부록 C. 확률 기초

세상에는 불확실성이 많다. 며칠간 계속 맑았다고 하여 내일도 맑다는 보장은 없다. 단지 맑을 가능성이 크다고 말할 수 있을 뿐이다. 고객이 숫자  를 썼다면 3인지 5인지 불확실하니 고객에게 물어보는 수밖에 없다. 윷을 던질 때 무엇이 나올지는 확률로만 말할 수 있을 뿐 단언할 수는 없다. 컴퓨터 비전이 처리할 데이터는 이와 같이 불확실한 세상에서 발생하므로 컴퓨터 비전 알고리즘은 불확실성을 다루는 학문인 확률과 통계를 잘 활용해야 한다.

C.1 확률 변수와 확률 분포

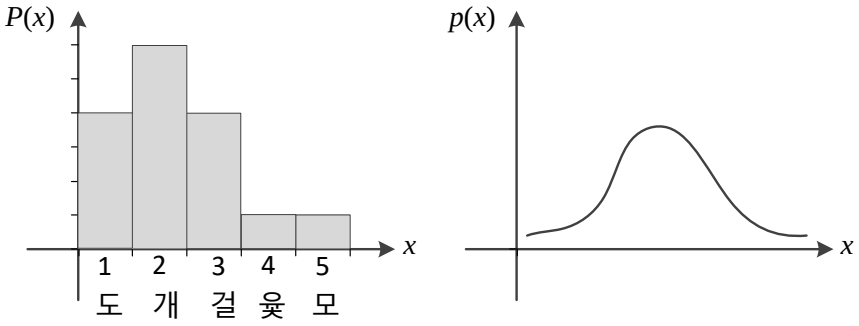
윷을 던지면 모, 윷, 걸, 개, 도의 다섯 가지 경우 중 하나가 발생한다. [그림 C-1]은 다섯 가지 경우를 보여준다.

[그림 C-1] 윷을 던졌을 때 나올 수 있는 다섯 가지 경우(왼쪽부터 도, 개, 걸, 윷, 모)

확률을 수식으로 표현하려면 다섯 가지 경우를 값으로 가지는 변수가 필요하다. 이러한 변수를 확률 변수$_{\text{random variable}}$라고 하며, x와 같이 소문자로 표기한다. 확률 변수가 가질 수 있는 값의 집합을 정의역이라 하는데, 윷놀이의 정의역은 {모, 윷, 걸, 개, 도}이다. 아래와 같이 정의역 전체에 걸쳐 확률을 표현한 것을 확률 분포$^{\text{probability distribution}}$라고 하며, 표기에 혼란이 없을 때는 $P(x=\text{모})$를 더 간략하게 $P(\text{모})$와 같이 쓴다.

$$P(x=\text{모})=\frac{1}{16}, P(x=\text{윷})=\frac{1}{16}, P(x=\text{걸})=\frac{4}{16}, P(x=\text{개})=\frac{6}{16}, P(x=\text{도})=\frac{4}{16}$$

윷놀이의 확률 분포는 [그림 C-2(a)]처럼 그래프로 표현할 수도 있다. 윷놀이는 정의역이 이산 값을 가지는데, 이처럼 이산일 때의 확률 분포를 확률질량 함수$^{\text{probability mass function}}$라고 한다. 키나 몸무게처럼 연속인 경우도 있는데, 연속인 확률분포를 확률밀도 함수$^{\text{probability density function}}$라고 한다. [그림 C-2(b)]는 연속인 경우이다. 확률은 항상 0보다 크거나 같아야 하며, 정의역에 정의된 확률을 모두 더하면 1이 되어야 한다. 즉, [그림 C-2]에서 그래프의 아래쪽 면적은 1이어야 한다.



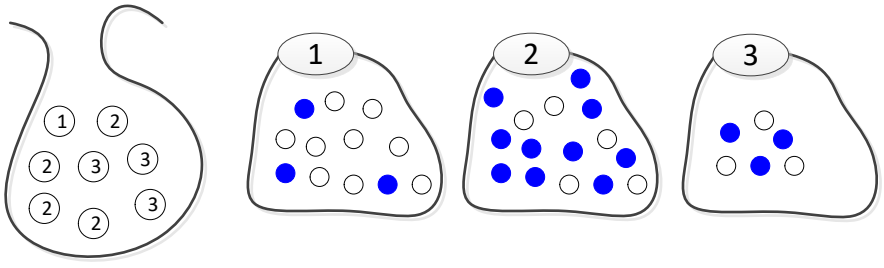
(a) 이산인 경우의 확률질량함수 (b) 연속인 경우의 확률밀도함수

[그림 C-2] 확률 분포 *** 오른쪽 p(x)에서 P를 대문자로

확률 변수가 벡터인 경우를 살펴보자. 붓꽃을 표현한 iris 데이터셋은 샘플이 4개의 특징(꽃받침의 길이, 꽃받침의 너비, 꽃잎의 길이, 꽃잎의 너비)으로 표현되므로, 확률 변수는 4차원 벡터이어야 한다. 이러한 변수를 확률 벡터[random vector]라고 하고 **x**로 표기한다. iris 데이터셋에서 특징은 길이와 너비를 뜻하므로 이산 값이 아니라 연속 값이다. 이제 정의역이 4차원 실수 공간인 $\mathbb{R}^4$가 되어, 확률 분포를 [그림 C-2(a)]처럼 쉽게 그릴 수 없을뿐더러 쉽게 계산할 수도 없다. 예를 들어, $\mathbf{x} = (5.9,3.4,0.8,6.1)$라면 $p(\mathbf{x})$를 어떻게 구할 수 있겠는가? 컴퓨터 비전은 확률 분포를 명시적으로 구하는 것을 시도하지 않는다. 데이터셋이 제공하는 부분적인 정보를 보고 예측 성능을 최대화하려고 시도할 뿐이다.

