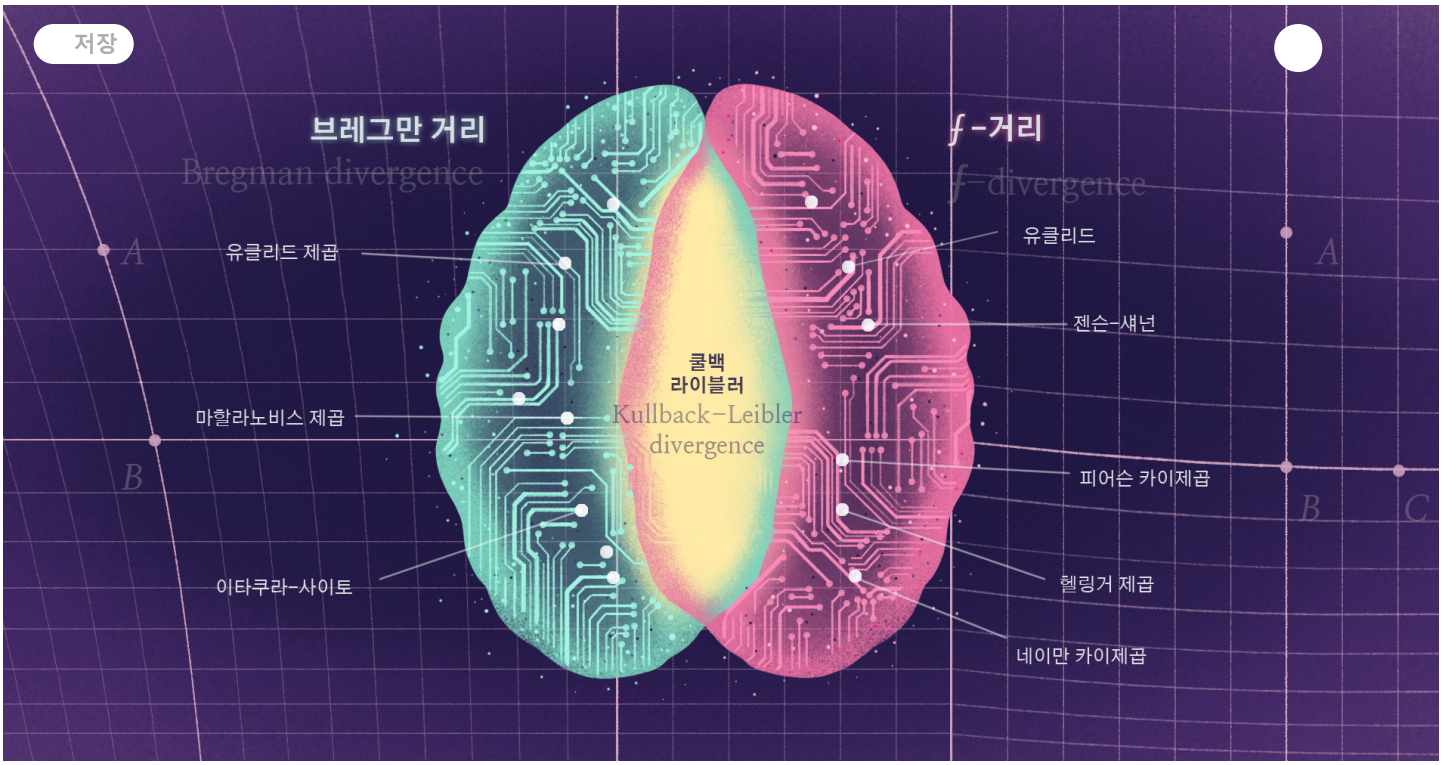
인공지능

정보기하학과 머신러닝 [2]: 충분통계량과 f-거리

2022년 12월 27일

조정효



지난 글(클릭 시 1편으로 연결)에서 우리는 확률모형 사이의 거리를 어떻게 정의할지 생각해 보았다. 특히 쿨백-라이블러 거리Kullback-Leibler divergence는 일반화된 피타고라스 정리를 만족하고, 투사projection라는 성질도 가지고 있음을 확인하였다. 이번 글에서는 쿨백-라이블러 거리의 또 다른 면모를 소개해 보려고 한다.

