워크북

교과목명 : 머신 러닝
차시명: 14차시

◈ 담당교수: **장 필 훈**

⊙ 세부목차

- 다항변수
- 디리클레분포
- 가우시안분포
- 감마분포

	학습에 앞서

■ 학습개요

이전시간에 이어 다양한 확률분포에 관해 배운다. 이전시간과는 다르게 다차원 데이터까지 다루기 위해 다항변수의 개념을 익히고, 그 변수의 확률분포인 다항분포를 배운다.

그 다음 가우시안에 대해 자세히 배운다. 가우시안 분포는 가장 널리 쓰이고, 연구결과도 가장 많은 분포이다. 다변량에까지 확장된 가우시안을 배우므로, 평균과 공분산을 벡터나 행렬로 계산하고 더 나아가 가우시안 분포의 평균과 분산추정을 모두 수식으로 풀어본다. 우선 분산을 알고 있는 상태에서 평균추정, 다음으로 평균을 알고 있는 상태에서 분산추정, 마지막으로 둘다 모르는 상태에서 추정을 모두 다루어본다. 그 과정중에 감마분포에 대해 더 자세히 보게 되고, 여러종류의 그래프를 통해 직관적인 감을 더 기르는데 집중한다.

■ 학습목표

1	고차원상의 데이터에서도 확률분포를 이야기 할 수도 있음을 이해하고 다항변수의 가우시안 분포가 식으로 어떻게 나타나는지 본다.
2	베이지안적 관점에서 확률분포들을 이해하고, 식으로 어떻게 전개하는지 확인한다.
3	디리클레 분포, 가우시안 분포, 감마분포, 베타분포등 여러 분포를 이해하고 각 분포의 기댓값,

	분산등을 계산하는 방법을 익힌다.
4	가우시안 분포의 베이지안 적용을 여러가지 문제를 통해 이해하고 실제로 식을 전개해본다.
5	매개변수들이 존재하는 분포(예:감마분포, 베타분포)들을 이해하고 매개변수가 변화함에 따라 분포 모양이 어떻게 변하는지 확인한다.

■ 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=2일때가 베타분포이다.
가우시안분포	정규분포라고도 한다. 지수족 분포중에 가장 흔하게 접할 수 있는 분포. 식으로 표현하면 다음과 같다. $\mathcal{N}(\mathrm{x}\vert\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$
감마분포	지수분포의 켤레사전확률분포. 두개의 매개변수를 가진다. 매개변수를 조절해서 지수분포를 얻을 수도 있고 카이제곱분포를 얻을 수도 있다. 매개변수의 조절에 따른 분포의 모양은 강의록 참고.
마할라노비스 거리	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.

학습하기

이번시간은 다항변수를 다루면서 주로 가우시안 분포에 관해 이야기해 보겠습니다.

다항변수의 대표적인 예는 1-hot encoding입니다. K개중 하나를 선택해야 하는 이산변수의 경우, 각 차원마다 0,1중 하나를 가지되, 모든 차원중 1을 가지는 차원이 단 하나인 경우를 말합니다. 예를들어 (0,0,1,0,0,0)은 1-hot encoding이지만 (0,0,0,0,0,0), (0,0,0,1,1,0)등은 아닙니다. 차원수는 임의대로 정할 수 있습니다. 예를들어 자연어처리에서 특정 단어 뒤에 나오는 단어가 무엇인지 출력으로 주는 분류기를 생각해보면, 출력차원이 후보단어들의 개수와 같게 됩니다. 그중에 하나만 1이고 그에 해당하는 단어가 다음 단어가 되는 것입니다. 수식으로 표현하면 $\sum_{\mathrm{k}=1}^{\mathrm{K}}x_k=1$이 됩니다.

이를 확률분포로 나타내면, $P(x_k=1)=\mu_k$일 때, $p(\mathbf{x}\vert\boldsymbol{\mu})=\prod_{k=1}^{K}\mu_k^{x_k}$ 입니다. 베르누이 분포를 두가지 이상인 경우로 확장한 것입니다.

기댓값은 확률벡터 그 자체가 됩니다. (value가 1 아니면 0이고, 1일 확률이 μ이므로 기댓값이μ가 되는

것은 당연합니다. 그러면 데이터집합 $\mathcal{D} = \{\mathbf{x_1}, \dots, \mathbf{x_N}\}$의 가능도함수는 다음과 같이 됩니다.

