Probabilistic Deep Learning:

2. Distributions & Exponential Family

ESC 2024 Winter Session 2주차

Contents

1
Distributions

Binary Variables

베르누이(*Bernoulli*) 분포

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

$$E[x] = \mu(2.3)\ var[x] = \mu(1-\mu)$$

관찰 데이터 D가 주어졌을 때 가능도 함수

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

Maximum Likelihood Estimator 구하기

$$lnp(D|\mu) = \sum_{n=1}^{N} lnp(x_n|\mu) = \sum_{n=1}^{N} x_n ln\mu + (1-x_n)ln(1-\mu)$$

$$\mu_M L = 1/N \sum_{n=1}^{N} x_n = m/N$$

이항 분포(*binomial distribution*)

$$Bin(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

The Beta Distribution

기존의 *MLE*가 아닌 베이지안 방식으로 파라미터를 추정하는 것을 살펴볼 것이다.

그 과정에서 필요하는 부분이 바로 파라미터의 사전(prior) 확률 값 p(μ)을 구하는 것이다.

베이지안 방식에서는 가능도 함수 값을 최대로 하는 파라미터 값을 구하는 문제가 아니라 사후 분포(posterior)를 최대화하는 문제로 풀게 된다.

이 때 사후 분포는 가능도 함수로 유추된 p(x|μ) 와 이 때의 p(μ) 사전 분포의 곱에 비례한다.

$$p(\mu|x) \propto p(x|\mu)p(\mu)$$

- 사후 분포는 확률 분포이다. 따라서 우리가 알고 있는 일반적인 분포의 형태로 만들어야 계산이 편하다. (예를 들면 정규분포, 이항분포 등)
- 사전 분포도 확률 분포이다. 따라서 사전 확률 분포를 도입하되 이것 또한 우리가 알고 있는 분포로 만들어야 계산이 편하다.
- 이런 경우 사후 분포와 사전 분포를 같은 분포의 형태로 만들어 사용하면 어떨까? 이러면 전체적으로 계산이 쉬워질 것이다.

▶ 이러한 속성을 **공액(conjugation)**이라고 한다.

(어떤 확률 분포들의 곱이 앞서 사용된 분포와 동일한 분포의 형태를 가지는 것)

이런 이유에서 **베타 분포(Beta distribution)**가 매우 유용하게 사용된다.

- 이항 분포의 형태를 취하는 가능도 함수를 사용하는 경우 사전 확률로 베타 분포를 사용하면 베타 분포의 사후 확률을 얻을 수 있다.
- 이는 이항 분포와 베타 분포가 서로 공액적 관계에 놓여있는 분포들이기 때문이다.
- 즉, (베타분포 = 이항분포 X 베타분포) 와 같은 형태의 식을 얻을 수 있다.
- 베타 분포는 다음과 같다.

$$Beta(\mu|a,b) = \Gamma(a+b)/\Gamma(a)\Gamma(b)\mu^{a-1}(1-\mu)^{b-1}$$

- 식에서 사용된 베타 분포는 앞서 설명한 이항 분포의 공액 분포를 설명하는 것으로, 주 변수는 바로 이항 분포의 ***u*** 가 된다.
- 따라서 이 베타 분포의 모수 a, b는 초모수(hyper-parameter) 또는 하이퍼-파라미터라고 부른다.

동전 던지기의 사후 분포(posterior)

$$p(\mu|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

공액(conjugation)적 특성이 나타나고 있다.

(모수의 사전 분포를 베타 분포로 썼더니 사후 분포도 베타 분포가 된다는 것)

사후 확률 분포식은 다음과 같다.

$$p(\mu|m,l,a,b) = \Gamma(m+a+l+b)/\Gamma(m+a)\Gamma(l+b)\mu^{m+a-1}(1-\mu)^{l+b-1}$$

위의 식에서 m 은 x=1 인 경우의 횟수, l 은 x=0 경우의 횟수를 의미한다.

The Beta Distribution

Sequential approach to learning

순차(sequential)적으로 확률 분포를 업데이트하는 모델

사후 분포를 다음 사건이 발생할 때, 사전 분포로 활용을 해도 문제가 없다.

따라서 데이터 한 건 한 건 반영될 때 마다 업데이트 모델을 만들어낼 수 있다.

- 예를 들어 동전 던지기라면 임의로 던진 동전 한개가 앞면일지 뒷면일지 예측하는 문제.
- 물론 부가적으로 모수값이 추정되어야만 앞면과 뒷면이 나올 확률값을 예측할 수 있지만, 어쨌거나 최종 필요한 것은 모수값은 아니다.
- 이런 모델이 왜 좋나면 값이 고정된 모수를 선택해서 예측하는 것이 아니라 모수 자체를 확률 변수로 놓고 영향을 줄 수 있는 모든 모수의 가능성을 염두해 둔 수식을 만들어 원하는 예측값을 만들어낼 수 있다.

