기계학습

15강 확률분포(3)

장필훈 교수

학습목차

1 확률분포(3)

01
확률분포(3)

1-1 감마 분포

- 감마사전분포와 가능도함수의 곱 = 사후분포

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

- 사전분포(감마분포)와 비교해보면,

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

1-1 감마 분포

- 사전분포의 매개변수 a_0
  - '$2a_0$개의 사전관측값'이라고 해석가능
  - 켤레사전분포의 매개변수들을 가상의 데이터포인트로 해석
    - 지수족 분포에서 많이 사용하는 방법

1-2 정규감마분포

- **예3**: 평균과 정밀도(1/분산)을 모두 모르는 경우

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

$$\propto \left\{\lambda^{\frac{1}{2}} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right\}^N \exp\left(\lambda\mu\sum_{n=1}^{N} x_n - \frac{\lambda}{2}\sum_{n=1}^{N} x_n^2\right)$$

- 위 가능도 함수와 같은 형태(μ와 λ에 대해 같은형태로 의존)를 가진 사전분포 $p(\mu,\lambda)$를 찾는다.

1-2 정규감마분포

- $p(\mu,\lambda) = p(\mu|\lambda)p(\lambda)$를 이용해서 구함. (수학잘하는 형들이 이미 해둠)

$$p(\mu,\lambda) = \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)$$

  - $\mu_0 = \frac{c}{\beta}, a = \frac{1+\beta}{2}, b = d - \frac{c^2}{2\beta}$
  - ‘정규감마’, ‘가우시안 감마’

1-3 다변량가우시안

- D차원 변수 $\mathbf{x}$에 대한 다변량 가우시안의 경우
  - 정밀도를 알 때 평균 μ에 대한 켤레사전분포는 가우시안.
    - 다변량일 때와 같다.
  - 평균이 알려져 있고 정밀도 행렬 Λ가 알려져 있지 않을 때 켤레 사전분포는 '위샤트 분포'
    Wishart distribution
    - 복잡…

1-4 스튜던트 t-분포

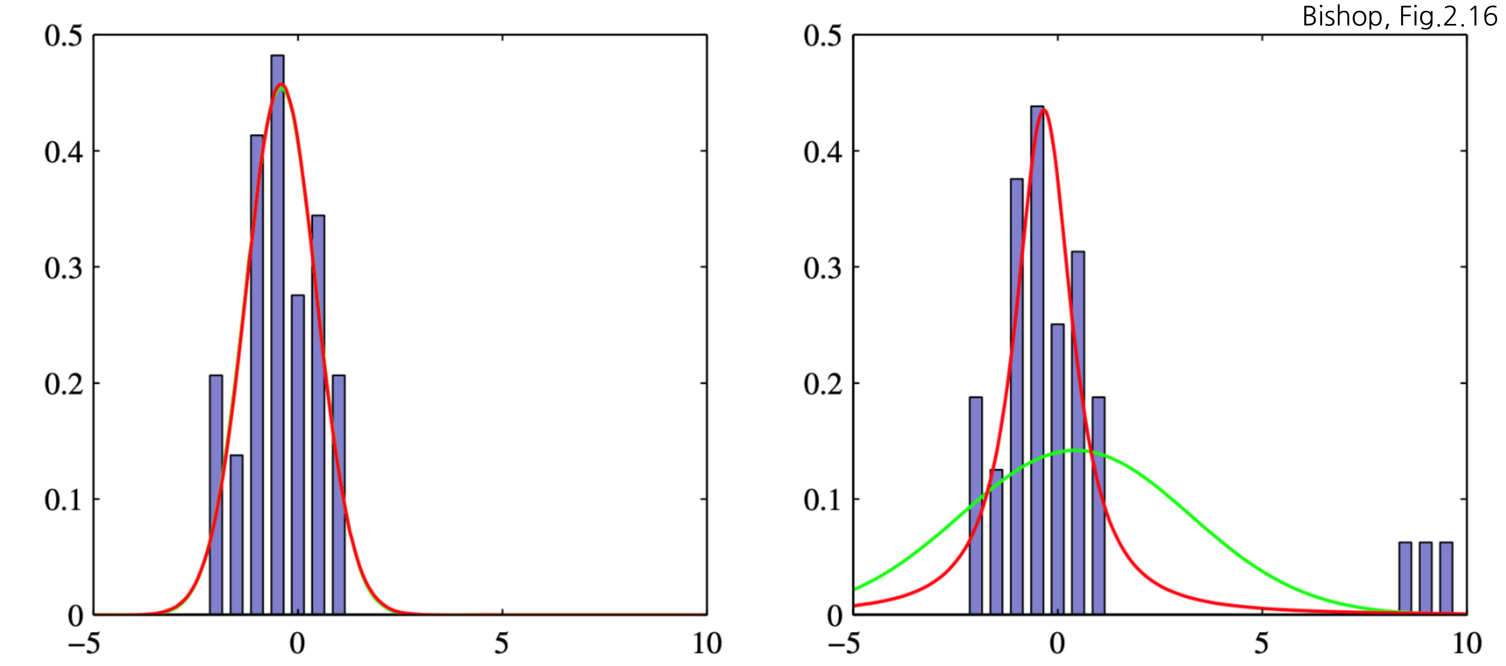
$$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\mathrm{Gam}(\tau|a, b)d\tau$$