충분통계량sufficient statistic이라고 들어 보셨는지? 정해진 확률 θ_0과 θ_1을 따라 0과 1이 나오는 베르누이 시행을 생각해보자. (1, 1, 0, 1, ..., 0)와 같은 10번의 시행 데이터가 있다. 이 결과를 보고 모형의 매개변수인 θ_0과 θ_1을 추론해보자. 이 추론 문제에서는 사실 10개의 이진 숫자들 $x = (x_1, x_2, \cdots, x_{10})$을 모두 기억할 필요는 없고, 10번의 시행에서 몇 번 1이 나왔는지 $T(x) = x_1 + x_2 + \cdots + x_{10}$를 아는 것으로 충분하다. 이처럼 x_i의 모분포를 추정하는데 필요한 모든 정보가 들어 있는 통계량 $T(x)$를 충분통계량[1]이라고 부른다. 포아송분포의 경우는 시행 데이터 $x = (x_1, x_2, \cdots, x_N)$의 평균 $T(x) = (x_1 + x_2 + \cdots + x_N)/N$이 포아송분포의 매개변수를 추론하는데 필요한 충분통계량이 된다. 그리고 정규분포의 경우는 데이터의 평균과 분산이 충분통계량이 된다.

정리를 해보면, 데이터 x에서 모형의 매개변수 θ를 추정하는데 필요한 모든 정보는 충분통계량 $T(x)$에 들어 있고, x의 나머지 정보는 θ와 무관한 것이다. 이 진술을 확률의 언어로 표현해보면 다음과 같이 쓸 수 있다.

$$\begin{aligned} P(x|\theta) &= P\big(x, T(x)|\theta\big) \\ &= P(T(x)|\theta)P(x|T(x),\theta) \\ &= P(T(x)|\theta)P(x|T(x)). \end{aligned} \tag{1}$$

첫째 줄에서는 $T(x)$는 주어진 x에서 온전히 얻을 수 있는 정보임을 이용해서 등식을 만들었고, 둘째 줄에서는 확률곱의 성질을 이용했고, 셋째 줄에서는 충분통계량 $T(x)$가 주어진 상황에서는 x의 나머지 정보는 θ와 무관함을 이용했다. 그래프로 표현해보면, $x \leftrightarrow T(x) \leftrightarrow \theta$의 관계가 있는 세 변수 사이에서 $T(x)$가 고정이 되면 x와 θ는 서로 독립이 된다.

충분통계량은 왜 언급했을까? 이는 확률모형의 변수변환을 생각해보기 위함이다. $P(x;\theta)$와 $P(x;\theta')$의 거리는 변수 x를 변수 $y=T(x)$로 바꾸면 어떻게 될까? 일반적으로 데이터는 변환을 거치면 항상 정보를 잃게 된다. 따라서 해상도가 낮아진 변환된 변수를 통해서 모형들 사이의 거리를 보면 원래 거리보다 줄어들게 될 것이다. 이 정보잃음 information monotonicity을 수식으로 표현하면 다음과 같다.

$$D[P(x;\theta)||P(x;\theta')] \geq D[P(y;\theta)||P(y;\theta')]. \tag{2}$$

여기서 등식이 되는 경우는 변환 후에도 θ를 추정하는데 아무런 정보도 잃지 않는 $y=T(x)$가 x에 대한 충분통계량이 될 때이다.

다음과 같은 형태의 거리 D_f를 정의해 보자.

$$D_f(\theta,\theta') \equiv \sum_x P(x;\theta) f\left[\frac{P(x;\theta')}{P(x;\theta)}\right] \tag{3}$$

여기서 $f(u)$는 $f(1)=0$을 만족하는 임의의 오목함수이다. 이 경우 같은 모형 사이의 거리는 $D_f(\theta,\theta)=0$을 만족하게 된다. 이런 거리 D_f를 f-거리라고 부른다. 이제 f-거리가 충분통계량 $y=T(x)$에 대해서 어떻게 불변이 되는지 살펴보자. 위 식 (1)을 이용하면 다음 비례관계를 얻을 수 있다.

$$\frac{P(x;\theta')}{P(x;\theta)} = \frac{P(y;\theta')}{P(y;\theta)}. \tag{4}$$

이 관계를 이용하면 f-거리는 충분통계량에 해당하는 변수변환에 대해서 불변임을 확인할 수 있다.

$$\begin{aligned} D_f(\theta,\theta') &= \sum_x P(x;\theta) f\left[\frac{P(x;\theta')}{P(x;\theta)}\right] \\ &= \sum_y \sum_{x\in y} P(x;\theta) f\left[\frac{P(x;\theta')}{P(x;\theta)}\right] \\ &= \sum_y P(y;\theta) f\left[\frac{P(y;\theta')}{P(y;\theta)}\right]. \end{aligned} \tag{5}$$

여기서 $P(y;\theta) = \sum_{x\in y} P(x;\theta)$를 이용하였다.

