14 강 확률분포(2)

◈ 담당교수: 장필훈

▣ 주요용어

용어	해설
디리클레분포	$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$, $\sum\mu_k = 1$ 으로 정의되는 확률분포. 매개변수에 따라 분포가 달라지고 k=2 일때가 베타분포이다.
가우시안분포	정규분포라고도 한다. 지수족 분포중에 가장 흔하게 접할 수 있는 분포. 식으로 표현하면 다음과 같다. $\mathcal{N}(\mathrm{x}\|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$
감마분포	지수분포의 컬레사전확률분포. 두개의 매개변수를 가진다. 매개변수를 조절해서 지수분포를 얻을 수도 있고 카이제곱분포를 얻을 수도 있다. 매개변수의 조절에 따른 분포의 모양은 강의록 참고.
마할라노비스거리	단순히 좌표계만 고려한 거리를 유클리디안 거리라고 한다면, 마할라노비스 거리는 분포를 고려한 거리를 뜻한다. 즉 좌표평면상의 데이터들이 가지는 분포를 고려하여 분산에 비해 얼마나 더 멀리 혹은 가까이 있는지를 나타내기 위해 쓴다. 따라서 분산을 고려할 필요가 없는 분포(예: 등방분포)에서는 유클리디안 거리와 차이가 없다.

▣ 정리하기

1. 확률변수 하나가 여러개의 차원을 가질 수 있으며, 그러한 확률변수를 다항변수라고 한다.

2. 다항변수가 1-hot encoding 으로 나타날 수 있다고 할 때, 각 차원의 매개변수(1 이 될 확률)을 μ_k라고 하면, 다항변수 하나의 분포는 다음과 같이 주어진다.

$$\mathrm{p}(\mathrm{x}|\mu) = \prod_{\mathrm{k}=1}^{\mathrm{K}}\mu_k^{x_k}$$

   a. x_k가 1 아니면 0 이므로, 0 이면 자연스럽게 무시되는 효과가 있다.

3. 여러개의 다항변수를 가진 데이터집합 $\mathcal{D}$를 고려해보면, 다음과 같은 형태를 가진다.

$$\mathrm{p}(\mathcal{D}|\mu) = \prod_{\mathrm{k}=1}^{\mathrm{K}}\mu_k^{m_k}, m_k = \sum_n x_{nk}$$

4. 다항분포의 매개변수 μ에 대한 사전분포가 디리클레 분포면, 사후분포도 디리클레 분포다.

5. 다음형태로 나타나는 분포가 가우시안 분포다.

$$\mathcal{N}(\mathrm{x}|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

   a. μ가 평균, σ^2이 분산에 해당한다.

6. 두 변수집합의 결합분포가 가우시안이라면,
    a. 하나의 변수집합에 대한 다른 변수집합의 조건부 분포도 가우시안이다.
    b. 각 변수집합의 주변분포도 가우시안이다.

7. 가우시안 분포의 최대 가능도함수를 이용해서 평균과 분산을 추정할 수 있다.
    a. 평균은 관측된 데이터포인트들의 평균이다
    b. 순차추정이 가능하다. 즉 데이터포인트 하나 추가될때마다 추정값을 업데이트 할 수 있다.

8. 가우시안분포에서 다음의 경우로 나누어 (매개변수의)베이지안 추론을 해볼 수 있다.
    a. 분산을 알고 있는 상태에서 평균추정
    b. 평균을 알고 있는 상태에서 분산추정
    c. 둘 다 모르는 상태에서 추정.