- 가우시안과 감마 사전분포가 주어졌을 때, 정밀도를 적분해서 없애면 x에 대한 주변분포를 구할 수 있다.
  - 의미: 같은 평균과 다른 정밀도를 가진 무한히 많은 가우시안 분포들을 합산한 분포
  - 매개변수를 조절함으로써 코시분포나 가우시안도 됨.

$$\mathrm{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2}$$

1-4 스튜던트 t-분포

- 가우시안을 무한히 혼합한것이어서 robust해진다.
  - =outlier에 대해 덜 예민하게 된다.



Bishop, Fig.2.16

(좌) 가우시안 분포로부터 추출한 30개의 데이터포인트들의 히스토그램. 최대가능도해를 이용하여 두 분포를 근사. t분포가 빨간색, 가우시안 분포가 녹색. 거의 동일하다.

가우시안분포는 t분포의 일종이다

(우) 세개의 outlier를 추가. t분포의 강건성이 돋보인다.

1-4 스튜던트 t-분포

- 최대가능도해는 EM으로 구한다.
- 다변량으로 확장도 가능하다.
  - 식은…

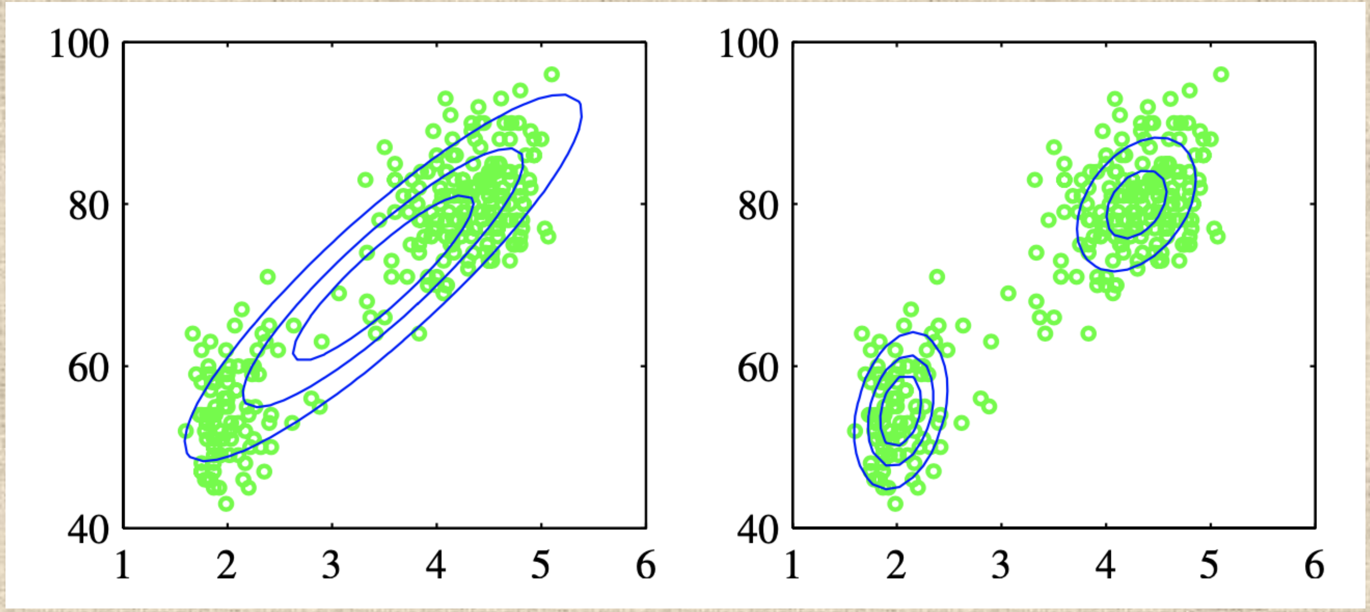
1-5 주기적 변수

- 가우시안을 모델로 사용하는 것이 적절하지 않은 경우
- 극좌표계를 이용해서 나타낸다.
- 가우시안 분포를 주기적 변수에 적용할 수 있도록 일반화한 분포: 폰 마이제스 분포(von Mises distribution) 혹은 원형 정규분포(circular normal distribution)