이제 두 분포 함수 $p(x)$와 $q(x)$사이에 정의된 몇 가지 f-거리를 살펴보자.[2]

	$f(u)$	$D_f(p,q)$
Euclidean distance	$\lvert u-1\rvert$	$\sum_x \lvert p(x)-q(x)\rvert$
Pearson χ^2-divergence	$\frac{1}{2}(u-1)^2$	$\frac{1}{2}\sum_x \frac{(p(x)-q(x))^2}{p(x)}$
Neyman χ^2-divergence	$\frac{1}{2}\frac{(u-1)^2}{u}$	$\frac{1}{2}\sum_x \frac{(p(x)-q(x))^2}{q(x)}$
Kullback-Leibler divergence	$u-1-\ln u$	$\sum_x p(x)\ln\frac{p(x)}{q(x)}$
Kullback-Leibler divergence	$u\ln u-(u-1)$	$\sum_x q(x)\ln\frac{q(x)}{p(x)}$

이제 f-거리가 충분통계량에 대해 실제로 불변인지 우리가 관심을 가지고 있는 쿨백-라이블러 거리를 통해서 살펴보자. 확률 θ_0와 θ_1을 따라 0과 1이 나오는 베르누이 시행을 다시 생각해보자. N번의 시행 데이터를 보고 매개변수 θ_0와 θ_1를 추론해보자. 이 문제에서 두 모형 (θ_0,θ_1)와 (θ_0',θ_1') 사이의 f-거리는 $x=(x_1,x_2,\cdots,x_N)$를 알 때

나, 이것의 충분통계량인 $y = x_1 + x_2 + \cdots + x_N$를 알 때나 같아야 한다. 즉,
$D_{KL}[P(x;\theta)||P(x;\theta')] = D_{KL}[P(y;\theta)||P(y;\theta')]$임을 확인해 보자.

$$
\begin{aligned}
&D_{KL}[P(x;\theta)||P(x;\theta')] \\
&= \sum_{x_1,\cdots,x_N} P(x_1,\cdots,x_N;\theta) \ln \frac{P(x_1,\cdots,x_N;\theta)}{P(x_1,\cdots,x_N;\theta')} \\
&= N \sum_{x_1 \in \{0,1\}} P(x_1;\theta) \ln \frac{P(x_1;\theta)}{P(x_1;\theta')} \\
&= N\left(\theta_0 \ln \frac{\theta_0}{\theta'_0} + \theta_1 \ln \frac{\theta_1}{\theta'_1}\right).
\end{aligned} \tag{6}
$$

위 계산에서는 $x_1, x_2, \cdots, x_N$이 서로 독립이라는 사실에서 다음 등식을 이용하였다.

$$
P(x_1, x_2, \cdots, x_N; \mu) = P(x_1;\mu)P(x_2;\mu)\cdots P(x_N;\mu). \tag{7}
$$

이번에는 충분통계량 y를 이용한 거리를 계산해보자.