확률값을 추정할 수 있는 식을 만들어낼 수 있다.

$$p(x=1|D) = \int_0^1 p(x=1|\mu)p(\mu|D)d\mu = \int_0^1 \mu p(\mu|D)d\mu = E[\mu|D]$$

- 식을 보면 그냥 지금까지 예측된 데이터로 인해 만들어진 x 에 대한 평균값을 구하기만 하면 끝이다.
- 위 식에서 데이터가 매우 커져서 m,l이 발산하면 결국 이 식은 *MLE* 결과와 동일해진다.
  - 즉, 사후 분포에서 사전 분포의 영향력이 사라짐.
- 데이터의 크기가 유한한 경우 *MLE* 로 얻어진 모수와 사전 분포로 얻어진 모수의 중간 어딘가에 사후 분포로 예측한 모수값이 오게 된다.
- 데이터가 많아지면 많아질수록 사후 분포의 피크는 점점 날카롭게 오르게 된다.
  - 결국 분산 값이 작아지게 되고 Non-uniform 한 분포의 형태를 가지게 된다.

▶ 우리가 궁금한 점은 베이지언 방식을 취할 때 관찰 데이터가 많아질수록 사후 분포의 불확실성은 줄어들게 되는지 여부이다.

- 만약 관찰 데이터 집합 D 의 파라미터 θ 를 추론한다고 하자.
- 이 때 p(θ,D) 의 결합 분포로 인해 만들어지는 분포는 다음과 같게 된다.
- 이 식을 *Law of total expectation* 이라고 한다.

$$E[\theta] = E_D[E[\theta|D]]$$

- 마찬가지로 분산에 대해서도 이렇게 전개 가능하다.
- 이 식을 *Law of total variance* 라고 부른다.

$$var[\theta] = E_D[var[\theta|D]] + var_D[E[\theta|D]]$$

- 왼쪽은 θ 에 대한 사전 분산값이다.
- 첫번 째 텀은 θ 의 사후 분산값에 대한 평균이다.
- 두번 째 텀은 θ 의 사후 평균에 대한 분산이다.
- 분산의 값은 양의 값을 가지기 때문에 첫번 째 텀은 왼쪽 값보다 작거나 같다.
- 즉 사후분포의 불확실성은 줄어든다.

Multinomial Variables

K개의 가능성 중 하나를 선택하는 문제

$x_k = 1$일때의 확률을 모수 μ_k를 이용하여 표현하면 $p(x_k = 1) = \mu_k$ 가 된다.

이 식으로 하나의 데이터 x 에 대한 확률 값은 다음과 같이 정의 가능하다.

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

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1 \qquad (2.27)$$

$$E[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^T = \boldsymbol{\mu} \qquad (2.28)$$

• 이제 모든 관찰값(즉, 샘플)에 대한 확률 값을 고려해보자. (가능도 함수는 이렇게 정의된다.)

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

- 식을 보면 가능도 함수는 K개의 종류를 가지는 샘플 N개에만 의존한다.

$$m_k = \sum_n x_{nk}$$

이를 **충분 통계**(*sufficient statistics*)라고 한다.

- 모수 정보를 모두 포함한 식을 충분통계량이라고 한다.
- 이런 경우 모수 대신 이 식을 이용해서 분포를 표현 가능하다.

이제 가능도 함수(likelihood)를 이용해서 모수를 추정하자.

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

$$\mu_k^{ML} = \frac{m_k}{N}$$

- 이제 m1,...mK를 가지는 결합 분포를 고려해보자. 그러면 식은 다음과 같아진다.

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

• 이게 바로 **다항분포**(*multinomial distribution*) 이다.

이항 분포를 K개의 그룹으로 확장한 것이다.

The Dirichlet Distribution

다항 분포에 어울리는 사전 분포

다항 분포에 대한 공액사전 분포로 디리슈레 분포가 사용된다.

• 다항 분포의 모수 추정을 위한 모수의 사전 분포로 다음과 같은 형태를 생각하면 된다.

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

디리슈레(*Dirichlet*)분포의 표현은 다음과 같다.

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

• 이제 가능도함수(likelihood)를 살펴보도록 하자.

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

• 디리슈레 분포가 공액 사전 분포로 사용되었기 때문에 사후분포(posterior distibution) 또한 디리슈레 분포를 따를 것이라는 걸 알 수 있다.

• 따라서 이를 계산하면 다음과 같다.

$$p(\boldsymbol{\mu}|D,\boldsymbol{\alpha}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) = \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{a_k+m_k-1}$$

식을 봐서도 짐작할 수 있지만 K=2 인 경우 베타 분포와 동일해진다.

2

The Gaussian Distribution

Multivariate Gaussian Distribution, Bayes' Theorem, Sequential Estimation

Multivariate Gaussian Distribution

· Gaussian Distribution (single variable x)

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

· Multivariate Gaussian Distribution (D-dimensional vector x)

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

· μ : D-dimensional mean vector

· Σ : $D \times D$ covariance matrix

· Mahalanobis distance

$$\Delta^2 = (x-\mu)^T\Sigma^{-1}(x-\mu)$$

· $\Sigma \rightarrow I$ 이면 Δ는 Euclidean distance (즉, 평균과의 거리 측정에 분산을 고려)

· x에 대한 가우시안의 함수적 종속성(Functional Dependence)을 담당

· quadratic form

Conditional Gaussian Distributions

WTS : $p(x_a, x_b)$가 가우시안 분포를 따르면 $p(x_a|x_b)$도 가우시안 분포를 따른다.

$pf)$ $\mathcal{N}(x|\mu, \sigma^2)$를 따르는 D차원의 벡터 x를 각각 M차원과 $D-M$차원의 두 개의 집합 $x = (x_a, x_b)$으로 나누어 확인한다.