1-6 가우시안 분포의 혼합

- 실제 데이터집합은 대체로 복잡하다. 예) ‘오래된 믿음’ old faithful

Bishop, Fig.2.21

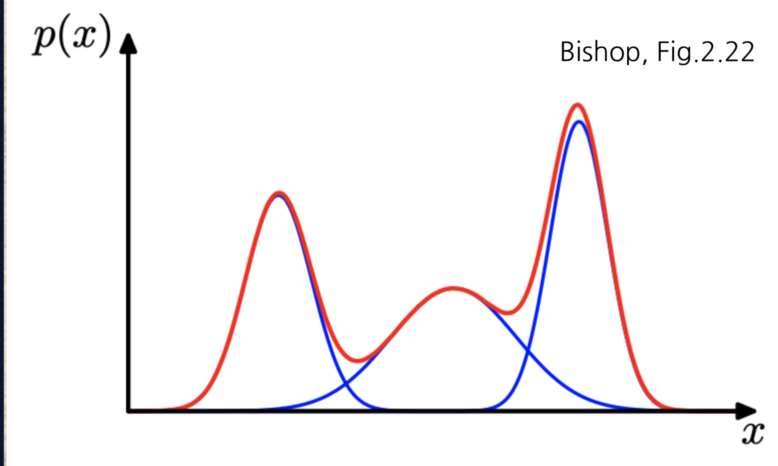


옐로스톤 국립공원의 간헐온천 분화에 관한 데이터. x축은 분화가 지속된 시간, y축은 다음 분화까지의 시간. 좌측은 단일 가우시안 분포를 이용해서 근사한것, 오른쪽은 가우시안 두개의 선형결합분포를 근사한 것

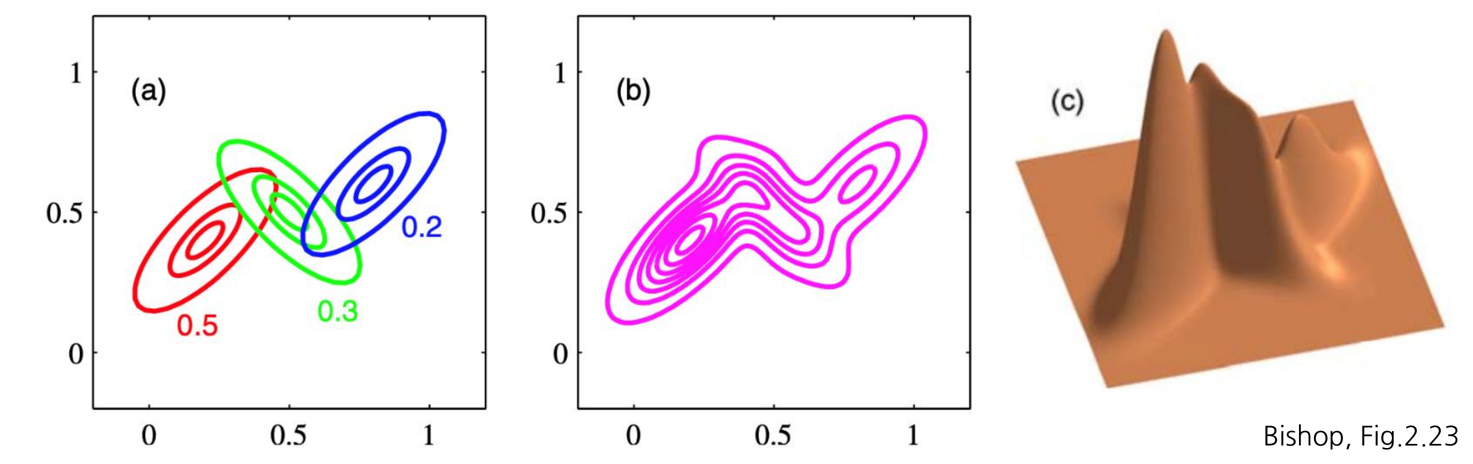
1-6 가우시안 분포의 혼합

- 가우시안 혼합분포 = K개의 가우시안 밀도의 중첩

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \sum_{k=1}^{K} \pi_k = 1, \qquad 0 \le \pi_k \le 1$$



Bishop, Fig.2.22



Bishop, Fig.2.23

1-6 가우시안 분포의 혼합

- 로그가능도 함수

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

  - 로그안에 합산… 곤란하다.
  - (이미 우리가 배웠음) 최대가능도방법의 해를 구하기가 매우 힘들므로, EM으로 구한다.