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

여기서 $m_k = \sum_n x_{nk}$(즉 $x_k = 1$인 관측값의 숫자)입니다.

(늘 하듯이) 최대가능도해를 구해봅시다. $\sum\mu_k = 1$의 제약조건하에서 위의 가능도함수의 로그가 가질 수 있는 최댓값을 구하면 됩니다. 이번에도 (늘 했듯이) 라그랑주 승수법을 사용합니다.

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda\left(\sum_{k=1}^{K}\mu_k - 1\right)$$ (익숙한) 라그랑지안 함수를 만듭니다.

μ에 대해 미분하고 0으로 두면 $\mu_k = -m_k/\lambda$를 얻고, 이를 제약조건에 대입하면 다음과 같이 최대가능도해를 얻습니다.

$$\mu_k^{ML} = \frac{m_k}{N}$$ 즉, 'N개 관측값중 $x_k = 1$인 경우의 비율'입니다. (천천히 생각해보시면 당연한 결과입니다)

이제 다항분포를 알아보겠습니다.

매개변수$\boldsymbol{\mu}$(벡터)와 관측숫자 N에 의해 결정되는 수량 m1, m2, …, mk가 있을 때

$\sum m_k = N$이 되고, 결합분포는 다항분포라고 부릅니다.

$$\mathrm{Mult}(m_1, \dots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \dots m_K}\prod_{k=1}^{K}\mu_k^{m_k}, \quad \binom{N}{m_1 m_2 \dots m_K} = \frac{N!}{m_1!\, m_2! \dots m_K!}$$

그럼 다항분포의 매개변수$\boldsymbol{\mu}$의 사전분포는 무엇을 사용하면 좋을까요? 이 질문은 켤레(conjugate)사전분포를 묻는 것과 같습니다. ('좋다'는건 단순히 우리가 계산하기 좋은것을 말하는 것이니까요) 켤레사전분포는 다음의 형태여야 합니다.

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K}\mu_k^{\alpha_k - 1}$$

$$\mathbf{0 \le \mu_k \le 1,\ \ \sum_k \mu_k = 1,\quad \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^T}$$

위의 형태대로 하면 사후분포도 같은 형태로 얻을 수 있음을 쉽게 확인할 수 있습니다. (곱해보면 되니까요) 비례(∝)관계 말고 정확한 확률분포를 알려면 정규화계수만 알면 됩니다. 그 결과가 다음과 같고, 이를 디리클레 분포라고 합니다.

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k - 1} \qquad \alpha_0 = \sum_{k=1}^{K}\alpha_k$$

다시말하면, 디리클레분포를 사전분포로 취하면, 사후분포 또한 디리클레 분포가 됩니다.('디리클레분포는 다항분포의 켤레사전분포다') 그리고 디리클레 분포에서 k=2일때는 (당연하게도) 베타분포가 됩니다. 우리가 앞서 배운 베타분포가 어떤 분포의 켤레분포였는지 생각해보세요.

디리클레분포는 파라미터에 따라 어떤 모양이 되는지 상상하기가 쉽지 않습니다. 강의에서 잠깐 언급

하고 넘어갔으니 참고 바랍니다.

<가우시안분포>

가우시안분포는 정규분포라고 합니다. 지금까지 너무 많이 다루어서 이제 지루할만큼 익숙하시리라 생각합니다. 단일변수의 경우 다음과 같고,

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

D차원의 경우 다음과 같습니다. 여기서 분산대신 공분산행렬이 들어가는 점을 눈여겨보세요. 변수도 모두 벡터입니다. 행렬의 절댓값은 행렬식을 뜻합니다.

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

위 식의 우측에 보이는 부분은 마할라노비스 거리입니다. 다차원 벡터의 거리를 측정할 때 단순히 차원간 차이를 제곱해서 더하는 수도 있지만, 데이터의 분포가 차원간에 차이가 있을 경우 그렇게 측정하면 두 데이터포인트 사이의 ‘거리’를 정확하게 표현하기가 어렵습니다. 특정차원의 경우 분포가 넓을 수 있고 다른 차원의 경우 좁다고 가정하면, 좁은 차원에서는 차이가 조금만 나도 차이로 받아들여야 하고 넓은 차원에서는 어느정도 거리도 무시할 수 있어야 ‘균형’있는 거리비교가 가능할 것입니다. 마할라노비스 거리($(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\sum^{-\mathbf{1}}(\mathbf{x}-\boldsymbol{\mu})$)는 공분산행렬($\sum$)을 이용해서 이 차이를 반영합니다. 만일 공분산행렬이 항등행렬(차원간 차이가 없음)이라면 자연스럽게 유클리디안 거리가 됩니다.