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
&\quad -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
&\quad -\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).
\end{aligned}
$$

- mean vector $\mu = \begin{pmatrix}\mu_a \\ \mu_b\end{pmatrix}$
- covariance matrix $\Sigma = \begin{pmatrix}\Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb}\end{pmatrix}$
- precision matrix $\Lambda = \Sigma^{-1} = \begin{pmatrix}\Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb}\end{pmatrix}$

x_b를 상수로 가정하고 전개하면, x_a에 대한 quadratic form

따라서 $p(x_a, x_b)$는 가우시안 분포를 따를 것이라고 예측할 수 있다.

Marginal Gaussian Distributions

WTS : $p(x_a, x_b)$가 가우시안 분포를 따르면 x_a에 대한 주변 확률 분포인 $p(x_a)$도 가우시안 분포를 따른다.

$pf)$ x_a에 대한 주변 확률 분포는 $p(x_a) = \int p(x_a, x_b) dx_b$이므로 $p(x_a, x_b)$를 x_b에 대해 나타내고 적분한다.

$$\mathbf{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)$$ 이라고 하면, $p(x_a, x_b)$는

$$-\frac{1}{2}\mathbf{x}_b^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^{T}\mathbf{m} = -\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m})^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}$$

이것을 x_b에 대해 적분하면 exp(quadratic)의 형태가 나오므로, $p(x_a)$는 가우시안 분포를 따른다.

Bayes' theorem for Gaussian variables

· Gaussian marginal distribution $\boldsymbol{p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})}$

· Gaussian conditional distribution $\boldsymbol{p(y|x) = \mathcal{N}(Ax + b|\mu, L^{-1})}$

· $\boldsymbol{z = (x, y)}$

WTS : $\boldsymbol{E[z] = \begin{pmatrix} \mu \\ A\mu - b \end{pmatrix}}$, $\boldsymbol{cov[z] = R^{-1} = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}}$, $\boldsymbol{E[y] = A\mu + b}$, $\boldsymbol{cov[y] = L^T + A\Lambda^{-1}A^T}$

$pf)$ by Bayes' theorem, $p(z) = p(x)p(y|x)$

로그를 씌워 전개하고 z에 대한 quadratic form으로 식을 변형한다.

$$-\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{y}$$

$$= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{R}\mathbf{z}$$

Maximum Likelihood for the Gaussian

· 관찰 데이터 집합 $X = (x_1, x_2, \cdots, x_n)^T$이 주어졌을 때, 데이터 x_i들은 서로 독립적으로 발현된다. (i.i.d)

· 각각의 관찰 데이터는 가우시안 분포를 따르게 되며, 이를 가능도 함수(likelihood)로 이용할 때에는 보통 로그를 취해 사용한다.

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

미분을 통해 최대값을 가질 때 평균과 공분산을 구하면,

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

Sequential Estimation

· 관찰 데이터 집합이 매우 큰 경우에 사용하기 좋다.

· 한 번에 한 개의 데이터 샘플을 연산하고 버리는 과정을 반복한다.

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\
&= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)})
\end{aligned}
$$

· $N-1$개의 데이터로부터 추정된 $\mu_{ML}^{(N-1)}$과 N번째 관측 데이터를 활용

· N이 증가할수록 새로 관측되는 데이터의 기여도는 감소

· 언제나 이런 방식으로 식을 유도할 수 있는 것은 아니다

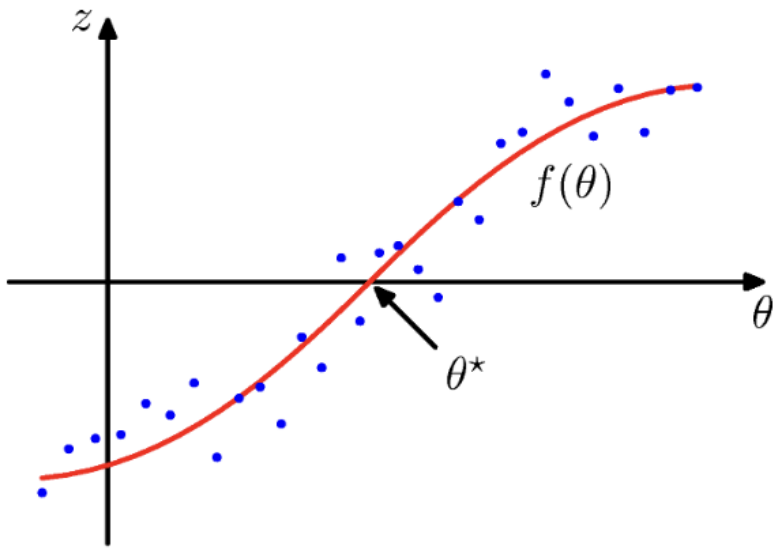
Sequential Estimation - Robbins & Monro Algorithm

· More general formulation of sequential learning

· 결합 분포 $p(\theta, z)$에 대한 랜덤 변수 θ와 z가 주어짐

· θ가 주어졌을 때, z에 대한 평균 함수를 정의하면 : $f(\theta) = E[z|\theta] = \int zp(z|\theta)dz$

· 이때 $f(\theta^*) = 0$을 만족하는 θ^*를 찾는 것이 목표가 된다.