1-7 지수족

- 더 큰 분류. 앞서 본 분포들의 대부분이 지수족(exponential family)의 일종이다.
- x에 대한 지수족의 분포는 다음과 같이 정의

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

  - $\mathbf{x}$:벡터,스칼라,연속,이산 모두 가능
  - η: 자연매개변수, $\mathbf{u}(\mathbf{x})$: $\mathbf{x}$에 관한 함수

1-7 지수족

- $g(\boldsymbol{\eta})$는 분포를 정규화하는 계수가 된다. 따라서,

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

- 지수족의 예: 베르누이분포

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^{x}(1-\mu)^{1-x}$$

$$= \exp(x \ln \mu + (1-x) \ln(1-\mu))$$

$$= (1-\mu) \exp\left\{\ln\left(\frac{\mu}{1-\mu}\right) x\right\}$$

1-7 지수족

- 지수족의 정의와 방금 구한 식을 비교하면,

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\} = (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}$$

  - $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$
  - μ에 대해 정리하면, $\mu = \frac{1}{1+\exp(-\eta)}$
    - ‘로지스틱 시그모이드’$(= \sigma(\cdot))$

1-7 지수족

- 따라서 베르누이 분포를 아래와 같이 나타낼 수 있다.

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x) \quad \left(\because 1 - \sigma(\eta) = \sigma(-\eta)\right)$$

- 그 외에, $u(x) = x,\ h(x) = 1,\ g(\eta) = \sigma(-\eta)$
- 다음으로 다항분포에 대해서,

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

1-7 지수족

- $x = \{x_1, \ldots, x_M)^T, \eta_k = \ln \mu_k, \boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^T$로 두면,

  $p(\mathbf{x}|\boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^{\mathbf{T}}\mathbf{x})$로 나타낼 수 있다.

  - $h(\mathbf{x}) = 1,\ g(\boldsymbol{\eta}) = 1,\ \mathbf{u}(\mathbf{x}) = \mathbf{x}, \sum_{k=1}^{M} \mu_k = 1$

- 결과적으로 분포는,

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

1-7 지수족

- 그려면 $\eta_k = \ln(\frac{\mu_k}{1-\sum_j \mu_j})$

- μ_k에 대해 정리하면, $\mu_k = \frac{\exp(\eta_k)}{1+\sum_j \exp(\eta_j)}$. 곧, softmax.

- 다시 정리하면, (베르누이분포 할때 과정과 동일)

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

1-7 지수족

- 가우시안 분포도 지수족이다.

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

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

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

1-7 지수족

- 지수족 분포에서 $\boldsymbol{\eta}$ 를 추정하기 위해 지수족 분포식을 미분.

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

- 예를들어, iid인 $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ 을 고려해보면,

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

$$\frac{\partial}{\partial \boldsymbol{\eta}} \ln\big(p(\mathbf{X}|\boldsymbol{\eta})\big) = 0$$ 계산하면, $$-\nabla \ln g(\boldsymbol{\eta}_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$$

1-7 켤레사전분포

- “베이즈 확률론에서 사후확률을 계산함에 있어
  사후 확률이 사전 확률 분포와 같은 분포 계열에 속하는 경우
  그 사전확률분포를 켤레 사전분포(Conjugate Prior) 라고 부른다.
  켤레 사전분포를 이용하면 사전확률분포의 파라미터를 업데이트하는
  방식으로 사후확률을 계산할 수 있게 되어 계산이 간편해진다.”(위키백과)

1-7 켤레사전분포

- 베르누이 분포는 베타분포가 켤레 사전분포
- 가우시안
  - 평균에 대한 켤레사전분포: 가우시안
  - 정밀도에 대한 켤레사전분포: 위샤트분포
- 일반적으로 켤레 사전분포를 찾는 것이 가능하다.
  - 모든 지수족분포는 다음형태의 켤레사전분포가 존재한다.

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

1-7 켤레사전분포

- 직접 곱해보면 나옴.

