워크북

교과목명 : 머신 러닝

차시명: 15차시

◈ 담당교수: **장 필 훈**

◉ 세부목차

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

학습에 앞서

■ 학습개요

확률분포의 마지막 시간으로 가우시안 분포에 대한 베이지안 추론을 마무리한다. (평균을 알 때 분산을 추정하는 문제에서) 켤레 사전분포인 감마분포에 대해 사전,사후분포일때 매개변수 변화를 관찰한다. 평균과 분산을 모두 모를 때 추정하는 문제에서 켤레확률분포의 모양(정규감마)을 확인하고, 계산을 통해 확인해본다. 정규감마로부터 도출되는 스튜던트분포에 관해서도 알아본다.

단변량 가우시안을 마치고 다변량 가우시안의 경우 어떤 켤레분포를 가지는지 알아본다(식으로 확인해보는 과정은 생략).

분포들의 혼합중 대표적으로 가우시안 분포의 혼합을 다시한번 살펴보고 그림을 통해 그 성질을 직관적으로 이해한다.

지수족이라는 포괄적인 분포형태를 배우고 우리가 알고 있는 분포들중 상당수가 지수족에 속함을 식을 통해 확인한다.

확률분포를 마치고, 비매개변수적 방법을 살펴본다. 대표적으로 K최근접이웃과 커널밀도 추정 그

두가지를 살펴보게 되고, 그 둘이 나오게 된 자연스러운 배경을 식으로 유도해본다.

■ 학습목표

1	감마분포, 정규감마분포의 성질 파악
2	다변량 가우시안 분포, 스튜던트t분포등을 살펴본다.
3	지수족에 대해 배우고 어떤 분포가 지수족에 속하는지 확인해본다.
4	비매개변수적 방법의 추정이 어떤 배경에서 크게 두가지로 갈리는지 배우고 이해한다.

■ 주요용어

용어	해설
감마분포	확률밀도함수가 $\frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k\Gamma(k)}, x>0$로 주어지는 분포.
정규감마분포	정규분포와 감마분포의 곱.
스튜던트분포	정규감마분포에서 정밀도에 대해 marginalize한 분포. 모든 분산에 대한 가우시안의 총 합을 나타낸다.
켤레사전분포	사전확률과 사후확률이 같은 분포계열에 속할 때 그 사전확률분포를 가리키는 말. 계산상 잇점이 있다.(가능도함수와 곱을 매번 계산하지 않고 매개변수만 업데이트 하는 방식으로 사후확률을 계산해낼수있다)

학습하기

(이전시간에 이어서)

감마사전분포와 가능도 함수의 곱은 다음과 같습니다.

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

사전분포와 비교해보면 다음과 같이 매개변수가 변했음을 알 수 있습니다.

$$a_N = a_0+\frac{N}{2}, \qquad b_N = b_0+\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0+\frac{N}{2}\sigma_{ML}^2$$

즉, 사전분포의 매개변수 a_0를 $2a_0$개의 사전관측값이라고 해석하면, 연속적 분포의 추정이 가능합

니다. 물론 맨 처음 가정했던 켤레분포의 매개변수는 '가상의' 데이터 포인트라고 이해할 수 있습니다.(지수족 분포에서 많이 사용하는 방법)

이제 마지막으로, 평균과 정밀도(1/분산)을 모두 모르는 경우를 보겠습니다. 이때는 가능도 함수가 다음과 같고,

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

이 가능도 함수와 같은 형태를 가진 사전분포를 사용하면 됩니다.(사후분포도 사전분포와 같은 형태를 가지게 하기 위해서 그렇습니다.) '정규감마'(=가우시안 감마)분포를 사용하면 된다는 것이 이미 알려져 있고, 다음과 같습니다.

$$p(\mu,\lambda)=\mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)\quad \mu_0=\frac{c}{\beta},a=\frac{1+\beta}{2},b=d-\frac{c^2}{2\beta}$$

이제 다변량 가우시안의 경우를 살펴보겠습니다.