C.2 베이즈 정리

대부분 상황에서는 여러 사건이 순차적으로 또는 동시에 일어난다. [그림 C-3]의 확률 실험 장치를 가지고 기초적인 확률 계산을 해보자. 먼저 주머니에서 카드를 1장 꺼내 번호를 확인한다. ①, ②, ③이라 표시된 병 중 카드 번호에 해당하는 병에서 공을 하나 꺼내 색을 말한다. 꺼낸 카드와 공은 꺼낸 곳에 다시 넣는다. 카드 번호와 공의 색을 나타내는 확률 변수를 각각 y와 x라 하자. 이들의 정의역은 $y \in \{①, ②, ③\}, x \in \{\text{파랑}, \text{하양}\}$이다.



[그림 C-3] 확률 실험 장치

곱 규칙과 합 규칙

이제 여러 가지 확률을 계산해보자. ①번 카드를 뽑을 확률은 8개 중 하나이므로 $\frac{1}{8}$이고 $P(y=①)=\frac{1}{8}$과 같이 쓴다. 표기상 혼란이 없다고 생각하면 $P(①)=\frac{1}{8}$과 같이 써도 된다. 이제 두 확률변수를 같이 생각하자. ①번 카드를 뽑고 ①번 병에서 하얀 공을 뽑을 확률은 $P(y=①, x=\text{하양})$과 같이 표기하는데, 두 사건이 결합된 확률이라는 의미에서 결합 확률$_{\text{joint probability}}$이라고 한다. 결합 확률은 아래 식으로 구할 수 있다. 이 식에서 $P(x=\text{하양}|\ y=①)$을 조건부 확률$_{\text{conditional probability}}$이라고 하는데, y=①이라는 사건이 이미 발생한 조건에서 x=하양이라는 사건이 발생할 확률을 뜻한다.

$$P(y=①, x=\text{하양})=P(x=\text{하양}|y=①)P(y=①)=\frac{9}{12}\frac{1}{8}=\frac{3}{32}$$

이 계산 식을 곱 규칙이라 부른다. 식 (C.1)은 곱 규칙의 일반적인 형태이다.

$$\text{곱 규칙: } P(y,x)=P(x|y)P(y) \qquad (C.1)$$

이제 하얀 공이 뽑힐 확률 $P(x=\text{하양})$을 생각하자. 카드를 뽑은 다음에 공을 뽑는 상황이므로 조금 복잡해지지만, 카드가 세 종류 뿐이니 각각을 고려하면 그리 어렵지 않다. 다음 식을 이용하여 $P(x=\text{하양})$을 구할 수 있다. 이 식에서는 확률 변수를 빼고 간략하게 표기하였다.

$$P(\text{하양})=P(\text{하양}|①)P(①)+P(\text{하양}|②)P(②)+P(\text{하양}|③)P(③)=\frac{9}{12}\frac{1}{8}+\frac{5}{15}\frac{4}{8}+\frac{3}{6}\frac{3}{8}=\frac{43}{96}$$

이 계산 식을 합 규칙이라고 한다. 식 (2.24)는 합 규칙의 일반적인 형태이다.

$$\text{합 규칙: } P(x)=\sum_{y}P(y,x)=\sum_{y}P(x|y)P(y) \qquad (C.2)$$

두 확률 변수가 식 (C.3)을 만족하면 둘은 독립이라고 말한다. 지금까지 살펴본 카드 번호를 나타내는 y와 공의 색을 나타내는 x는 독립이 아니다. y에 따라 x가 결정되기 때문에 서로 강한 관련성이 있을 것이라는 사실을 쉽게 유추할 수 있고, 독립이 아니라는 사실을 직관적으로 이해할 수 있다. 게다가 $P(①, \text{하양})=\frac{3}{32}$인데, $P(①)=\frac{1}{8}$, $P(\text{하양})=\frac{43}{96}$이므로 $P(①, \text{하양})\neq P(①)P(\text{하양})$인 반례를 쉽게 제시할 수 있다. 실험 방법을 바꾸어, 뽑은 카드 번호와 무관하게 병을 임의로 골라 공을 뽑는다면 x와 y는 독립이 된다. 만일 x와 y가 발의 크기와 키라면 둘은 독립이 아닐 것이고, x와 y가 발의 크기와 성적을 나타낸다면 둘은 독립일 것이다.

$$P(x,y)=P(x)P(y) \qquad (C.3)$$

베이즈 정리

일반적으로 x와 y가 같이 일어날 결합 확률이나 y와 x가 같이 일어날 결합 확률이 같으므로 다음 식이 성립된다.

$$P(y,x)=P(x|y)P(y)=P(x,y)=P(y|x)P(x)$$

이 식을 정리하면 식 (C.4)가 된다. 이 식이 유명한 베이즈 정리Bayes' theorem이다.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \qquad (C.4)$$

베이즈 정리를 [그림 C.3]의 확률 실험에 적용하면, 다음 질문에 대한 합리적인 답을 구할 수 있다.

> "하얀 공이 나왔다는 사실만 알고 어느 항아리에서 나왔는지 모르는데, 어느 항아리인지 추정하라"

어느 항아리이든 가능성이 있으므로, *P*(①|하양), *P*(②|하양), *P*(③|하양)을 계산한 다음 가장 큰 값을 가진 항아리 번호를 선택하는 전략을 써야 한다. 이 전략을 수식으로 쓰면 식 (C.5)가 된다.

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x = \text{하양}) \qquad (C.5)$$

식 (C.5)에 베이즈 정리를 대입하면 다음과 같은 식을 얻는다.