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

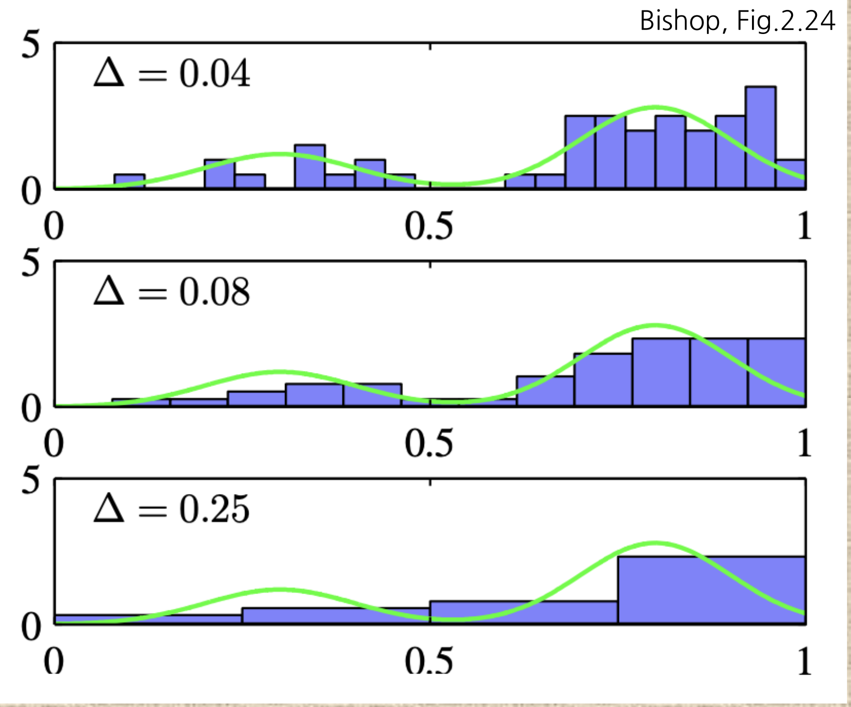
1-8 비매개변수적 방법

- 분포의 형태에 대해 가정을 최소화: non-parametric 밀도추정
  - 비매개변수적 방법이기 때문에 빈도학파의 방법론이 대다수
- 가장 직관적이고 쉬운 방법 : 히스토그램 밀도 추정
  - 데이터를 구간으로 나누고 구간에 속한 데이터 숫자 세기.
  - 구간별 정규화를 거치고 나면 확률로 해석 가능하다.

$$p_i = \frac{n_i}{N\Delta_i}$$

1-8 비매개변수적 방법

- 히스토그램 밀도추정(계속)
  - 구간 너비에 따라 결과가 달라진다.(단점)



◀ 녹색선이 원 분포. 히스토그램이 추출된 데이터로 그린것. 구간이 좁으면 원 분포에 없는 구조(뾰족한 부분)가 생긴다. 너무 넓으면 양봉형태를 표현하는데 실패한다.

1-8 비매개변수적 방법

- 히스토그램 밀도추정(계속)
  - 적절한 구간너비(Δ)를 결정하는 것이 쉽지 않다.
  - 구간의 가장자리로 인해 *불연속*면이 생긴다.
  - 고차원데이터를 다루는 것이 거의 불가능에 가깝다.
    - D차원 공간이라고 하면 변수를 M개의 구간으로 나누었을 때 총 구간'조각'수는 M^D. 따라서 데이터가 어마어마하게 많이 필요하다.

1-8 비매개변수적 방법

- 특정구역 $\mathcal{R}$에 해당하는 확률질량

$$P = \int_{\mathcal{R}} p(\mathbf{x})d\mathbf{x} \qquad (p(\mathbf{x})\text{를 추정하는것이 목표})$$

- $p(\mathbf{x})$로부터 N개를 관측. 각각의 데이터는 구역 $\mathcal{R}$에 속할 확률이 P. 그러면 총 K개의 포인트가 구역에 존재할 확률=

$$\mathrm{Bin}(K|N,P) = \frac{N!}{K!\,(N-K)!}P^K(1-P)^{N-K}$$

1-8 비매개변수적 방법

- $K \approx NP$
- $\mathcal{R}$의 부피를 V라 하고, 한 구역내에서 $p(\mathbf{x})$는 대략 상수.
  따라서 $P \approx p(\mathbf{x})V$
- 그러면, $p(\mathbf{x}) = \frac{K}{NV}$ 를 얻음.
- K를 고정시키고 V를 데이터로부터 얻든지(KNN),
  그 반대로 할 수 있음(커널밀도추정).