· 주어진 θ에 대한 관찰 데이터 z값이 순차적으로 하나씩 업데이트 된다고 가정

· 아래 수열에 데이터를 순차적으로 입력하여 θ^*을 추정할 수 있다. (Regression)

$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1} z(\theta^{(N-1)})$$

단, a_N이 다음 조건을 만족해야 한다

1) $\lim_{N\to\infty} a_N = 0$: θ가 특정 값에 수렴

2) $\sum_{N=1}^{\infty} a_N = \infty$: θ^*를 찾기도 전에 임의 값에 수렴하지 않도록

3) $\sum_{N=1}^{\infty} a_N^2 < \infty$ 축적되는 노이즈는 유한 (수렴성을 보장)

Bayesian inference for the Gaussian

다양한 상황에서 가우시안 분포가 주어졌을 때의 베이지안 추론 방식 (Single Gaussian random variable x)

1) 분산(σ^2)을 알고 있을 때 평균값(μ)의 추론

N개의 관찰 데이터 $\mathrm{X} = (x_1, \cdots, x_N)^T$가 주어졌을 때 가능도 함수(likelihood function)는 다음과 같다

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

이때 가능도 함수가 μ에 대한 이차 형식이므로,

이때 μ에 대한 사전 확률 함수(prior) $p(\mu)$를 가우시안 분포를 따르게끔 고른다면 (이것을 **conjugate prior distribution**이라고 함)

사후 확률 분포(posterior)는 μ에 대한 exp(quadratic) 형태의 두 곱이므로 마찬가지로 가우시안 분포를 따른다

따라서 사전 확률 분포를 $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$**로** 잡아주면, 사후 확률 분포는 $p(\mu|X) \propto p(X|\mu)p(\mu)$에 따라

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

(이때 $\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$, $\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$)

Bayesian inference for the Gaussian

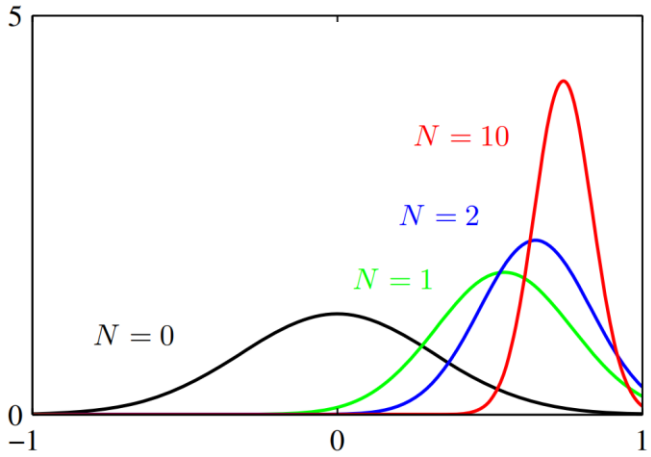
다양한 상황에서 가우시안 분포가 주어졌을 때의 베이지안 추론 방식 (Single Gaussian random variable x)

1) 분산(σ^2)을 알고 있을 때 평균값(μ)의 추론

1. 베이지안 추론을 통해 얻어진 평균값 $\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$과 MLE를 통해 얻은 평균값 $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$의 비교

   · $N = 0$이면 $\mu_N = \mu_0$ (최초 설정한 평균값)

   · $N \to \infty$이면 $\mu_N \to \mu_{ML}$ (MLE의 결과값)



2. 베이지안 추론을 통해 얻어진 분산 $\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$에 대한 고찰

   · $N = 0$이면 $\sigma_N = \sigma_0$

   · $N \to \infty$이면 $\sigma_N \to 0$ (정확도가 계속 증가함을 의미)

3. D차원에 대한 평균값의 추론은 sequential update formula 활용

$$p(\boldsymbol{\mu}|D) \propto \left[p(\boldsymbol{\mu}) \prod_{n=1}^{N-1} p(\mathbf{x}_n|\boldsymbol{\mu}) \right] p(\mathbf{x}_N|\boldsymbol{\mu})$$

· N번째 데이터를 식에서 분리

· 앞서 살펴본 sequential estimation 방법을 그대로 적용

Bayesian inference for the Gaussian

2) 평균값을 알고 있을 때 분산의 추론

- 앞서 평균을 구할 때 분산값을 고정되어 있다고 가정했던 것처럼, 분산을 추론할 때에는 고정된 평균값을 가정한다.
- 실제 계산에서는 공분산의 역수(정확도, precision)를 구하는 것이 편리하므로, $\lambda = 1/\sigma^2$으로 정의한다.
- 마찬가지로 계산의 편의성을 위해 conjugate prior distribution을 사용하는데, 이는 감마 분포 $Gam(\lambda|a_0, b_0)$를 도입한다.

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

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du, \qquad E[\lambda] = \frac{a}{b}, \qquad var[\lambda] = \frac{a}{b^2}$$

