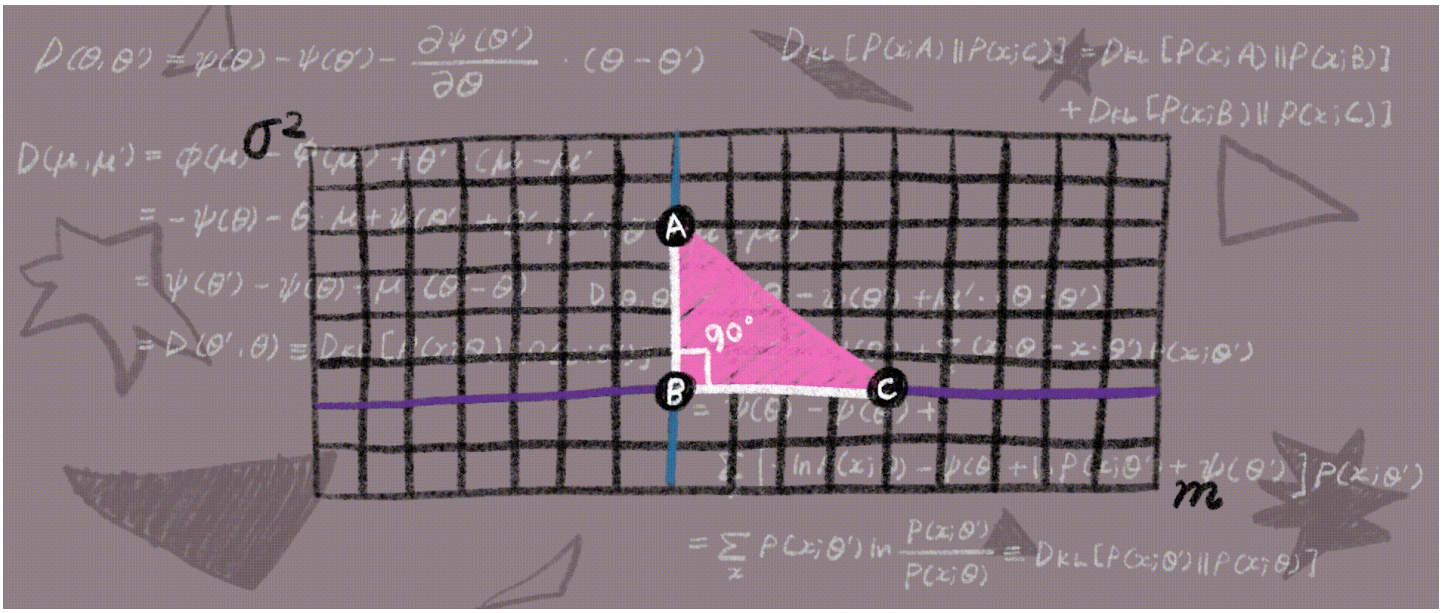
인공지능

정보기하학과 머신러닝 [1]: 확률모형 사이의 거리

2022년 10월 11일

조정효



머신러닝의 모형은 크게 분류모형과 생성모형으로 구별된다. 이들은 확률을 써서 일반적으로 표현할 수 있다. 분류모형은 입력 x에 대한 출력 y를 표현하는 조건부 확률 $P(y|x;\theta)$에 해당한다. 여기서 θ는 이 확률모형의 매개변수parameter를 뜻한다. 생성모형은 주어진 데이터 x가 나타날 확률 $P(x;\theta)$에 해당한다. 확률모형을 이용해서 확률이 높은 x를 선택하는 행위가 바로 샘플 생성이 된다. 여기서 이들 확률모형을 그래프로 표현하는 것이 바로 신경망 기반의 머신러닝이다.

데이터의 분포 $\hat{P}(x)$가 주어졌을 때 이를 설명하는 확률모형 $P(x;\theta)$의 매개변수 θ를 찾는 행위를 우리는 "학습"이라고 부른다. 여기서 우리는 자연스럽게 확률들 또는 확률모형들 사이의 거리를 정의할 필요가 생기고, 거리를 줄이는 최적화를 하게 된다. 앞으로 "정보기하학과 머신러닝"이라는 주제로 세 번에 걸친 연재를 통해서 확률모형들 사이의 거리와 최적화에 대해서 소개해 보려고 한다.

점들 사이의 거리를 정의하고 이로부터 피타고라스 정리가 성립하는 유클리드 공간에 우리는 익숙하다. 과연 확률모형들이 살고 있는 공간에서 거리는 어떻게 정의하는 것이 자연스러울까? 그 공간에서도 피타고라스 정리 같은 것이 존재할까? 이번 첫 연재에서는 이 질문에 대한 대답을 얻어 보겠다.