1-8 커널밀도추정

- 구역R(한 변의 길이가 h인 입방체)에 포함되는 데이터포인트 수K라 하면,

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right), \qquad k(\mathbf{u}) = \begin{cases} 1, i = 1 \dots D, |u_i| \le 1/2 \\ 0, \quad \text{그 외} \end{cases}$$

  - $k(\mathbf{u})$: 커널함수. Parzen window라고 부르기도 함.
  - $\mathbf{x}$를 중심으로 한변이 h인 입방체 안에 $\mathbf{x}_n$이 존재하면 1.

1-8 커널밀도추정

- $\mathrm{p}(\mathbf{x}) = K/NV$에 대입하면,

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x_n}}{h}\right) \quad \because V = h^D$$

- 불연속면이 존재. 매끄러운 확률분포를 위해 더 매끄러운 커널을 사용한다. 예-가우시안함수

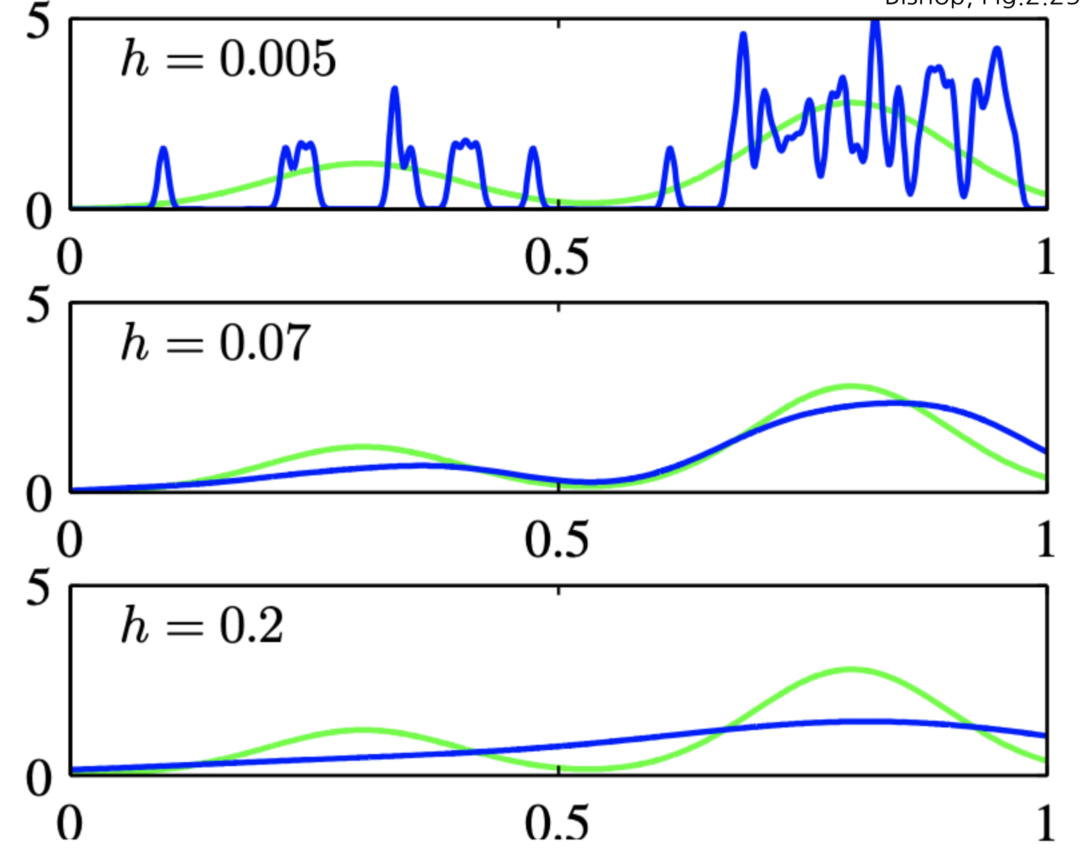
$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{D/2}}\exp\left(-\frac{\|\mathbf{x}-\mathbf{x_n}\|^2}{2h^2}\right)$$

1-8 커널밀도추정

- 각각의 데이터포인트에 가우시안을 위치시키고 모두 합한 후 정규화 한 것과 같음.

▶h가 평활매개변수로 작용한다. h가 너무 작으면 노이즈가 심한 모델을 얻고, 너무 크면 원 분포의 특성(양봉)을 표현하지 못한다.