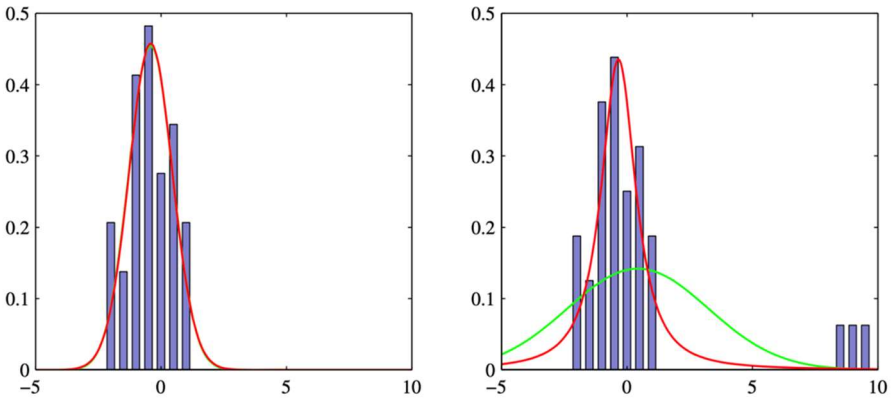
정밀도를 알 때 평균에 대한 켤레사전분포는 단변량일 때와 같이 가우시안입니다.

평균이 알려져 있고 정밀도행렬이 알려져 있지 않을 때 켤레사전분포는 '위샤트 분포'임이 알려져 있습니다. 복잡하기도 하고 크게 중요한것도 아니고 해서 생략했습니다.

가우시안과 감마 사전분포가 주어졌을 때, 정밀도를 적분해서 없애면 x에 대한 주변분포를 구할 수 있습니다. 식으로 나타내면 다음과 같고, 의미는 '같은 평균과 다른 정밀도를 가진 무한히 많은 가우시안 분포들을 합산한 분포'가 됩니다. (이것을 스튜던트t분포라고 합니다)

$$p(x|\mu,a,b)=\int_0^{\infty}\mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)d\tau$$

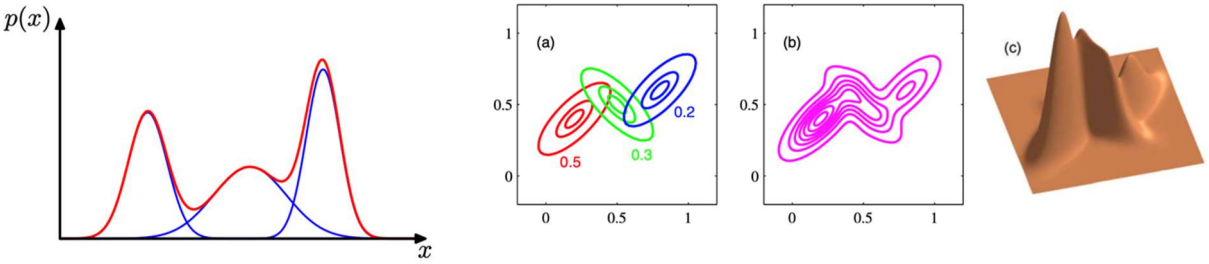
그렇게 가우시안 분포를 무한히 혼합하면, robust해집니다. 아래 그림[Bishop. Fig.2.16]에서 보면,



가우시안 분포로부터 추출한 30개의 데이터포인트로 가우시안 분포와 스튜던트t분포를 근사한 것입니다. 물론 최대가능도법을 이용한 것입니다. t분포(붉은색) 가우시안분포(녹색)가 outlier가 없을 때는 거의 차이가 없지만, outlier가 존재할때는 t분포가 훨씬 더 영향을 받지 않는 것을 관찰할 수 있습니다.

<가우시안분포의 혼합>

지금까지 주로 하나의 분포만 살펴보았지만, 실제의 데이터를 모델링 하기에 하나의 분포는 부족할 때가 많습니다. 그럴때 여러개의 분포를 선형결합해서 사용할 수 있습니다. 가우시안 혼합분포는 그중 한 예로, K개의 가우시안 밀도의 중첩을 뜻합니다.



위 그림[좌:Bishop,Fig.2.22, 우:Bishop,Fig.2.23]을 보면 더 직관적으로 이해하기 쉽습니다.

