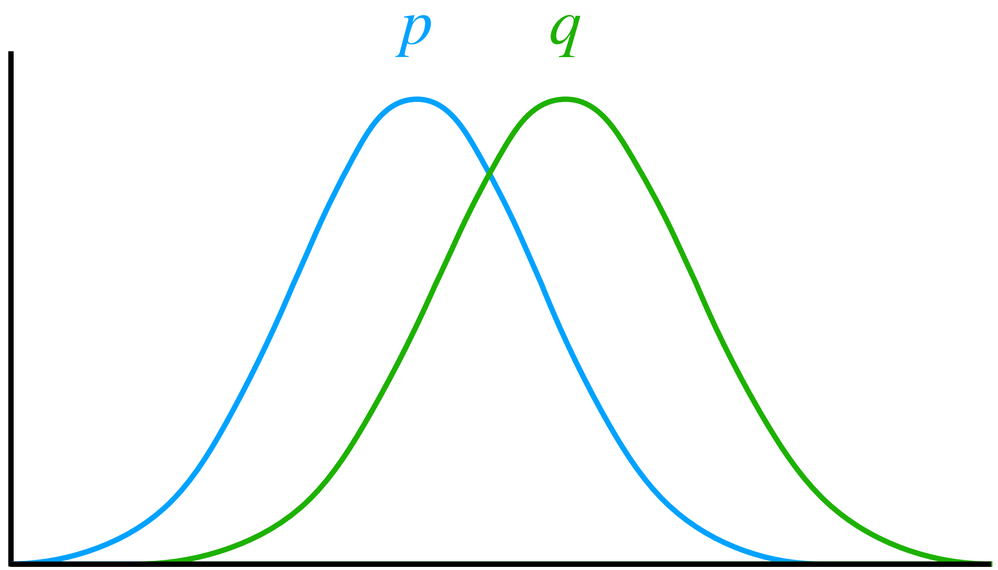
ICT이노베이션스퀘어 AI복합교육 고급 언어과정

자연어처리를 위한 Kullback–Leibler divergence

현청천

2021.04.19

KL-divergence (연속확률분포)



$$D_{KL}(p \parallel q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$

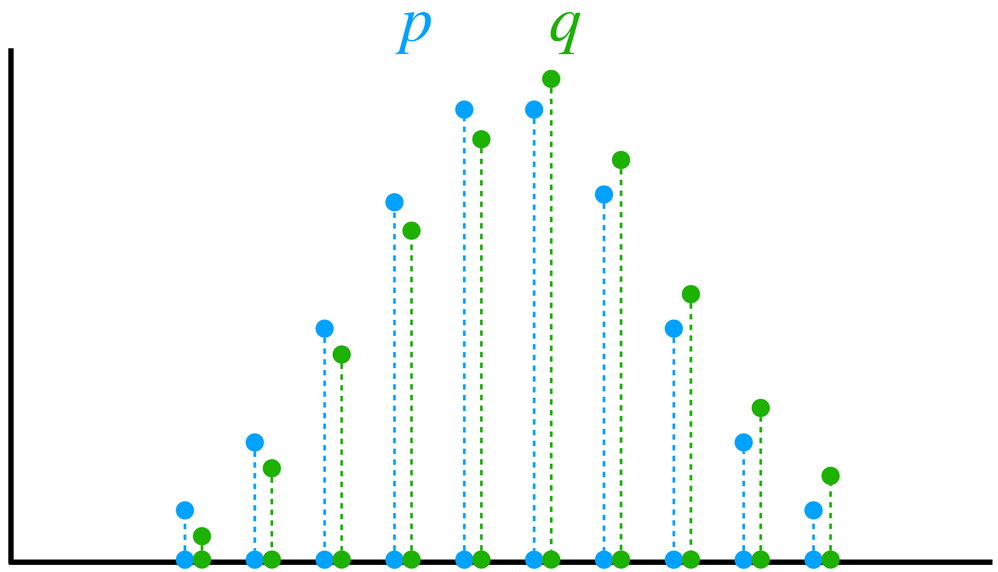
$$= -\int_x p(x) \log q(x) dx + \int_x p(x) \log p(x) dx$$

$$= H(p, q) - H(p)$$

$$D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$$

두 확률분포 p, q의 정보량의 차이

KL-divergence (이산확률분포)



$$D_{KL}(p \parallel q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} dx$$

$$= -\sum_{x} p(x) \log q(x) dx + \sum_{x} p(x) \log p(x) dx$$

$$= H(p, q) - H(p)$$

두 확률분포 p, q의 정보량의 차이

KL-divergence

p

A	50%	0
B	25%	1 0
C	12.5%	1 1 0
D	12.5%	1 1 1

$\log_2 \frac{1}{0.5} = 1$

$\log_2 \frac{1}{0.25} = 2$

$\log_2 \frac{1}{0.125} = 3$

q

A	25%	0 0
B	25%	0 1
C	25%	1 0
D	25%	1 1

$$H(p) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

$$0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 = 1.75$$

$$D_{KL}(p \parallel q) = H(p, q) - H(p) = 0.25$$

$$0.5 \times 2 + 0.25 \times 2 + 0.125 \times 2 + 0.125 \times 2 = 2$$

$$H(p, q) = \sum_x p(x) \log_2 \frac{1}{q(x)}$$

KL-divergence

p

A	50%	0
B	25%	1 0
C	12.5%	1 1 0
D	12.5%	1 1 1

$\log_2 \frac{1}{0.5} = 1$

$\log_2 \frac{1}{0.25} = 2$

$\log_2 \frac{1}{0.125} = 3$

q

A	50%	0
B	25%	1 0
C	12.5%	1 1 0
D	12.5%	1 1 1

$$H(p) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

$$0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 = 1.75$$

$$D_{KL}(p \parallel q) = H(p, q) - H(p) = 0$$

$$0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 = 1.75$$

$$H(p, q) = \sum_x p(x) \log_2 \frac{1}{q(x)}$$

KL-divergence

A, B 두 글자가 발생하는 경우

A 발생 확률에 따른 두 확률분포의

KL-divergence

$q(A)$

$p(A)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0	0.04	0.12	0.23	0.37	0.55	0.79	1.15	1.76
0.2	0.04	0	0.03	0.09	0.19	0.33	0.53	0.83	1.36
0.3	0.15	0.03	0	0.02	0.08	0.18	0.34	0.58	1.03
0.4	0.31	0.1	0.02	0	0.02	0.08	0.19	0.38	0.75
0.5	0.51	0.22	0.09	0.02	0	0.02	0.09	0.22	0.51
0.6	0.75	0.38	0.19	0.08	0.02	0	0.02	0.1	0.31
0.7	1.03	0.58	0.34	0.18	0.08	0.02	0	0.03	0.15
0.8	1.36	0.83	0.53	0.33	0.19	0.09	0.03	0	0.04
0.9	1.76	1.15	0.79	0.55	0.37	0.23	0.12	0.04	0

감사합니다.