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x = \text{하양}) = \underset{y}{\operatorname{argmax}} \frac{P(x = \text{하양}|y)P(y)}{P(x = \text{하양})}$$

y를 ①, ②, ③으로 바꾸며 계산하면 아래와 같다. 결국 $P(③|\text{하양})$이 가장 크므로 ③번 항아리에서 나왔을 가능성이 가장 크다. 이 계산에서 분모에 있는 $P(x = \text{하양})$은 생략해도 된다. 왜냐하면 $P(x = \text{하양})$은 y와 무관하여 세 계산식에서 같은 값을 가지는데, 알고자 하는 것은 세 항아리의 상대적인 우열이기 때문이다.

$$P(①|\text{하양}) = \frac{P(\text{하양}|①)\text{P}(①)}{P(\text{하양})} = \frac{\frac{9}{12}\frac{1}{8}}{\frac{43}{96}} = \frac{9}{43}$$

$$P(②|\text{하양}) = \frac{P(\text{하양}|②)\text{P}(②)}{P(\text{하양})} = \frac{\frac{5}{15}\frac{4}{8}}{\frac{43}{96}} = \frac{16}{43}$$

$$P(③|\text{하양}) = \frac{P(\text{하양}|③)\text{P}(③)}{P(\text{하양})} = \frac{\frac{3}{6}\frac{3}{8}}{\frac{43}{96}} = \frac{18}{43}$$

지금까지 푼 항아리 추정 문제에서 생각할 점이 있다. 얼핏 생각하면 ①번 항아리의 하얀 공 확률이 제일 높으므로 ①번이라고 답하려는 유혹이 있을 수 있다. 또한, ②번 항아리가 뽑힐 확률이 제일 높으므로, ②번이라고 답하려는 유혹도 있을 수 있다. 하지만 앞에서 한 것처럼, 항아리의 확률과 공의 확률을 모두 고려해야 합리적인 답을 구할 수 있다. 내기를 한다면 ③번이라고 말해야 돈을 벌 수 있다.

베이즈 정리의 의미를 더 깊이 생각해보자. 공의 색깔 x를 관찰했다면 이미 사건이 벌어진 것이다. 이때 어느 병에서 나왔는지를 나타내는 확률 $P(y|x)$는 사건 발생 후의 확률이므로 사후 확률posterior probability이라고 하고, $P(y)$는 사건 x와 무관하게 미리 알 수 있는 확률이므로 사전 확률prior

probability이라고 한다. 그리고 $P(x|y)$를 우도likelihood라고 한다. 아래 식은 이런 관계를 보여준다. 분모에 있는 $P(x)$는 종종 무시할 수 있다. 예를 들어, 하얀 공이 나온 항아리를 알아맞히는 앞 문제에서는 알고자 하는 값이 세 항아리의 상대적인 확률이지, 절대적인 값은 아니므로 분모 $P(x = \text{하양})$을 무시해도 된다.

$$\overbrace{P(y|x)}^{\text{사후확률}} = \frac{\overbrace{P(x|y)}^{\text{우도}}\ \overbrace{P(y)}^{\text{사전확률}}}{P(x)}$$

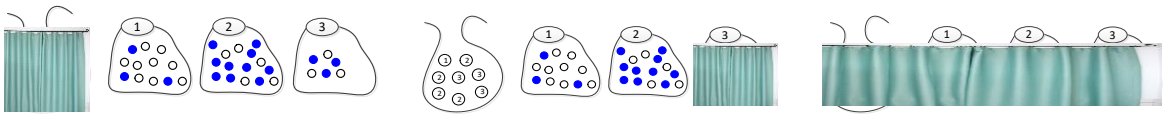
보통 조건부 확률은 |의 오른쪽에 이미 일어나서 알고 있는 사건을 쓰고 왼쪽에 추정해야 할 사건을 쓴다. 예를 들어, $P(x = \text{하양}|y = ①)$은 ①번 항아리가 뽑힌 상황에서 하얀 공이 뽑힐 확률을 추정하라는 뜻이다. 하지만 식 (C.6)과 같이 우도는 위치가 뒤바뀌었다. |의 왼쪽에 알고 있는 사건이 있고, 오른쪽에 추정해야 할 사건이 있다. 따라서 우도 추정을 역 확률 문제라 부르기도 한다. 때로는 우도를 $\mathcal{L}$ 기호를 사용하고 x와 y의 위치를 바꾸어 $\mathcal{L}(y|x)$처럼 표기한다.

$$\text{우도: } P(\underbrace{x}_{\text{알고 있음}} \mid \underbrace{y}_{\text{추정해야 함}}) = \mathcal{L}(y|x) \qquad \text{(C. 6)}$$

C.3 최대 우도

[그림 C-3]의 확률 실험에서는 주머니와 세 항아리 내부를 볼 수 있어 모든 확률 값을 아는 상태에서 확률 추론을 하였다. 단지 뽑은 카드 번호가 알려지지 않아 하얀 공이 나온 항아리를 추정하는 문제가 생겼고, 베이즈 정리를 이용하여 발생 가능성이 가장 큰 항아리를 추정하는 문제를 풀었다.

이제 [그림 C-4]처럼 일부 또는 전부가 가려진 상황에서 가려진 곳에 있는 매개변수를 추정하는, 더 복잡한 문제를 생각해보자. [그림 C-4(a)]는 카드를 담은 주머니가 가려져 있어 카드 ①, ②, ③의 확률을 추정해야 한다. 세 확률을 더하면 1이므로, 카드 ①과 ②의 확률 p_1과 p_2만 추정하면 된다. [그림 C-4(b)]에서는 ③번 항아리에 들어 있는 파란 공의 확률 q_3을 추정해야 한다. 파란 공의 확률을 구하면 하얀 공의 확률은 1-q_3으로 구할 수 있다. [그림 C-4(c)]는 전체가 가려져 있어 가장 복잡한 상황이다. 추정해야 하는 매개변수가 5개다. 추정해야 하는 매개변수의 집합은 Θ로 표기한다. [그림 C-4]에서는 편의상 주머니와 관련된 확률은 p, 병과 관련된 확률은 q로 표기하였다.