우리는 특히 지수족exponential family의 확률모형을 중심으로 이야기를 풀어 보려고 한다.

$$P(x;\theta) = \frac{\exp(-\theta \cdot x)}{Z}$$

여기서 모든 x에 대한 정규화 상수 $Z \equiv \sum_x \exp(-\theta \cdot x)$는 통계물리에서는 분배함수partition function라고 불린다. 지수분포 또는 볼츠만분포가 바로 이런 형태를 띤다. 그리고 정규분포도 조금 변형을 해보면 지수족이라는 것을 알 수 있다.

$$\begin{aligned} P(y) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{(y-m)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{m}{\sigma^2}y - \frac{1}{2\sigma^2}y^2 - \frac{m^2}{2\sigma^2} - \frac{1}{2}\ln 2\sigma^2 - \frac{1}{2}\ln\pi\right] \\ &= \exp\left[-\theta_1 x_1 - \theta_2 x_2 - \psi(\theta_1,\theta_2)\right] = P(x;\theta) \end{aligned}$$

여기서 $x = (x_1, x_2) = (y, y^2)$ 그리고 $\theta = (\theta_1, \theta_2) = (-m/\sigma^2, 1/2\sigma^2)$는 2차원 벡터이다. 정규화 상수 Z는 지수 안에 $\psi(\theta_1,\theta_2) = \ln Z = \theta_1^2/(4\theta_2) - (1/2)\ln\theta_2 + (1/2)\ln\pi$로 넣었다. 사실 이런 대표적인 분포들을 포함해서 우리가 알고 있는 대부분의 분포들이 지수족이다.[1] 이것은 우연은 아니고 정보이론을 통해서 이해할 수 있는데 이번 글에서는 생략하겠다.

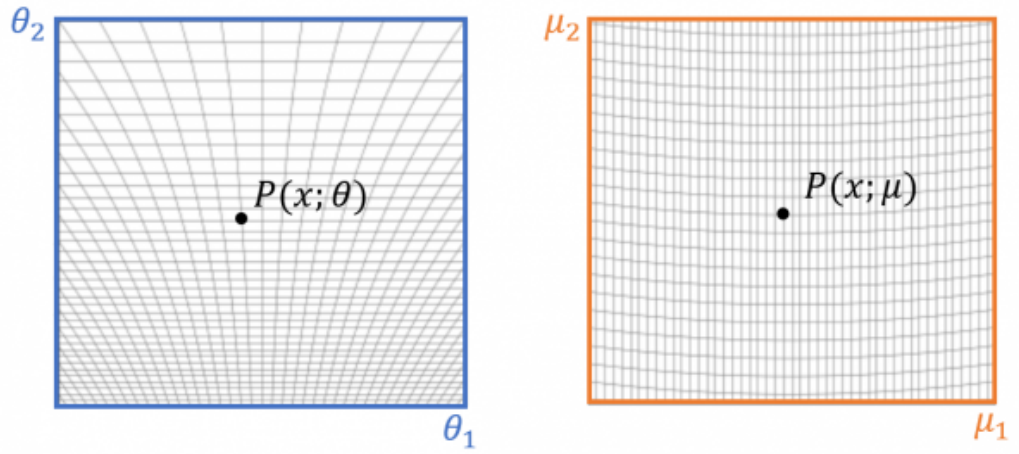
위에서 정의한 정규화 상수 또는 분배함수 Z에 로그를 취하면 누적생성함수cumulant generating function $\psi(\theta) \equiv \ln Z(\theta)$가 된다. 즉 누적생성함수를 미분하면 누적cumulant을 쉽게 얻을 수 있다. 가령 한번 미분하면 1차 누적에 해당하는 평균값을 얻는다.

$$\begin{aligned} -\frac{\partial\psi}{\partial\theta_i} &= -\frac{1}{Z}\frac{\partial Z}{\partial\theta_i} = \sum_x \frac{x_i \exp(-\theta\cdot x)}{Z} \\ &= \sum_x x_i P(x;\theta) = \mathbb{E}[x_i] \equiv \mu_i \end{aligned}$$

따라서 $P(x;\theta)$에 대한 모든 누적을 구할 수 있는 $\psi(\theta)$를 안다는 것은 확률모형 $P(x;\theta)$을 안다는 사실과 동등하다. $\psi(\theta)$는 국소적으로 오목함수이므로, 각 θ 위치에서 이 함수의 기울기 μ는 구별되는 값을 가지게 된다. 벡터 μ의 각 성분은 $\mu_i = -\partial\psi/\partial\theta_i$로 정의된다. 이제 우리는 θ를 변수로 가지는 함수 $\psi(\theta)$ 대신 μ를 변수로 가지는 함수 $\phi(\mu)$를 상상해 보자. 이는 추상적인 변수 θ로 표현되는 확률모형 $P(x;\theta)$ 대신, x의 평균값에 해당하는 관찰가능한 변수 μ로 표현되는 확률모형 $P(x;\mu)$를 사용하겠다는 의도이다. θ와 켤레conjugate관계에 있는 μ를 변수로 가지는 함수 $\phi(\mu)$는 르장드르 변환을 통해 얻을 수 있다.