다차원 가우시안 분포일때 기댓값을 구해보면 다음과 같습니다.

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}d\mathbf{x} \\ &= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})d\mathbf{z} \\ &= \boldsymbol{\mu}\end{aligned}$$

우함수이고, $-\infty \sim \infty$적분하면 $\mathbf{z}$ 항이
대칭성으로 없어짐

분산을 구하기 위해 이차모멘트를 구해보겠습니다.

$$\begin{aligned}\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] &= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\mathbf{x}^{\mathrm{T}}d\mathbf{x} \\ &= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^{\mathrm{T}}d\mathbf{z}\end{aligned}$$

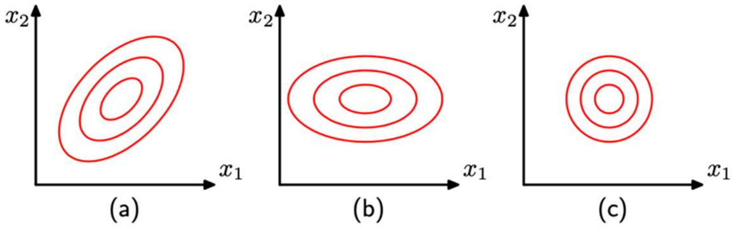
교차항(μz^T ,$\mathrm{z}\mu^{\mathrm{T}}$)은 사라지고 정규화되어 있으므로 $\mu\mu^T = 1$, zz^T항은 공분산 행렬이 됩니다.

조금 더 자세하게 강의에서 다루었지만 line by line으로 더 자세한 전개를 원하시는 분도 쉽게 자료를 구할 수 있으니 공부하는 데 따로 어려움은 없을것이라 예상합니다만, 자세히 이해하기 위해서는, 행렬과 벡터의 연산이라서 행렬계산을 미리 익혀두기를 추천합니다.

결국 $E[xx^T] = \mu\mu^T + \Sigma$이 되고 이를 바탕으로 공분산을 구하면 다음과 같습니다.(다차원이기 때문에 ‘분산’이 아니라 ‘공분산’을 구합니다)

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] = \mathbf{\Sigma}$$

이 공분산에 따라 (2차원) 가우시안이 어떻게 나타나는지 다음 그림을 보세요.

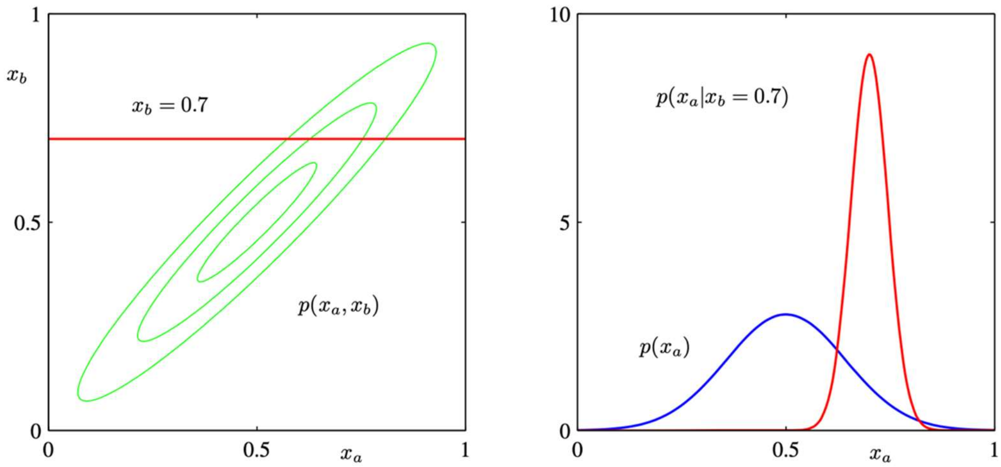


[Bishop. Fig. 2.8]

공분산이 각각 a: 일반적인 공분산 행렬, b: 대각행렬, c:등방성 공분산(항등행렬의 상수배)일때 입니다.