(a) $\Theta = \{p_1, p_2\}$ (b) $\Theta = \{q_3\}$ (c) $\Theta = \{p_1, p_2, q_1, q_2, q_3\}$

[그림 C-4] 매개변수가 감추어진 여러 가지 상황

실험을 여러 번 반복하여 공의 색깔, 즉 데이터 집합 X={●○○●○●○○●●○○}를 얻었다고

가정하자. 공의 순서는 중요하지 않다. [그림 C-4(b)]처럼 ③번 항아리만 가려진 상황에서 매개변수 Θ를 추정하는 문제를 생각하자. 이때 추정해야 하는 것은 파란 공의 확률 q_3이다. 파란 공의 확률을 정하고 나면 하얀 공의 확률은 1-q_3으로 구할 수 있으므로 매개변수는 2개가 아니라 1개이다. 문제를 다음과 같이 쓸 수 있다.

> "데이터 X가 주어졌을 때, X를 발생시켰을 가능성을 최대로 하는 매개변수 $\Theta = \{q_3\}$의 값을 찾아라."

이 문제는 최적화 문제로서 식 (C.7)과 같이 쓸 수 있다. $P(X|q_3)$은 식 (C.6)과 같은 형태이므로 우도이다. 이 식을 우도를 최대화하는 해를 구한다는 뜻에서 최대 우도 추정MLE; Maximum Likelihood Estimation이라고 한다.

$$\hat{q}_3 = \underset{q_3}{\operatorname{argmax}} P(X = \{●○○●○●○○●●○○\}|q_3) \qquad (C.7)$$

최대 우도를 일반적인 표기로 다시 쓰면 식 (C.8)이 된다. 매개변수는 상황에 따라 다르므로 매개변수 집합을 Θ로 표기하였다.

$$\text{최대 우도 추정:}\quad \hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} P(X|\Theta) \qquad (C.8)$$

데이터 $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$은 독립 동일 분포iid; independent and identically distributed이다. 따라서 식 (C.9)처럼 샘플을 독립적으로 다룰 수 있다. $P(X|\Theta)$를 $P_\Theta(X)$로 표기하기도 한다.

$$P(X|\Theta) = P(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n|\Theta) = \prod_{i=1}^{n} P(\mathbf{x}_i|\Theta) \qquad (C.9)$$

식 (C.9)은 수치적인 문제를 일으킬 수 있다. 샘플 개수 n은 보통 수천을 훨씬 넘는데, 확률을 n번 곱하면 너무 작은 값이 되어 컴퓨터로 계산할 때 버림이 될 가능성이 크다. 따라서 식 (C.10)의 최대 로그우도 추정maximum log likelihood estimation을 주로 사용한다. log는 단조증가 함수이므로 식 (C.8)로 구한 최적해 $\hat{\Theta}$과 식 (C.10)으로 구한 $\hat{\Theta}$은 같다.

$$\text{최대 로그우도 추정:}\quad \hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log P(\mathbb{X}|\Theta) = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(\mathbf{x}_i|\Theta) \qquad (C.10)$$

C.4 평균과 분산

데이터셋에 있는 샘플 전체를 대표할 수 있는 몇 가지 요약 정보를 추출할 수 있다면 매우 유용할 것이다. 이 절에서는 가장 널리 쓰이는 요약 정보로서 평균과 분산에 대해 설명한다.

평균과 분산

식 (C.11)은 n개의 샘플을 가진 데이터의 평균과 분산을 구하는 공식이다. x_i는 i번째 샘플을 뜻한다. 분산variance의 제곱근인 σ를 표준편차standard deviation라고 한다.

$$\left.\begin{aligned} \text{평균}\ \ \mu &= \frac{1}{n}\sum_{i=1,n} x_i \\ \text{분산}\ \ \sigma^2 &= \frac{1}{n}\sum_{i=1,n} (x_i-\mu)^2 \end{aligned}\right\} \qquad \text{(C. 11)}$$

[예시 C-1]

윷을 12번 던져 개, 걸, 모, 걸, 도, 걸, 도, 개, 개, 걸, 개, 개가 나왔다고 가정하자. 데이터는 다음과 같다. 도는 1칸, 개는 2칸, 걸은 3칸, 윷은 4칸, 모는 5칸을 갈 수 있으므로, 다섯 가지 경우를 갈 수 있는 칸 수로 대치한다.

$$X=\{x_1=2, x_2=3, x_3=5, x_4=3, x_5=1, x_6=3, x_7=1, x_8=2, x_9=2, x_{10}=3, x_{11}=2, x_{12}=2\}$$

식 (C.11)을 적용하면 다음과 같다.

$$\mu = \frac{1}{12}(2+3+5+3+1+3+1+2+2+3+2+2) = 2.4167$$
$$\sigma^2 = \frac{1}{12}((2-2.4167)^2+(3-2.4167)^2+(5-2.4167)^2+\cdots) = 1.0764$$

예시 끝.