Bishop, Fig.2.25



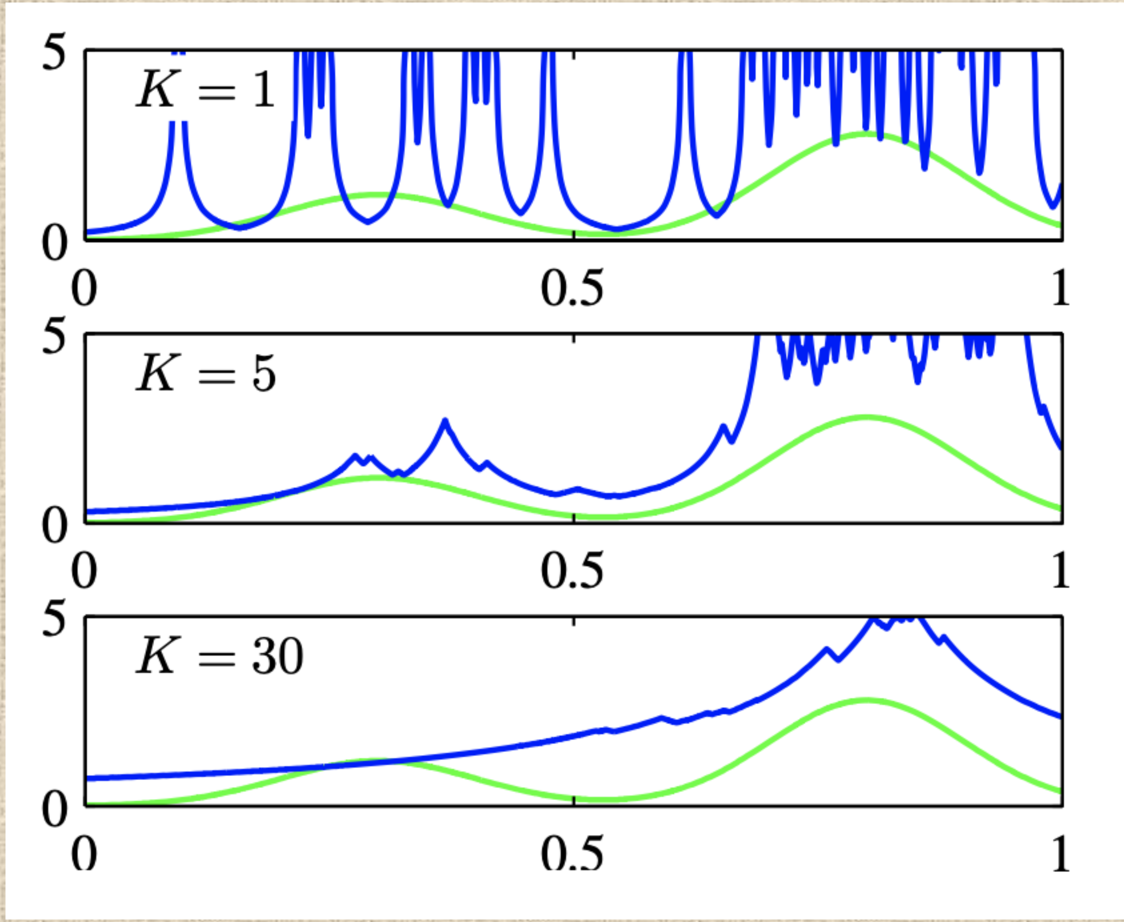
1-8 커널밀도추정

- 커널의 조건
  - $k(\mathbf{u}) \geq 0$
  - $\int k(\mathbf{u})d\mathbf{u} = 1$
- 추정에 복잡한 계산이 필요 없다.
  - 모든 데이터포인트를 저장하기만 하면 된다.
  - 데이터 포인트의 크기와 복잡도증가가 선형.

1-8 K최근접이웃

- 커널추정의 단점: h가 모든 커널에 대해 동일
- 포인트 $\mathbf{x}$주변의 작은 구(sphere)를 가정.
- 구의 크기는 일정하지 않다.
  - 구가 정확히 K개의 포인트를 포함할때까지 구를 키움.
  - K값이 평활화(degree of smoothing)의 정도를 결정한다.
- 이렇게 구한 모델은 모든공간에 대해 적분하면 발산
  - 확률모델이라고 할 수 없다.

1-8 K최근접이웃



◀커널밀도추정에 사용되었던 데이터로 K최근접이웃밀도 추정.