$$\psi(\theta) + \phi(\mu) = -\theta \cdot \mu$$

이제 확률모형들이 사는 공간을 상상해보자. 본래공간primal space에서는 각 지점의 확률모형이 θ로 표현되고, 쌍대공간dual space에서는 각 지점의 확률모형이 μ로 표현된다.[그림1]



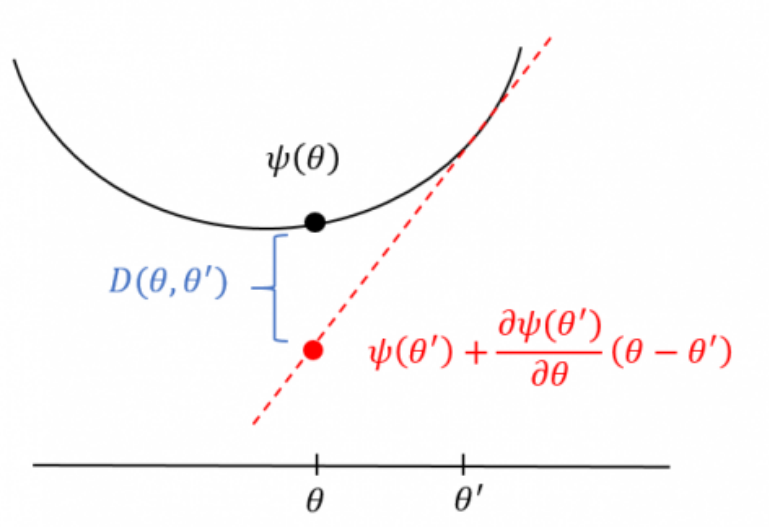
[그림1] 본래공간(왼쪽)과 쌍대공간(오른쪽)에서의 확률모형

조정효

다음으로 이 공간에서 확률모형들 사이의 거리를 정의해 보자. $\psi(\theta)$함수는 θ공간에서는 국소적으로 오목함수convex function이므로 브레그만 거리Bregman divergence를 자연스럽게 생각해 볼 수 있다. 브레그만 거리는 다음처럼 정의가 된다.

$$D(\theta, \theta') = \psi(\theta) - \psi(\theta') - \frac{\partial \psi(\theta')}{\partial \theta} \cdot (\theta - \theta')$$

즉 θ 위치에서의 함수값과, θ' 위치로부터의 1차 테일러 근사값 사이의 차이에 해당한다.[그림2] 오목함수에서는 $\theta = \theta'$가 되는 곳에서만 거리가 $D = 0$이 된다.



[그림2] 브레그만 거리

조정효

이제부터는 지수족 확률모형의 브레그만 거리를 계산해 보겠다.

$$
\begin{aligned}
D(\theta,\theta') &= \psi(\theta)-\psi(\theta')+\mu'\cdot(\theta-\theta') \\
&= \psi(\theta)-\psi(\theta')+\sum_x (x\cdot\theta - x\cdot\theta')P(x;\theta') \\
&= \psi(\theta)-\psi(\theta')+ \\
&\quad \sum_x \Big[-\ln P(x;\theta)-\psi(\theta)+\ln P(x;\theta')+\psi(\theta')\Big]P(x;\theta') \\
&= \sum_x P(x;\theta')\ln\frac{P(x;\theta')}{P(x;\theta)} \equiv D_{KL}[P(x;\theta')||P(x;\theta)]
\end{aligned}
$$

위 계산에서 둘째 줄에서는 $\mu'=\sum_x xP(x;\theta')$, 셋째 줄에서는 $P(x;\theta)=\exp(-\theta\cdot x-\psi(\theta))$, 넷째 줄에서는 $\sum_x P(x;\theta')=1$을 이용했다. 재미있게도 이렇게 구한 브레그만 거리는 정보이론에서 상대적 엔트로피로 불리는 쿨벡-라이블러 거리Kullback-Liebler divergence가 된다.

이번에는 쌍대공간에서의 브레그만 거리도 같은 방식으로 유도해 보겠다.

$$
\begin{aligned}
D(\mu,\mu') &= \phi(\mu)-\phi(\mu')+\theta'\cdot(\mu-\mu') \\
&= -\psi(\theta)-\theta\cdot\mu+\psi(\theta')+\theta'\cdot\mu'+\theta'\cdot(\mu-\mu') \\
&= \psi(\theta')-\psi(\theta)+\mu\cdot(\theta'-\theta) \\
&= D(\theta',\theta) \equiv D_{KL}[P(x;\theta)||P(x;\theta')]
\end{aligned}
$$

위 계산에서 둘째 줄에서는 르장드르 변환 $\psi(\theta)+\phi(\mu)=-\theta\cdot\mu$을 이용하였다.