평균 벡터와 공분산 행렬

앞에서는 확률 변수가 하나뿐인 상황에서 평균과 분산을 계산하였다. 하지만 컴퓨터 비전이 사용하는 데이터는 여러 개의 특징으로 구성된 특징 벡터이다. 따라서 특징 벡터를 확률 벡터로 간주하고 확률 벡터의 평균과 분산을 계산하는 방법으로 확장해야 한다. 특징이 d개인 데이터를 가정하고 확률 벡터를 $\mathbf{x} = (x_1, x_2, \cdots, x_d)$ 로 표기하고 i번째 샘플을 $\mathbf{x}_i$ 로 표기한다. 평균 벡터는 $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_d)$로 표기한다. 식 (C.12)는 평균 벡터를 구하는 공식이다. n은 샘플의 개수이다.

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1,n} \mathbf{x}_i \qquad \text{(C. 12)}$$

확률 벡터에는 d개의 확률 변수가 있어 분산이 조금 복잡하다. 개별 변수의 분산이 있고 서로 다른 변수 간에 공분산이 있다. 따라서 분산은 식 (C.13)과 같은 d*d 행렬을 이룬다. 이 행렬을 $\boldsymbol{\Sigma}$로 표기하고 공분산 행렬$_{\text{covariance matrix}}$이라고 한다.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & & \sigma_{2d} \\ \vdots & & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix} = \begin{pmatrix} {\sigma_1}^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & {\sigma_2}^2 & & \sigma_{2d} \\ \vdots & & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & {\sigma_d}^2 \end{pmatrix} \qquad \text{(C. 13)}$$

식 (C.14)는 공분산 행렬을 구하는 공식이다. $(\mathbf{x}_i - \boldsymbol{\mu})$는 d*1 행렬이므로 $(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{x}_i - \boldsymbol{\mu})$는 d*d 행렬이 된다.

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1,n} (\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{x}_i - \boldsymbol{\mu}) \qquad \text{(C. 14)}$$

공분산 행렬의 요소 σ_{ij}는 i번째 특징과 j번째 특징의 상호 변화 양상을 표현한다. i번째 특징이 커질 때 j번째 특징도 따라 커지면 양수를 가지게 되고, 그런 경향이 강할수록 값이 크다. 반대로 i번째 특징이 커질 때 j번째 특징이 작아지면 음수를 가진다. 예를 들어, 사람의 키라는 특징과 신발의 크기라는 특징은 양의 공분산을 가지며, 콜레스테롤 수치와 건강한 정도는 음의 공분산을 가진다.

[예시 C-2]

8명의 키와 몸무게, 학점을 조사한 결과 아래와 같은 데이터를 얻었다고 가정한다.

$$X = \{\mathbf{x}_1 = (170\ \ 60\ \ 4.1), \mathbf{x}_2 = (165\ \ 55\ \ 3.0), \mathbf{x}_3 = (174\ \ 75\ \ 2.8), \mathbf{x}_4 = (169\ \ 67\ \ 2.9), \mathbf{x}_5 = (155\ \ 49\ \ 3.1), \mathbf{x}_6 = (172\ \ 63\ \ 3.6), \mathbf{x}_7 = (166\ \ 58\ \ 3.7), \mathbf{x}_8 = (168\ \ 61\ \ 4.0)\}$$

먼저 평균 벡터를 구하면 $\boldsymbol{\mu} = (167.375\ \ 61.0\ \ 3.4)$다. 첫 번째 샘플 $\mathbf{x}_1$을 식 (C.14)에 적용하면 다음과 같다.

$$(\mathbf{x_1} - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{x_1} - \boldsymbol{\mu}) = \begin{pmatrix} 2.625 \\ -1.0 \\ 0.7 \end{pmatrix} (2.625\ \ -1.0\ \ 0.7) = \begin{pmatrix} 6.891 & -2.625 & 1.838 \\ -2.625 & 1.0 & -0.7 \\ 1.838 & -0.7 & 0.49 \end{pmatrix}$$

나머지 7개 샘플도 비슷한 계산을 한 다음, 결과를 모두 더하고 8로 나누면 다음과 같은 공분산 행렬을 얻게 된다.

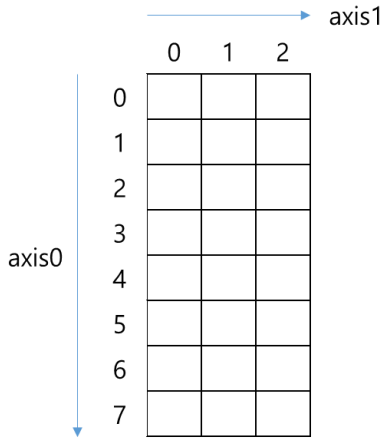
$$\boldsymbol{\Sigma} = \begin{pmatrix} 29.484 & 34.5 & 0.325 \\ 34.5 & 53.25 & -0.825 \\ 0.325 & -0.825 & 0.23 \end{pmatrix}$$

키와 몸무게 사이의 공분산 σ_{12}는 34.5로서 양수이므로 키가 커지면 몸무게도 커지는 경향을 나타낸다. 키와 학점 사이의 공분산 σ_{13}은 0.325로서 작은 양수이고, 몸무게와 학점 사이의 공분산 σ_{23}은 -0.825으로서 작은 음수이다. 데이터를 더 많이 수집하면 σ_{13}와 σ_{23}은 0에 가까워질 것이다.

예시 끝.