다만 위와같이 분포를 구성하면 로그가능도함수가 다음과 같이 되어 최대가능도법의 해를 구하기가 매우 어렵게 됩니다.

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

그래서 EM방법으로 구합니다.

<지수족>

지수족 함수의 일반적 성질을 살펴보겠습니다. 가우시안도 지수족의 일종입니다.

지수족의 분포는 다음과 같이 정의됩니다. (eta는 매개변수이고, u(x),h(x),g(η)등은 함수를 뜻합니다. (g(η)는 분포를 정규화함)

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}, \quad g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\}\,dx = 1$$

베르누이 분포도 지수족에 속합니다. 왜냐하면 다음과 같이 전개할 수 있기 때문입니다.

$$\begin{aligned} p(x|\mu) &= \mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \\ &= \exp(x\ln\mu + (1-x)\ln(1-\mu)) \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned}$$

지수족의 정의와 위 식을 비교해보면, $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$, $u(x) = x, h(x) = 1, g(\eta) = \sigma(-\eta)$임을 알 수 있습니다. σ는 시그모이드 함수를 뜻합니다.

그럼 다항분포의 경우는 어떨까요

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M}\mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln\mu_k\right\}$$

위의 경우도 잘(?)정리할 수 있습니다. $x = \{x_1, \dots, x_M)^T, \eta_k = \ln \mu_k, \boldsymbol{\eta} = (\eta_1, \dots, \eta_M)^T$로 두면, $p(\mathbf{x}|\boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^{\mathbf{T}}\mathbf{x})$로 타나낼 수 있고, 다음과 같이 변형 가능합니다.

$$\exp\left\{\sum_{k=1}^{M-1} x_k \ln\left(\frac{\mu_k}{1-\sum_{j=1}^{M-1}\mu_j}\right) + \ln\left(1-\sum_{k=1}^{M-1}\mu_k\right)\right\}$$

그러면 $\eta_k = \ln\left(\frac{\mu_k}{1-\sum_j \mu_j}\right)$로 둘 수 있습니다. 최종적으로 다음과 같이 됩니다.

$$p(\mathbf{x}|\boldsymbol{\eta}) = \left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}\exp(\boldsymbol{\eta}^{\mathbf{T}}\mathbf{x})$$

가우시안 분포도 지수족입니다. 가우시안 분포는 다음과 같이 변형 가능하기 때문입니다.

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

$$\boldsymbol{\eta} = \begin{pmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{pmatrix} \quad g(\boldsymbol{\eta}) = (-2\eta_2)^{\frac{1}{2}}\exp\left(\frac{\eta_1^2}{4\eta_2}\right)$$

$$h(x) = (2\pi)^{-\frac{1}{2}} \qquad \mathbf{u}(x) = \begin{pmatrix}x\\ x^2\end{pmatrix}$$

그럼 지수족 분포에서, η를 최대가능도법으로 추정해 보겠습니다.

다음 가능도함수를η에 대해 미분하고 0으로 두면,

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left(\boldsymbol{\eta}^T \sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right)$$

다음과 같이 됩니다.

$$-\nabla \ln g(\boldsymbol{\eta}_{ML}) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)$$

지수족 분포의 일반적인 켤레사전분포도 구할 수 있습니다.

다음과 같습니다.

$$p(\mathbf{X}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp(\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi})$$

맞는지는 곱해보면 됩니다

$$p(\boldsymbol{\eta}|\mathbf{X},\boldsymbol{\chi},\nu) \propto g(\boldsymbol{\eta})^{\nu+N}\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\left(\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n) + \nu\boldsymbol{\chi}\right)\right)$$