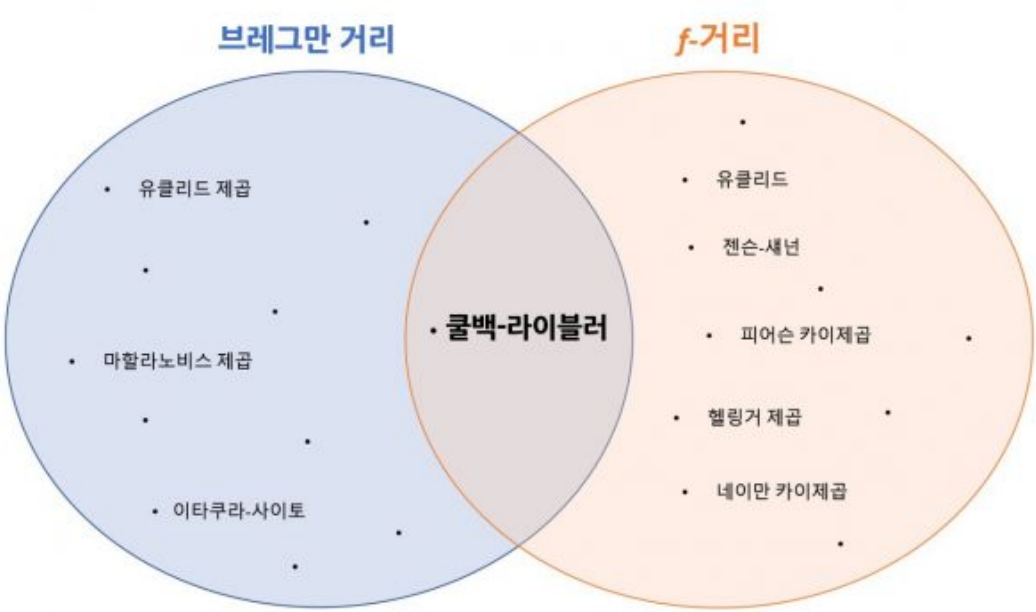
$$
\begin{aligned}
&D_{KL}[P(y;\theta)||P(y;\theta')] \\
&= \sum_{y} P(y;\theta) \ln \frac{P(y;\theta)}{P(y;\theta')} \\
&= \sum_{y=0}^{N} \binom{N}{y} \theta_0^{(N-y)} \theta_1^{y} \ln \frac{\binom{N}{y} \theta_0^{(N-y)} \theta_1^{y}}{\binom{N}{y} \theta_0^{`(N-y)} \theta_1^{`y}} \\
&= N\left(\theta_0 \ln \frac{\theta_0}{\theta'_0} + \theta_1 \ln \frac{\theta_1}{\theta'_1}\right).
\end{aligned} \tag{8}
$$

이 결과는 위의 식 (6)과 일치하는 것이다. 즉, 쿨백-라이블러 거리는 충분통계량에 해당하는 변수변환에 대해서 불변임을 확인한 것이다. 마지막 계산과정에서는 이항분포의 다음 항등식을 이용하였디.

$$\sum_{y=0}^{N} y \binom{N}{y} \theta_0^{(N-y)} \theta_1^y = \theta_1 \frac{\partial (\theta_0 + \theta_1)^N}{\partial \theta_1} = N\theta_1 \tag{9}$$

$$\sum_{y=0}^{N} (N-y) \binom{N}{y} \theta_0^{(N-y)} \theta_1^y = \theta_0 \frac{\partial (\theta_0 + \theta_1)^N}{\partial \theta_0} = N\theta_0. \tag{10}$$

지난 글에서는 쿨백-라이블러 거리가 오목함수에서 정의되는 브레그만 거리 가운데 하나임을 확인했다. 이번 글에서는 쿨백-라이블러 거리가 f-거리 가운데 하나임을 확인했다. 흥미롭게도, 쿨백-라이블러 거리는 f-거리가 되는 유일한 브레그만 거리로 알려져 있다.[그림1] 두 글을 정리해 보자면, 쿨백-라이블러 거리는 쌍대적으로 평평한 공간에서 일반화된 피타고라스의 정리를 만족하고, 또 충분통계량에 대해서도 불변인 아주 특별한 거리이다.



[그림1] 쿨백-라이블러 거리

머신러닝을 공부해 본 사람은 여기저기서 쿨백-라이블러 거리를 만난 적이 있을 것이다. 두 확률모형을 비교할 때, 또는 모형과 데이터의 분포를 비교할 때, 왜 하필 쿨백-라이블러 거리를 써야 하는지에 대한 하나의 설득력 있는 대답이 되었기를 바란다. 앞으로는 믿고 안심하고 쓸 수 있기를. 다음 연재에서는 이런 정보기하학적 개념이 실제 머신러닝의 최적화에 어떻게 멋지게 이용될 수 있는지를 한번 소개해 보겠다.

참고문헌

1. 충분통계량 https://en.wikipedia.org/wiki/Sufficient_statistic
2. F-거리 https://en.wikipedia.org/wiki/F-divergence