다음 코드는 numpy 모듈을 이용하여 [예시 C-2]를 수행한다. [1]행은 numpy 모듈을 불러온다. [2]행은 데이터를 설정한다. [3]행의 mean 함수는 평균 벡터를 계산한다. 이때 axis 인수를 생략하면 벡터를 이어 붙여 1차원 벡터로 만들고 1차원 벡터의 요소 평균을 구하기 때문에 원하는 값을 얻을 수 없다. [그림 C-5]는 axis=0 인수를 설정한 이유를 설명한다. 축에 대한 상세한 설명은 [그림 A-5]를 참조한다. [4]행은 cov 함수는 공분산 행렬을 구해준다. 이때 rowvar=False 인수를 설정했는데, 그렇게 해야 샘플이 행에 배치되어 있다고 해석하여 axis0을 기준으로 공분산 행렬을 계산해준다. 만일 rowvar 인수를 생략하면, 기본값이 True이므로 8차원 샘플이 3개 있다고 간주하고 8*8 공분산 행렬을 계산한다. 세번째 인수 bias=True를 생략하면 기본값으로 bias=True를 사용하게 되어 식 (C.14)에서 n 대신 $n-1$로 나누는 계산을 한다.

```
In [1]: import numpy as np
In [2]: X=np.array([[170,60,4.1], [165,55,3.0], [174,75,2.8], [169,67,2.9], [155,49,3.1], [172,63,3.6],
[166,58,3.7], [168,61,4.0]])
In [3]: m=np.mean(X,axis=0)
In [4]: cv=np.cov(X,rowvar=False,bias=True)
In [5]: print(m)
[167.375    61.          3.4    ]
In [6]: print(cv)
[[29.484375 34.5         0.325    ]
 [34.5       53.25      -0.825    ]
 [ 0.325     -0.825      0.23     ]]
```



[그림 C-5] 2차원 배열의 축 번호 매기기

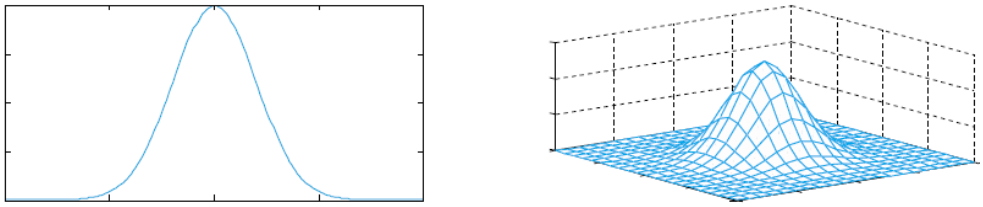
C.5 유용한 확률분포

확률 분포로 널리 사용되는 함수로 가우시안 분포, 베르누이 분포, 이항 분포가 있다. 이들 분포는 일정한 모양을 가지는데, 1개 또는 2개의 매개변수로 모양을 조절할 수 있다. 먼저 가장 널리 쓰이는 가우시안 분포부터 살펴보자.

가우시안 분포

가우시안 분포Gaussian distribution는 평균과 분산을 나타내는 2개의 매개변수 μ와 σ^2으로 규정하며, 식 (C.15)로 정의한다. 정규 분포normal distribution라고도 하며 $N(x;\mu,\sigma^2)$과 같이 표기하는데, 세미콜론 앞에는 확률 변수, 세미콜론 뒤에는 매개변수를 적는 표기법을 사용한다. 때때로 확률 변수를 생략하고 $N(\mu,\sigma^2)$과 같이 쓰기도 한다. [그림 C-6(a)]는 가우시안 분포이다. 최댓값을 가지는 지점이 μ이다. σ^2은 분포의 퍼진 정도를 나타내는데 σ^2이 클수록 봉우리의 높이가 낮고 좌우로 멀리 퍼진다. 자연에서 측정한 여러 가지 데이터가 가우시안 분포와 유사하다. 사람의 키, 영상에 나타나는 잡음, 수능시험 성적 등을 예로 들 수 있다.

$$N(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \qquad (C.15)$$



(a) 1차원 (b) 2차원

[그림 C-6] 가우시안 분포

특징 벡터가 d차원인 가우시안 분포는 평균 벡터 $\boldsymbol{\mu}$와 공분산 행렬 $\boldsymbol{\Sigma}$라는 매개변수로 모양이 규정된다. 식 (C.16)은 다차원 가우시안 분포이다. [그림 C-6(b)]는 2차원 가우시안 분포이다. 1차원과 마찬가지로 $\boldsymbol{\mu}$는 최댓값을 가진 지점이며, $\boldsymbol{\Sigma}$는 퍼진 정도를 나타낸다.

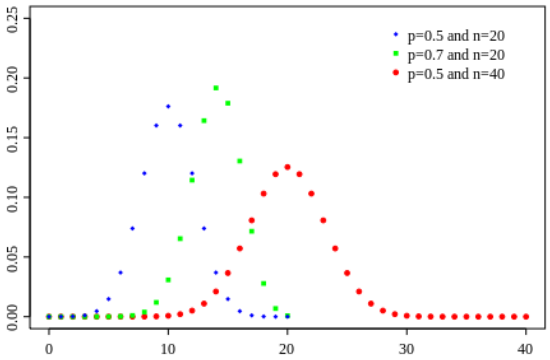
$$N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{|\boldsymbol{\Sigma}|}\sqrt{(2\pi)^d}}exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \qquad (C.16)$$

베르누이 분포와 이항 분포

확률 변수 x가 1(성공) 또는 0(실패)의 두 가지 값만 가질 수 있는 이진 변수이고, 성공 확률은 p이며 실패 확률은 1-p인 분포를 베르누이 분포Bernoulli distribution라고 한다. 베르누이 분포는 식 (C.17)로 정의할 수 있다. 매개변수는 p 하나이다.

$$Ber(x;p)=p^x(1-p)^{1-x}=\begin{cases}p, & x=1\text{일 때}\\ 1-p, & x=0\text{일 때}\end{cases} \qquad (C.17)$$

성공 확률이 p인 베르누이 실험을 n번 수행할 때 성공할 회수가 x인 확률 분포를 이항 분포binomial distribution라고 한다. 매개변수는 p와 n으로 2개이다. [그림 C-7]은 매개변수에 따른 이항 분포의 모양이다. 예를 들어, 녹색 곡선은 p=0.7이고 n=20인 경우의 이항 분포인데, x=14일 확률이 가장 높고 좌우로 멀어질수록 값이 낮아진다.