가우시안 분포는 유니모달(봉우리가 하나)이 기본이기 때문에 멀티모달에 대해 적절한 근사치를 제공하지 않습니다. 잠재변수를 이용해서 이 문제를 해결하는데, 가우시안 혼합모델을 사용하거나 선형동적 시스템(분포가 매개변수에 의해 변하도록)을 이용합니다.

두 변수집합의 결합분포가 가우시안이라고 가정하면, 하나의 변수집합에 대한 다른 변수집합의 조건부분포도 가우시안이고, 각 변수집합의 주변분포도 가우시안입니다. 개념적으로는 다음 그림과 같습니다.



[Bishop.Fig.2.9]

왼쪽그림에서 붉은선 단면을 잘라 그린것이 오른쪽 붉은선입니다. (왼쪽 그림은 지도의 ‘등고선’개념에 빗대어 생각하시면 편합니다) 오른쪽 그림의 파란색 선은 marginal distribution입니다(왼쪽 빨간선과 관계 없습니다.) 위에서 주변(marginal)확률은 x축으로 사영한 것이라고 생각하면 됩니다. 주변화의 정의 자체가 다른 변수(여기서는 x_b)의 전구간에 대해 적분해서 그 변수를 없애는 것이기 때문입니다.

파란선과 빨간선이 모두 가우시안분포인것을 관찰할 수 있습니다. 상세한 계산은 강의시간에 다루었습니다. 눈에 보기에 대~충 가우시안같아 보여도 실제로 확인해보기 전에는 확신할 수없지요 ☺ 강의를 참고하시기 바랍니다.

다변량 가우시안 분포의 로그 최대가능도는 다음과 같이 주어집니다.(단순히 확률분포에 로그를 한 것입니다.)

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

이제 (항상 했듯이) 미분=0 으로 놓고 해를 구하면,

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

즉, 관측된 데이터포인트들의 평균벡터입니다. 직관적이죠.

공분산에 대해 최대화 하는것도 같은 방법을 사용하지만 계산이 매우매우 복잡합니다. (강의시간에도 다루지 않았기 때문에 다른 자료를 참고하셔야 합니다.)

위에서 우리가 구한 μ_{ML}을 잘 보면, 순차적으로 계산이 가능함을 알 수 있습니다.(사실 모든 평균이 그렇지요)

$$\mu_{ML}^{(N)} = \frac{1}{\mathrm{N}}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\mu_{ML}^{(N-1)} = \mu_{ML}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \mu_{ML}^{(N-1)})$$

이제 가우시안분포에서의 베이지안 추론을 해보겠습니다. 위에서 우리는 최대가능도 '값'을 구했는데요, 이것을 '분포'의 형태로 구해보는것은 어떨까요. μ에 관한 최대가능'분포'라고 표현해도 좋을 것입니다. 일단은 사전함수의 형태를 가정해야 하는데 이도 역시 가우시안을 가정하도록 하겠습니다. 왜냐하면 이렇게 해야 사후분포도 같은 가우시안을 얻기 때문입니다. (켤레분포에서 우리는 비슷한 사례들을 보았습니다. 그리고 가우시안(지수족)이 곱셈에 대해 닫혀있다는 이야기도 앞서 나왔었습니다)

먼저, 분산을 알고 있는 상태에서 평균을 추정해보겠습니다. 가능도 함수는 다음과 같고,

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N}p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

사전분포도 가우시안을 가정합니다.

$$p(\mu) = \mathcal{N}(\mu|\mu_0,\sigma_0^2)$$

사후분포는 $p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu)$이므로, 다음과 같은 새로운 가우시안을 얻습니다.

$$p(\mu|\mathbf{X}) = \mathcal{N}(\mu|\mu_N,\sigma_N^2)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

증명은 강의시간에 다루었습니다. 강의를 참고바랍니다.

식으로부터 μ_N이 N에 따라 어떻게 변하는지 관찰할 수 있습니다. N이 0에서 무한대로 변할때 μ_N이 어떻게 변할까요. 분산은 어떤가요. 어렵지 않으니 천천히 생각해보시고 강의에서 이야기한것과 같은지

비교해보시기 바랍니다.