<비매개변수적 방법>

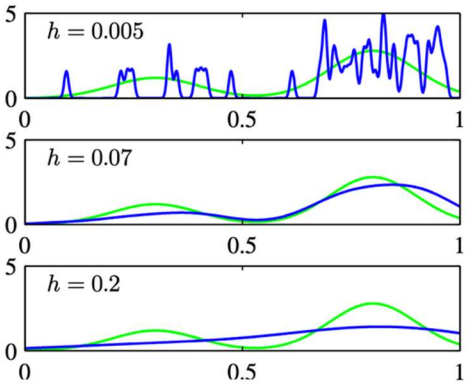
분포를 사용하지 않는 방법을 편의상 비매개변수적 방법이라고 생각하면 쉽습니다. 방법이야 수도 없겠지만, 대표적으로 두가지(커널밀도추정, k최근접 이웃)를 살펴보겠습니다. 이 두가지 방법의 배경이 될 수 있는 비매개변수적 방법의 일반론에 대해서는 강의에서 짧게 언급했습니다.

먼저, 한변의 길이가 h인 입방체(구역 R)에 포함되는 데이터포인트의 수를 K라고 하면 다음과 같이 나타낼 수 있고,

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right), \qquad k(\mathbf{u}) = \begin{cases} 1, i = 1 \ldots D, |u_i| \leq 1/2 \\ 0, \qquad\qquad |u_i| > 1/2 \end{cases}$$

k(u)는 커널함수(파젠창 Parzen window)입니다.

강의에서 커널밀도추정에 관해 자세한 설명을 넣었지만, 직관적으로는 특정크기 공간 안에 데이터포인트가 얼마나 많이 들어있는지 추정해내는 방법입니다.(그런 구역의 모음으로 전체 분포를 추정해냅니다) 그렇게 이해하면 아래 그림도 이해하기 쉽습니다.



[Bishop. Fig.2.25]

h는 구역의 크기를 나타낸다고 보면 됩니다. 구역이 커질수록 데이터포인트의 분포가 더 부드럽게 나오겠지요. outlier의 역할이 좀 더 smoothing되기 때문입니다. 왼쪽 그림에서 녹색이 원 분포, 그리고 파란색이 원분포로부터 추출한 데이터로부터 커널밀도추정을 한것인데, h를 너무 작게 잡으면 데이터포인트 하나하나의 영향을 너무 많이 받아서 분포 전체가 잘 추정되지 않는것을 볼 수 있습니다.(너무 큰 경우도 양봉이 아니라 단봉인것처럼 나타난 것을 관찰할 수 있습니다. 적당해야 좋습니다.)

k최근접 이웃은 여러분이 익숙하게 알고 있는 k-nearest neighbor맞습니다. 이경우는 위의 커널밀도추정과 반대로 공간의 크기가 유동적으로 변합니다. 일정수(K)의 데이터포인트를 포함할때까지 공간의 크기를 축소하거나 확장하는 것입니다. 이 방법의 경우 분류문제에 응용하는 것도 쉬운데 강의시간에 조금 더 자세히 설명했습니다.

그동안 수강하느라 수고했습니다. 부족한 강사가 너무 큰 주제를 다룬것 아닌가 종종 생각했는데, 그럼에도 최대한 잘 전달하려고 노력했습니다. 잘 되었는지 모르겠습니다. 수고하셨습니다.

연습문제

1. (가우시안의 베이지안 관점 해석에서) 평균을 아는 상태에서 분산을 추정할 때, 켤레 사전분포

는 감마분포이다.

a. O

b. 강의록 참고. 식으로 보일 수 있다.

2. (가우시안의 베이지안 관점 해석에서) 분산을 아는 상태에서 평균을 추정할 때, 켤레 사전분포는 감마분포이다.

a. X

b. 이전시간 강의록 참고. 평균에 대한 켤레사전분포는 가우시안분포이다.

3. (가우시안의 베이지안 관점 해석에서) 평균과 분산을 모두 모르는 경우 켤레 사전분포는 가우시안 감마분포이다.

a. O

b. 단순히 가우시안×감마 분포로 이해해도 된다. 강의록 참고.

4. 다변량가우시안의 경우 켤레 사전분포는 단변량 가우시안과 동일하다.

a. X

b. '본질적으로' 동일하다/아니다의 애매한 문제를 차치하고서, 평균을 아는 상태에서 정밀도 행렬을 추정하는 문제에서 다변량 가우시안의 켤레 사전분포는 '위샤트 분포'라고 한다.