[그림 C-7] 이항 분포

이항 분포는 식 (C.18)로 정의한다. C_n^x는 n개의 서로 다른 물건에서 x개를 뽑는 가지 수를 나타내는 이항 계수로, $C_n^x = \frac{n!}{x!(n-x)!}$이다. 예를 들어, [그림 C-7]의 p=0.7, n=20인 경우의 이항 분포에서 x=14일 확률은 $\frac{20!}{14!6!}0.7^{14}0.3^6 = 0.1916$이고, x=20일 확률은 $\frac{20!}{20!0!}0.7^{20}0.3^0 = 0.0007979$이다.

$$B(x; n, p) = C_n^x p^x (1-p)^{1-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{1-x} \qquad \text{(C. 18)}$$

베르누이 분포는 n=1일 때의 이항 분포로서 이항 분포의 특수한 형태임을 쉽게 알 수 있다. 즉, $Ber(x; p) = B(x; 1, p)$이다.

C.6 정보 이론

윷놀이를 하다가 '모가 나왔다'와 '개가 나왔다'라는 메시지를 들었을 때 어느 것이 더 많은 정보를 전달할까? '고비 사막에 눈이 왔다'와 '대관령에 눈이 왔다'라는 뉴스를 들었을 때 어느 쪽이 더 많은 정보를 전달할까? 분명 두 경우 모두 앞쪽 메시지가 더 많은 정보를 전달한다. 그렇다면 메시지가 가진 정보량을 수량화할 수 있을까? 가령 '고비 사막에 눈이 왔다'라는 메시지는 '대관령에 눈이 왔다'라는 메시지보다 1.95배 정보량이 많다고 말할 수 있을까? 정보 이론은 이와 같은 질문에 명쾌하게 답을 준다.

자기 정보와 엔트로피

정보 이론에서는 메시지의 정보량을 확률을 사용하여 측정한다. 확률이 낮은 사건일수록 더 많은 정보를 전달한다. '고비 사막에 눈이 왔다'라는 뉴스를 들으면 사람들은 '놀라운데'라거나 '굉장한 뉴스네' 라는 반응을 보인다. 많은 정보를 취득한 셈이다. 반면, '대관령에 눈이 왔다'라는 뉴스를 들으면 '또 왔네'라거나 '올 때 됐지'라는 반응을 보인다.

이렇게 근거를 마련하였으니, 이제는 어떤 사건이 일어날 확률을 추정할 수 있다면 그 사건에 대한 정보량을 측정할 수 있다. 확률 변수를 x라 하고 x의 정의역을 $\{a_1, a_2, \ldots, a_k\}$라 하자. 사건 a_i가 발생하면 해당 메시지가 전달되므로 사건과 메시지를 같은 의미로 섞어 사용한다. 정보 이론에서는 식 (C.19)를 이용하여 사건 a_i의 정보량 $h(a_i)$를 측정한다. 이 정보량을 자기 정보$_{\text{self-information}}$라고 한다. 이 식에서는 $P(x{=}a_i)$를 줄여 $P(a_i)$로 썼다. 가령 윷놀이에서 모가 나온 사건의 자기 정보는 $h(\text{모}) = -\log_2 \frac{1}{16} = 4$이고, 개가 나온 사건의 자기 정보는 $h(\text{개}) = -\log_2 \frac{6}{16} = 1.415$이다. 따라서 자기 정보라는 기준에 따르면 모가 나왔다는 메시지는 개가 나왔다는 메시지보다 2.8269배만큼 정보량이 많다고 할 수 있다.

$$\mathrm{h}(a_i) = -\log_2 P(a_i) \quad \text{또는} \quad \mathrm{h}(a_i) = -\log_e P(a_i) \qquad \text{(C. 19)}$$

밑이 2인 로그 함수를 사용하는 경우 자기 정보의 단위는 비트$_{\text{bit}}$이다. 확률이 1/2일 때 1비트의 정보량을 가진다. 모가 나온 사건의 자기 정보는 4비트이다. 밑이 e인 자연 로그를 사용하는 경우

확률이 1/e=0.3679일 때 1만큼의 정보량을 가지는데, 이때 단위는 나츠$_{nat}$이다. 모가 나온 사건의 자기 정보는 2.7726나츠이다.

자기 정보가 특정 사건 a_i의 정보량을 측정하는 반면, 엔트로피$_{entropy}$는 확률 분포의 무질서도 또는 불확실성$_{uncertainty}$을 측정한다. 식 (C.20)과 식 (C.21)은 각각 이산 확률 분포와 연속 확률 분포의 엔트로피를 정의한다. 단위는 자기 정보와 마찬가지로 밑이 2이면 비트, 밑이 e이면 나츠이다. $\mathbb{R}$은 실수 공간을 뜻한다.

$$\text{이산 확률분포 } \mathrm{H}(x) = -\sum_{i=1,k} P(a_i)\log_2 P(a_i) \quad \text{또는} \quad \mathrm{H}(x) = -\sum_{i=1,k} P(a_i)\log_e P(a_i) \quad (\mathrm{C}.20)$$

$$\text{연속 확률분포 } \mathrm{H}(x) = -\int_{\mathbb{R}} P(a)\log_2 P(a) \quad \text{또는} \quad \mathrm{H}(x) = -\int_{\mathbb{R}} P(a)\log_e P(a) \quad (\mathrm{C}.21)$$

[예시 C-3]

윷을 나타내는 확률변수를 x라 할 때 x의 엔트로피는 다음과 같다.

$$\mathrm{H}(x) = -\left(\tfrac{4}{16}\log_2\tfrac{4}{16} + \tfrac{6}{16}\log_2\tfrac{6}{16} + \tfrac{4}{16}\log_2\tfrac{4}{16} + \tfrac{1}{16}\log_2\tfrac{1}{16} + \tfrac{1}{16}\log_2\tfrac{1}{16}\right) = 2.0306\text{비트}$$

주사위는 눈이 6개인데 모두 1/6이라는 균일한 확률을 가진다. 이 경우 엔트로피를 계산하면 다음과 같다.

$$\mathrm{H}(x) = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 2.585\text{비트}$$

예시 끝.