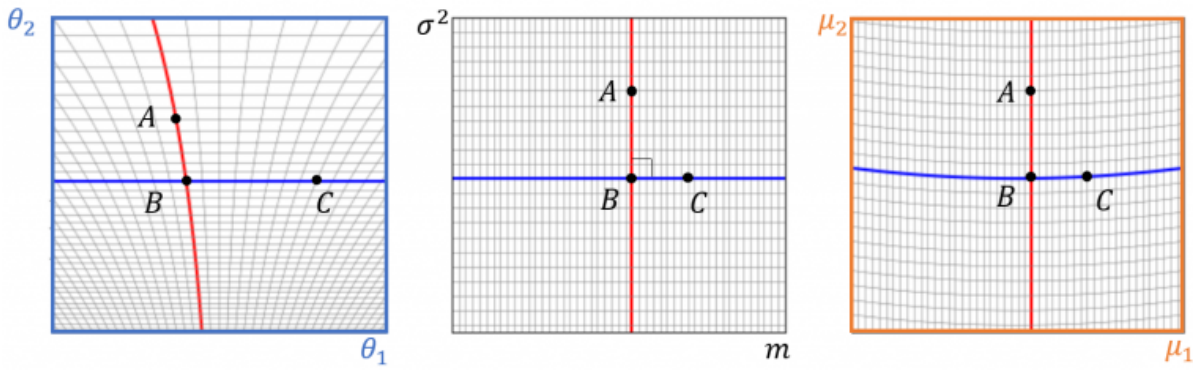
이렇게 정의된 지수족 함수의 브레그만 거리는 일반화된 피타고라스 정리를 만족한다. 가령 평균과 분산이 (m_1,σ_1)인 정규분포 A와 (m_2,σ_2)인 정규분포 C를 생각해 보자. 이때 이 두 확률모형과 피타고라스 정리를 만족하는 확률모형은 평균과 분산이 (m_1,σ_2)인 정규분포 B가 됨을 간단히 보일 수 있다. 즉 다음 식이 성립함을 뜻한다.

$$\begin{aligned}D_{KL}[P(x;A)||P(x;C)] &= D_{KL}[P(x;A)||P(x;B)]\\ &+ D_{KL}[P(x;B)||P(x;C)]\end{aligned}$$

정규분포의 정의를 이용해서 위 등식을 보이는 것은 숙제로 남기겠다.

확률모형들이 살고 있는 공간에서 "직교함"의 의미를 살펴보자. 여기서 재미난 사실은 확률모형들 사이의 측지선인 AB와 BC가 θ로 구성된 본래공간에서 직교하지 않고, μ로 구성된 쌍대공간에서도 직교하지 않는다.[그림3] 쌍대공간의 AB와 본래공간의 BC가 서로 직교하는 것을 "쌍대적으로 평평한dually flat" 공간의 직교라고 한다.[2]

이렇게 본래공간과 쌍대공간을 같이 염두에 두면서 직교한 성질을 이용하면, 투사projection를 이용할 수 있다. 우리는 2차원 유클리드 공간에서 투사를 통해서 (x,y)에서 가장 가까운 x축 위의 점 $(x,0)$을 쉽게 얻을 수 있다. 확률모형이 살고 있는 쌍대적으로 평평한 공간에서도 투사를 할 수 있다. 평균과 분산이 (m_1,σ_1)인 정규분포 확률모형에서 분산만 바꾸는 것을 허락하면서 (m_2,σ_2) 모형에 가장 가까운 모형을 찾아 보자. 투사를 통해 얻는 모형은 (m_1,σ_2)이 됨을 바로 알 수 있다.



[그림3] 쌍대적으로 평평한 공간

황준오

이번 연재에서는 확률모형 가운데 가장 보편적인 지수족을 중심으로 모형 간의 거리를 자연스럽게 정의하면 정보이론에 뿌리를 둔 쿨벡-라이블러 거리를 자연스럽게 유도해 낼 수 있음을 살펴 보았다.

그리고 이 거리를 사용하면 유클리드 공간의 피타고라스 정리와 투사의 성질도 일반화해서 사용할 수 있음을 확인하였다. 다음 연재에서는 쿨벡-라이블러 거리가 데이터의 압축에 무관한 충분통계량과 관련이 있는 f-거리가 됨을 소개해 보겠다.

참고문헌

1. https://en.wikipedia.org/wiki/Exponential family
2. Shun-ichi Amari, “Information geometry and its applications”, Springer (2016).