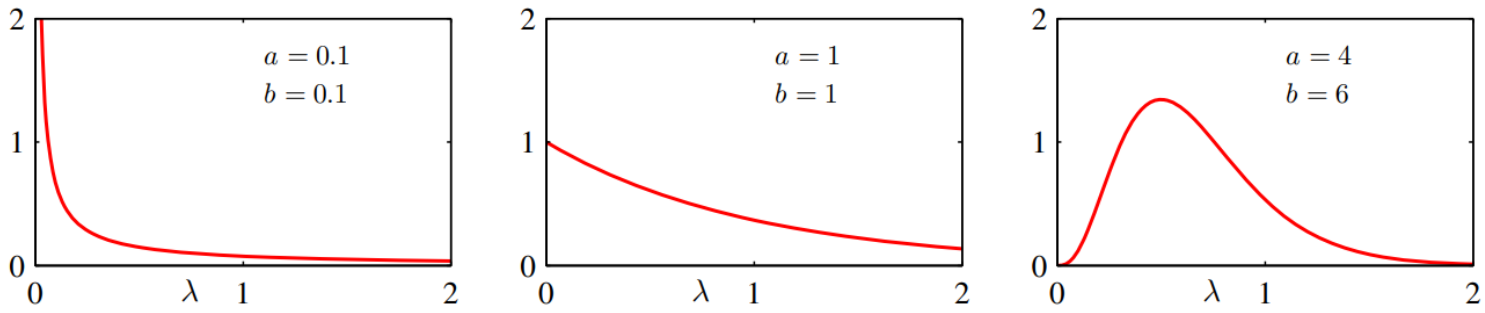


Figure 2.13 Plot of the gamma distribution $\mathrm{Gam}(\lambda|a,b)$ defined by (2.146) for various values of the parameters a and b.

Bayesian inference for the Gaussian

2) 평균값을 알고 있을 때 분산의 추론

이때 λ에 대한 가능도 함수(likelihood function)의 형태는 다음과 같다

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

사전 확률 함수로 감마 분포 $Gam(\lambda|a_0, b_0)$를 도입했으므로 여기에 가능도 함수를 곱해 사후 확률 분포(posterior)를 추론하면

마찬가지로 감마 분포 $Gam(\lambda|a_N, b_N)$ 를 따르게 된다. ($\because$ Conjugacy)

이때 a_N, b_N은 아래와 같다.

$$a_N = a_0 + \frac{N}{2}$$

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

- $N = 2a_0$이면 $a_N = 2a_0, b_N = 2b_0$
- effective number : $N = 2a_0$
  - 관찰 데이터의 영향력이 지정된 사전 확률의 영향력을 넘어서는 지점

Bayesian inference for the Gaussian

3) 평균과 분산을 둘 다 모를 때의 두 값에 대한 추론

먼저 가능도 함수에서 μ와 λ에 대한 의존도를 확인해보자면, 식을 완전히 분리할 수 없기에 전개를 통해 살펴보아야 한다.

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

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

이때 추론해야 하는 모수는 2개이므로 평균과 분산을 동시에 랜덤 변수로 고려한 $p(\mu,\lambda)$가 사전 확률이 된다.

여기서 $p(\mu,\lambda)$의 분포를 가능도 함수에서의 μ와 λ에 대한 의존성을 그대로 유지하게끔 설정하자. (Conjugacy)

$$p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\beta}\exp\{c\lambda\mu - d\lambda\}$$

$$= \exp\left\{-\frac{\beta\lambda}{2}(\mu-c/\beta)^2\right\}\lambda^{\beta/2}\exp\left\{-\left(d-\frac{c^2}{2\beta}\right)\lambda\right\}$$

- c, d, β는 상수

Bayesian inference for the Gaussian

3) 평균과 분산을 둘 다 모를 때의 두 값에 대한 추론

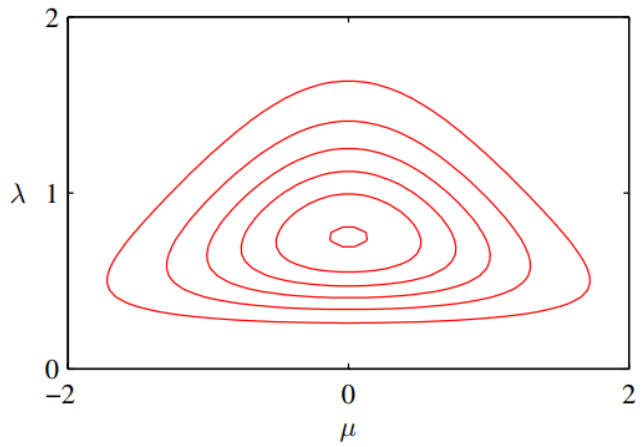
이때 앞서 살펴본 결합 확률 분포식에서 $p(\mu,\lambda) = p(\mu|\lambda)p(\lambda)$이 성립하므로, 다음과 같이 기술할 수 있다.

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

대입하여 전개해보면, $\mu_0 = \frac{c}{\beta}, a = \frac{(1+\beta)}{2}, b = d - \frac{c^2}{2\beta}$ 이 되고,

이러한 분포의 형태를 normal - gamma 또는 Gaussian - gamma distribution이라고 한다.

Figure 2.14 Contour plot of the normal-gamma distribution (2.154) for parameter values $\mu_0 = 0$, $\beta = 2$, $a = 5$ and $b = 6$.



Bayesian inference for the Gaussian

3) 평균과 분산을 둘 다 모를 때의 두 값에 대한 추론 - Multivariate Gaussian distribution $\mathcal{N}(x|\mu, \Lambda^{-1})$ for a D-dimensional vector x

- Conjugate prior로 아래의 Wishart distribution을 도입한다.(ν는 자유도, B는 정규화 상수)

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu) = B|\mathbf{\Lambda}|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{\Lambda})\right)$$

- 마찬가지로 Conjugate prior를 다음과 같이 기술할 수도 있다. (normal - Wishart 또는 Gaussian - Wishart 분포)

$$p(\boldsymbol{\mu}, \mathbf{\Lambda}|\boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\mathbf{\Lambda})^{-1})\,\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu)$$

Periodic Variables

Gaussian Distribution은 보편적으로 많이 활용되는 분포이나, 특정 경우에 대해서는 전혀 어울리지 않을 수 있다.