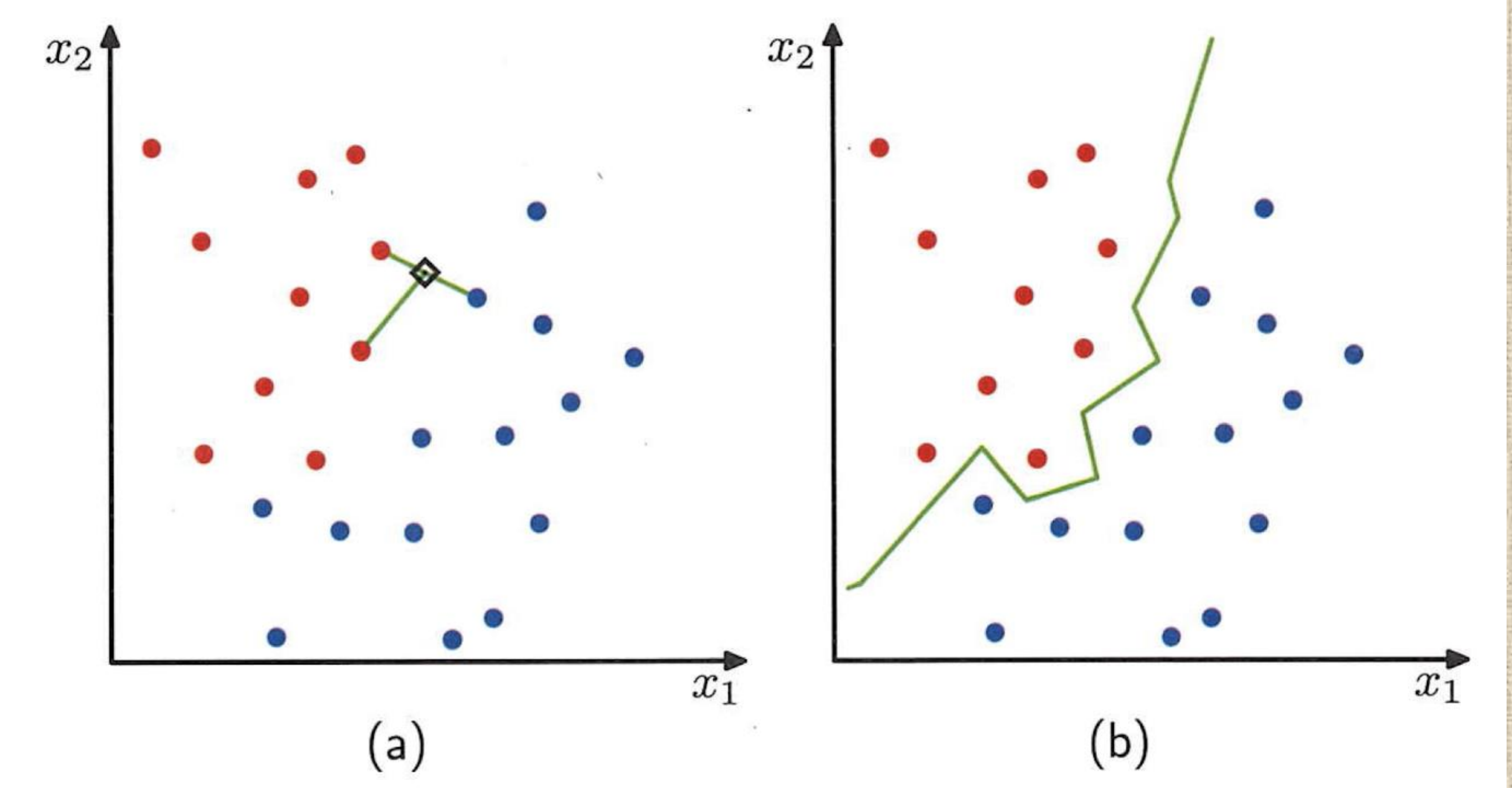
커널밀도추정에서 h(평활매개변수)가 했던 역할을 K가 하고 있다.

1-8 K최근접이웃

- 분류문제에 응용
  - 시험포인트 **x**주변에 K개의 이웃을 조사해서 그 포인트 중 가장 많은 포인트가 속한 클래스로 **x**를 분류.
  - K=1인 경우 '최근접 이웃'방법이 된다.
  - 작은 K값을 사용하면 각 클래스에 해당하는 많은 구역.
  - 큰 K값을 사용하면 더 적은 수의 큰 구역.
- (최근접이웃방법의 오류) < (최적분류기(=답안)의 오류*2)

1-8 K최근접이웃

▶ (a)의 경우 K=3일때이다.
(b)의 경우 K=1.
즉, '최근접 이웃'방법.
이때 선택경계는 서로다른 클래스에 속하는 두 점을 수직으로 이등분하는 초평면들로 이루어진다.



종강

그동안
수고하셨습니다.