앞의 예시를 보면 윷보다 주사위의 엔트로피가 더 높다. 왜 그럴까? 주사위는 모든 사건이 동일한 확률을 가진다. 즉, 어떤 사건이 일어날지 윷보다 예측하기 어렵다. 다른 말로 표현하면 주사위가 윷보다 더 무질서하고 불확실성이 더 크다. 따라서 엔트로피가 더 높다. 사실 주사위처럼 모든 사건이 동일한 확률을 가질 때 엔트로피가 최고이다.

정의역의 크기가 크면 엔트로피도 커지는데, 주사위는 정의역이 6개의 요소를 가진 데 반해 윷은 5개 요소이므로 정의역이 커서 엔트로피가 커진 탓도 있다. 정오각형 주사위를 가정하여 윷과 정의역의 크기를 맞추어도, 엔트로피는 $-\frac{1}{5}\log_2\left(\frac{1}{5}\right) * 5 = 2.3219$가 되어 여전히 주사위의 엔트로피가 윷보다 높다.

교차 엔트로피와 쿨백-라이블러 발산

식 (C.20)으로 정의되는 엔트로피는 한 확률 분포의 무질서 정도를 측정한다. 그런데 때로 두 확률 분포 간의 엔트로피를 측정할 필요가 있다. 식 (C.22)는 서로 다른 두 확률 분포 P와 Q 사이의 교차 엔트로피$_{cross\ entropy}$를 정의한다. 이때 두 확률 분포는 같은 확률 벡터에 대해 정의되어 있어야 한다.

$$\mathrm{H}(P,Q) = -\sum_{x} P(x)\log_2 Q(x) = -\sum_{i=1,k} P(a_i)\log_2 Q(a_i) \qquad (\mathrm{C}.22)$$

식 (C.22)를 다음과 같이 전개할 수 있다.

$$\mathrm{H}(P,Q) = -\sum_{x} P(x)\log_2 Q(x)$$
$$= -\sum_{x} P(x)\log_2 P(x) + \sum_{x} P(x)\log_2 P(x) - \sum_{x} P(x)\log_2 Q(x)$$
$$= \mathrm{H}(P) + \sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)}$$

마지막 식의 두 번째 항을 쿨백-라이블러 발산Kullback-Leibler divergence이라 하고, 줄여서 KL 발산이라고 부른다. KL 발산은 식 (C.23)과 같이 정의한다. 위의 유도식에 따르면 교차 엔트로피와 KL 발산은 식 (C.24)를 만족한다.

$$\mathrm{KL}(P \parallel Q) = \sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)} \qquad (\mathrm{C}.23)$$

P와 Q의 교차 엔트로피 $\mathrm{H}(P,Q) = \mathrm{H}(P) + \sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)} = P$의 엔트로피 $+ P$와 Q의 KL 발산 $(\mathrm{C}.24)$

식 (C.23)으로 정의되는 KL 발산은 두 확률 분포가 얼마나 다른지 측정한다. P와 Q가 같을 때 0이 된다는 사실을 쉽게 알 수 있다. 따라서 거리 개념을 내포한다. 하지만 $\mathrm{KL}(P \parallel Q) \neq \mathrm{KL}(Q \parallel P)$이므로 엄밀한 수학적 정의에 따르면 거리가 아니므로 발산이라고 부른다.

[예시 C-4]

정상적인 주사위와 찌그러진 주사위가 있는데, 정상적인 주사위의 확률 분포는 P, 찌그러진 주사위의 확률 분포는 Q를 따르며, P와 Q가 다음과 같이 분포한다고 가정하자.

$$P(1) = \frac{1}{6}, P(2) = \frac{1}{6}, P(3) = \frac{1}{6}, P(4) = \frac{1}{6}, P(5) = \frac{1}{6}, P(6) = \frac{1}{6}$$
$$Q(1) = \frac{3}{12}, Q(2) = \frac{1}{12}, Q(3) = \frac{1}{12}, Q(4) = \frac{1}{12}, Q(5) = \frac{3}{12}, Q(6) = \frac{3}{12}$$

확률 분포 P와 Q 사이의 교차 엔트로피와 KL 발산은 다음과 같다.

$$\mathrm{H}(P,Q) = -\left(\frac{1}{6}\log_2\frac{3}{12} + \frac{1}{6}\log_2\frac{1}{12} + \frac{1}{6}\log_2\frac{1}{12} + \frac{1}{6}\log_2\frac{1}{12} + \frac{1}{6}\log_2\frac{3}{12} + \frac{1}{6}\log_2\frac{3}{12}\right) = 2.7925$$

$$\mathrm{KL}(P \parallel Q) = \frac{1}{6}\log_2\frac{2}{3} + \frac{1}{6}\log_2 2 + \frac{1}{6}\log_2 2 + \frac{1}{6}\log_2 2 + \frac{1}{6}\log_2\frac{2}{3} + \frac{1}{6}\log_2\frac{2}{3} = 0.2075$$

[예시 C-3]에서 P의 엔트로피 $\mathrm{H}(P)$는 2.585이었다. 따라서 식 (C.24)가 성립함을 알 수 있다.

예시 끝.