5. 가우시안감마분포에서 정밀도에 대한 marginal distribution을 구하면 스튜던트 t-분포가 나온다.

a. O

b. 강의록 참고. 설명 그대로 가우시안 감마를 정밀도에 대해 모든 구간(0~∞)에서 적분하면 스튜던트 분포를 얻지만, 수업에서 과정을 자세히 살펴보지는 않았으므로 이렇다는 사실만 알고 있으면 충분하다.

6. 베르누이 분포는 지수족 분포이다

a. O

b. 강의록 참고. 지수족의 일반적인 형식으로 표현되는지 확인하면 된다. 우리가 알고 있는 대부분의 분포가 지수족에 속한다.

7. 지수족 분포의 혼합분포도 지수족 분포이다

a. X

b. 지수족 분포의 혼합분포의 경우 일반적으로 지수족의 형태로 치환가능하지 않다. 따라서 지수족이라고 할 수 없다.

8. KNN의 경우 K=1인 경우 경계선이 복잡하게 형성되고 이상치(outlier)에 더 민감하게 된다.

a. O

b. 강의록 참고. K가 평활화계수역할을 하게 되고 값이 커질수록 더 부드러운 경계를 형성하며 이상치에 덜 반응하게 된다.(=robust하게 된다)

정리하기

1. (가우시안의) 평균을 아는 상태에서 분산을 추정할 때, 켤레 사전분포는 감마분포이다.
    a. 사전분포와 사후분포를 비교해보면 (같은 감마분포이고) 다음과 같은 변화를 관찰 할 수 있다.

    $$a_N = a_0 + \frac{N}{2}, \quad b_N = b_0 + \frac{N}{2}\sigma_{ML}^2$$

    b. 사전분포의 매개변수 a_0는 $2a_0$개의 사전관측값이라고 해석할 수 있다.
2. 평균과 분산을 모두 모르는 상태라면 켤레사전분포는 다음의 정규감마(가우시안 감마)이다.

$$\mathcal{N}\left(\mu\middle|\mu_0, \frac{1}{\beta\lambda}\right)\mathrm{Gam}(\lambda|a,b)$$

3. 다변량 가우시안의 경우
    a. 분산을 알 때 평균에 대한 켤레사전분포는 가우시안이다.
        i. 단변량일때와 같다.
    b. 평균을 알 때 공분산행렬에 대한 켤레사전분포는 '위샤트 분포'라고 한다.
4. 같은 평균과 다른 정밀도를 가진 무한히 많은 가우시안 분포를 합산하면 스튜던트분포(t분포)가 된다.
    a. 가우시안을 무한히 혼합한 것이기 때문에 outlier에 대해 robust하게 된다.
    b. 최대가능도해는 EM을 이용해 구한다.
5. 극좌표계를 이용해서 주기적 변수로 확률변수로 나타낼 수 있다.
6. 실제 데이터집합은 복잡하기 때문에 단일모델로 나타내기 어려운 경우가 많고, 그럴때는 혼합모델을 사용한다.
7. $h(x)g(\eta)\exp\left(\eta^T u(x)\right)$로 나타낼 수 있는 모든 분포를 지수족 분포라고 한다.
    a. 일례로, 베르누이 분포를 지수족분포의 형태로 나타낼 수 있다.
    b. 다항분포도 지수족 분포의 형태로 나타낼 수 있다.
    c. 가우시안분포도 지수족이다.
8. 켤레사전분포를 이용할 수 있으면, 사후확률을 계산할 때 파라미터만 업데이트 하는 방법으로 계산할 수 있다.
9. 모든 지수족 분포에 적용 가능한 켤레사전분포를 정의할 수 있다.
10. 해석적 분포를 사용할 수 없으면, 파라미터 없이 밀도추정을 해야 한다.
    a. 예: 히스토그램 밀도추정
        i. 구간너비에 따라 결과가 달라진다.
        ii. 고차원 데이터를 다루기 부적합하다.
    b. 크게, 커널밀도추정과 K최근접 이웃 두가지로 구분해볼 수 있다.

참고하기

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

다음 차시 예고

종강. 수고하셨습니다.