Periodic variable이란 일정한 단위를 두고 값이 반복되는 형태의 함수값으로, 위에 대한 적절한 예시가 된다.

Ex 1. 특정 위치에서의 바람의 방향

Ex 2. 하루 단위, 또는 연간 단위의 주기를 갖는 모델

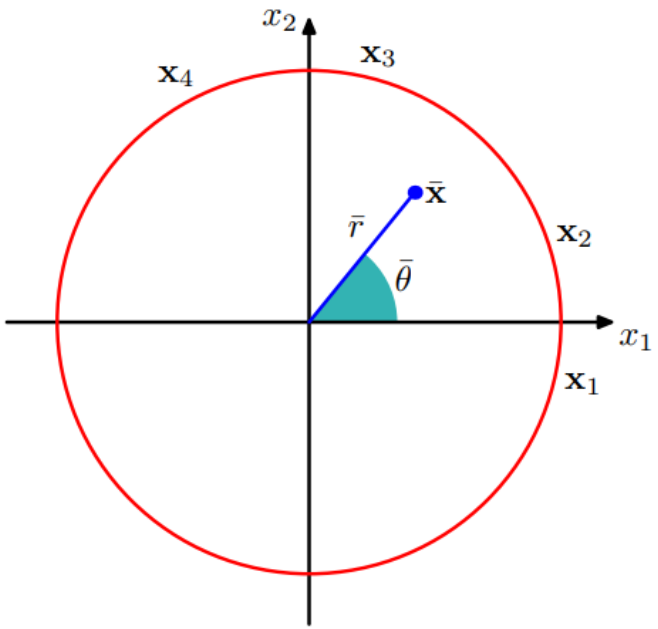
주기성 변수의 관찰 데이터를 $D = \theta_1, \theta_2, \cdots \theta_n$이라고 하고, 단위 원 내의 한 점으로 나타낸다.

이때 이 점들에 대한 평균 값은 $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$ (데카르트 좌표에 의한 값들의 평균을 나타냄)

대입해서 전개해보면 $\bar{x} = (\bar{r}\,cos\bar{\theta},\, \bar{r}\,sin\bar{\theta}) = (\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n, \frac{1}{N}\sum_{n=1}^{N}\sin\theta_n)$

$$\therefore\ \bar{\theta} = tan^{-1}\left\{\frac{\sum sin\,\theta_n}{\sum cos\,\theta_n}\right\}$$

이 결과는 주기성 변수에 대한 MLE 결과와도 일치한다. (von Mises 분포 참고)

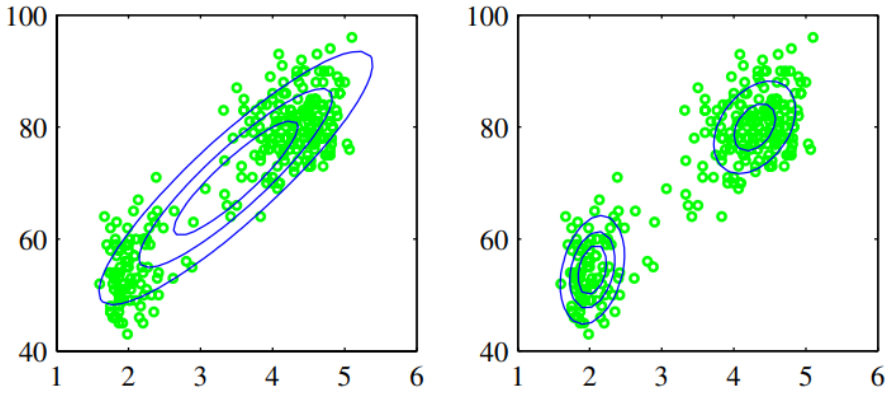


Mixtures of Gaussians

현실적으로 가우시안 분포만을 적용하기 어려운 경우가 다수 존재한다.

오른쪽 자료는 옐로스톤 국립공원에 있는 간헐천의 화산 폭발 데이터로,
크게 두 개의 지배적인 집단을 형성하고 있다.

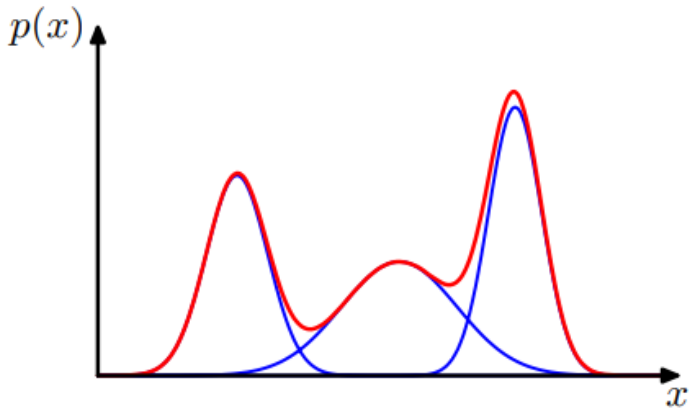
따라서 왼쪽 그림처럼 하나의 가우시안 분포로 데이터를 표현한 것은 잘못된 예시.
∴ 오른쪽 그림처럼 가우시안 분포의 중첩을 통해 더 정확한 표현이 가능함.