두번째로 평균을 알고 있는 상태에서 분산추정문제를 풀어보려고 하면, (분산의 더 정확하게는 분산의 역의) 사전분포를 감마분포로 잡아야 한다는 것을 알 수 있습니다.

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

사후분포를 계산하면 다음과 같습니다.

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1} \lambda^{\frac{N}{2}} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

(다음시간에는 여기서부터 이어서 하겠습니다.)

감마분포의 모양과 대략의 성질에 관해서는 강의를 참고하시기 바랍니다.

결국 분포중의 하나이기 때문에 대략의 모양,성질만 이해하는 것으로 충분합니다.

	연습문제

1. 다항변수에서 μ의 최대가능도해는 N개 관측값 중 1이 발견될 비율과 같다. 즉 $\mu_k^{ML} = \frac{m_k}{N}$이다.
   a. O
   b. 강의록 참고. $\sum \mu_k = 1$의 조건에서 라그랑주 승수법으로 구한다.
2. 다항분포의 매개변수 μ_k의 사전분포는 디리클레 분포이다.
   a. O
   b. 강의록 참고. 연역적으로 도출하지 않고 디리클레 분포형태를 사전분포 삼았을 때 사후분포도 디리클레분포형태가 나옴을 보인다. '디리클레 분포는 다항분포의 켤레사전분포'
3. 디리클레 분포는 다음으로 정의된다. 여기서 k=2일때가 베타분포다.

   $\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$,$\sum\mu_k = 1$
   a. O
   b. 베타분포는 다음과 같은 분포인데, 정확히 디리클레분포에서 k=2일때이다.

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

4. 여러개의 확률변수들의 합에 해당하는 확률변수는 그 숫자가 증가함에 따라 점점 가우시안이 된다.
   a. O
   b. 중심극한정리를 설명한 것이다.
5. 마할라노비스 거리는 데이터의 분산을 고려한 거리이다. 따라서 유클리디안 거리를 포함하는 개

념이다.

   a. O
   b. 마할라노비스 거리는 다음과 같이 정의된다.

      $(\mathrm{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu), x$와 μ는 모두 vector

      여기서 중간에 있는 Σ이 공분산을 나타내고, 해당 공분산이 항등행렬이면 유클리디안 거리가 된다.

6. 두 변수집합의 결합분포가 가우시안이면 하나의 변수집합에 대한 다른 변수집합의 조건부확률분포도 가우시안이고, 각 변수집합의 주변분포도 가우시안이다.
   a. O
   b. 강의록참고. 직접 식을 전개해봄으로써 모두 확인가능
7. 가우시안 분포에서 분산을 알고 있는 경우 평균을 추정한다고 하자. 관찰 데이터 수가 0이면 사후평균은 사전평균과 같고, 관찰 데이터 수가 무한대에 가까워지면 사후평균은 최대가능도 해와 같아진다.
   a. O
   b. 강의록 참고. 사전평균을 μ_0, 최대가능도해를 μ_{ML}이라 하면, 사후평균은 $\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$로 나타낼 수 있고, 이때 N값이 0에 가까워지거나 무한대에 가까워질 때 지문의 설명이 옳음을 확인할 수 있다.
8. 가우시안 분포에서 분산을 알고 있는 경우 평균을 추정한다고 하자. 관찰 데이터 수가 증가할때 분산은 감소한다.
   a. O
   b. 강의록참고. N이 증가할 때 분산의 역은 증가하고 따라서 분산은 감소한다.

정리하기

1. 확률변수 하나가 여러개의 차원을 가질 수 있으며, 그러한 확률변수를 다항변수라고 한다.
2. 다항변수가 1-hot encoding으로 나타날 수 있다고 할 때, 각 차원의 매개변수(1이 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 1아니면 0 이므로, 0이면 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.

참고하기

Bishop, C. M. "Bishop-Pattern Recognition and Machine Learning-Springer 2006." Antimicrob. Agents Chemother (2014): 03728-14.

다음 차시 예고

- 감마분포
- 정규감마분포
- 스튜던트 t분포
- 가우시안분포의 혼합
- 지수족
- 비매개변수적 방법
  - 히스토그램 밀도추정
  - 커널밀도추정
  - K최근접이웃