌다가 한쪽으로 치우치거나 부적절한 확률분포를 적용하면 완전히 엉뚱한 추정 결과가 나오는 치명적인 약점이 있다.
6. 반면 이러한 약점을 보완하기 위한 방법으로 베이즈 추정법이 있는데, 이 방법은 지금까지의 관찰 결과(상상한 가설)를 근거로, 사전분포(확률)를 가정한다.
7. 그런 후에 관찰을 통해 얻은 데이터는 사전분포에 따라 얻어진 결과이므로, 그에 대한 조건부확률을 구하면 된다라는 접근 방법으로 사후확률(조건부확률)을 구한다.
8. 시행 횟수를 늘려야만 신뢰할 수 있는 최대가능도추정법과 의심스러운 전제 조건을 도입해야 하는 베이즈 추정법, 둘 중 어느 것을 사용하더라도 결국 어디까지나 결국 어디까지나 추정에 불과하다. 통계에서는 그러한 한계를 명확히 인지한 상태에서 데이터를 다루려는 자세가 무엇보다 중요하다.

1. ML Lec 07-1 - Learning rate, data preprocessing, overfitting
   `https://www.youtube.com/watch?v=1jPjVoDV_uo`