⇨ **혼합 가우시안 모델 (Mixture Gaussian Distribution)**

- 선형 결합을 통해 매우 복잡한 밀도의 표현이 가능
- 충분한 개수의 결합으로 선형결합의 계수뿐만 아니라 평균과 공분산을 조절

하여 거의 대부분의 연속 밀도를 임의의 정확도로 근사 가능함



Mixtures of Gaussians

K개의 중첩 형태는 다음과 같으며, 이를 혼합 가우시안 분포(Mixtures of Gaussians)이라고 한다.

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

- $\mathcal{N}(x|x_k, \Sigma_k)$: component
- π_k : mixing coefficients ($\sum \pi_k = 1, \;\; 0 \leq \pi_k \leq 1 \;\; \because nomalization$)

Marginal density를 전개하면,

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

위 식과 비교하면 $p(k) = \pi_k$, $p(x|k) = \mathcal{N}(x|x_k, \Sigma_k)$으로 고려할 수 있고,

Posterior probability $p(k|x)$의 값을 유도해낼 수 있다. (이 값은 responsibility라고 불린다.)

(이에 대한 자세한 내용은 Ch.9의 EM 알고리즘에서 다룬다.)

3

The Exponential Family

Maximum Likelihood and Sufficient Statistics, Conjugate Priors, Noninformative Priors

The Exponential Family

지수족에 속하는 분포는 다음과 같은 형태의 일반화된 형태로 표현 가능하다.

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

η는 natural parameters(자연 모수)

$\mathbf{u}(\mathbf{x})$는 $\mathbf{x}$에 대한 함수

$g(\eta)$는 확률분포에 대한 정규화 계수 (확률의 전체 합을 1로 맞추는데 사용됨)

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

베르누이 분포는 지수족에 속할까?

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

$p(x|\mu)$를 적절히 변형하면 다음과 같이 표현 가능하다.

$$p(x|\mu) = exp\{x\ ln\mu\ + (1-x)ln(1-\mu)\} = (1-\mu)exp\{ln(\frac{\mu}{1-\mu})x\}$$

이때, $\eta,\ g(\eta),\ u(x),\ h(x)$를 아래와 같이 설정.

$\eta = ln(\frac{\mu}{1-\mu})$

$g(\eta) = \frac{1}{1+exp(-\eta)}$

$u(x) = x$

$h(x) = 1$

위와 같이 설정해주면 베르누이 분포를 다음과 같이 표현 가능

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

따라서, 베르누이 분포는 지수족에 속한다.

The Exponential Family

다항 분포는 지수족에 속할까?

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

이때, $\eta,\ g(\eta),\ u(x),\ h(x)$를 아래와 같이 설정.

$\eta_k = ln \; \mu_k, \quad \eta = (\eta_1, \dots, \eta_M)^T$

$g(\eta) = 1$

$u(x) = \mathbf{x}$

$h(x) = 1$

위와 같이 설정해주면 다항 분포를 다음과 같이 표현 가능

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

따라서, 다항 분포는 지수족에 속한다.

정규 분포는 지수족에 속할까?

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

이때, $\eta,\ g(\eta),\ u(x),\ h(x)$를 아래와 같이 설정.

$\eta = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$

$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$

$h(x) = (2\pi)^{-1/2}$

$g(\eta) = (-2\eta_2)^{1/2} exp(\eta_1^2/4\eta_2)$

위와 같이 설정해주면 다항 분포를 다음과 같이 표현 가능

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

따라서. 정규 분포는 지수족에 속한다.

Maximum Likelihood and Sufficient Statistics

지수족 분포에서 MLE를 사용하여 파라미터 벡터 η를 추정해보자.

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

위 식을 η에 대해서 미분해보면 다음과 같다.

$$\nabla g(\eta) \int h(\mathbf{x}) exp\{\eta^T \mathbf{u}(x)\}\, d\mathbf{x} + g(\eta) \int h(\mathbf{x}) exp\{\eta^T \mathbf{u}(x)\} \mathbf{u}(x)\, d\mathbf{x} = 0$$

정리하면 다음과 같은 식을 얻을 수 있다.

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

서로 독립인 관찰 데이터 $\mathbf{X} = \mathbf{x}_1, ..., \mathbf{x}_N$이 주어졌을 때, likelihood function은 다음과 같다.

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

위 식에 로그를 붙인 뒤, η에 대해 미분한 값이 0이 된다고 전개하면 η_{ML}을 구할 수 있다.

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

위 식을 보면 η_{ML}이 $\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$에만 의존한다는 것을 알 수 있다.
이때, $\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$을 분포에 대한 $sufficient\ statistics$ (충분 통계) 라고 한다.
충분 통계란 모수 값을 완전히 설명할 수 있는 최소한의 함수식이다.

따라서 모든 데이터를 다 저장해서 사용하지 않고, 위 식에 나온 결과만을 사용하여 모수 추정에 활용하게 된다.

EX1) 베르누이 분포에서 $\mathbf{u}(\mathbf{x})$는 $\mathbf{x}$이므로, 단지 x의 합만을 유지하면 됨.
EX2) 가우시안 분포에서 $\mathbf{u}(\mathbf{x})$는 $(x, x^2)^T$이므로, 단지 x의 합과 x^2의 합만을 유지하면 됨.

N이 무한대로 갈 경우, $\frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$이 $E[\mathbf{u}(\mathbf{x})]$가 된다. 따라서, $\eta_{ML} = \eta$ 임을 알 수 있다.

Conjugate Priors

지수족에 속하는 분포의 $conjugate\ prior$는 다음과 같이 표현할 수 있다.

$$p(x|\chi, v) = f(\chi, v)g(\eta)^v exp\{v\eta^T\chi\}$$

$f(\chi, v)$는 정규화 계수

$g(\eta)$는 앞서 사용된 식과 동일 (확률의 전체 합을 1로 맞추는데 사용됨)

앞서 살펴봤던 $likelihood$는 다음과 같다.

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

$conjugate\ prior$와 $likelihood$를 곱해서 만든 $posterior$는 다음과 같다.

$$p(\eta|\mathbf{X}, \chi, v) \propto g(\eta)^{v+N} exp\{\eta^T(\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}) + x\chi)\}$$

$prior$와 $posterior$가 동일한 형태의 식이므로 conjugacy를 확인할 수 있다.

이때, v는 effective number

Noninformative Priors

확률 추정의 문제 중 일부는 사전 분포의 모수 값을 손쉽게 지정 가능하다.
그러나 **사전 분포의 형태를 전혀 예측하기 어려운** 경우도 존재한다.
이때, 사전 분포로 ***Noninformative priors***를 사용한다.

가장 쉬운 방법은 $p(x|\lambda)$처럼 모수 λ가 주어진 경우, $p(\lambda) = const$ 로 설정하는 것
ex) λ가 K개의 상태를 가지는 이산 변수라면 $p(\lambda)$의 확률 값으로 1/K 사용

그러나 2가지의 문제점이 존재한다.

1. λ의 범위가 unbounded인 경우

이 경우, 단순한 상수 값을 사용하게 되면 분포를 전체 λ 영역에 적분하게 되면 값이 ∞이 된다.
이처럼, 확률 함수의 합이 1이 아니라 ∞이 되는 경우, ***improper prior***라고 함

2. 비선형 변환되는 변수를 분포에 적용하는 경우

$h(\lambda) = const$ 라고 가정하자. 이 경우, $\lambda = \eta^2$라고 한다면 $h(\eta^2)$ 역시 상수이다.
새로운 함수 $\hat{h}(\eta) = h(\eta^2)$ 라고 한다면, $\hat{h}(\eta)$ 역시 상수이다.

그러나, 위와 같은 자연스러운 변환은 확률 식에서 통용되지 않는다.
$p_\lambda(\lambda) = const$ 라고 할 때, η의 식으로 전개하면 다음과 같다.

$$p_\eta(\eta) = p_\lambda(\lambda) \left| \frac{d\lambda}{d\eta} \right| = p_\lambda(\eta^2) 2\eta \propto \eta$$

위의 식을 보면, λ에 대해서는 상수였던 확률이 η에 대해서는 상수가 아니라 선형 함수이다.

따라서, 상수 값을 가지는 사전 분포는 알맞은 상황에만 사용해야 한다.

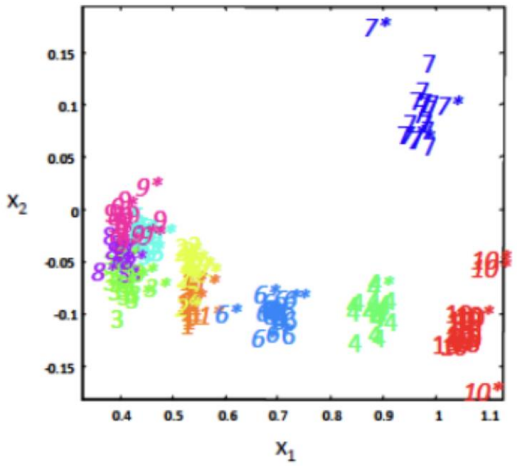
4

Nonparametric Methods

Kernel Density Estimators, Nearest-Neighbor Methods

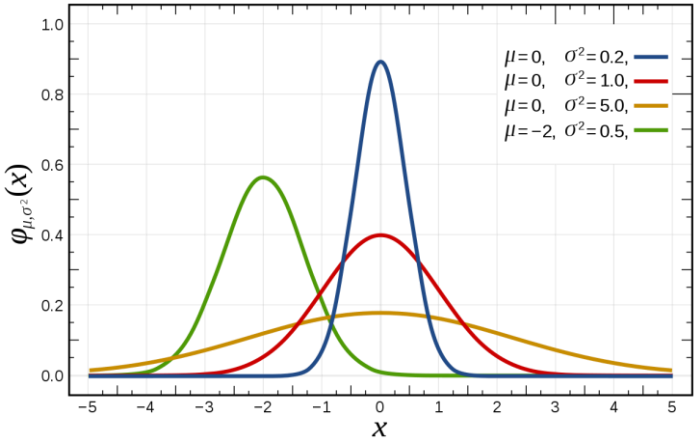
Parametric vs. Nonparametric Approach

- Parametric approach: data로부터 추정할 수 있는 적은 수의 파라미터를 가지는 특정한 형태의 함수를 확률 분포로 사용해 확률 밀도를 추정한다.
  - 표현할 수 있는 분포에 제한이 있어, 선택한 분포가 data generating 분포를 잘 설명하지 못하면 예측력이 떨어지는 한계가 있다.

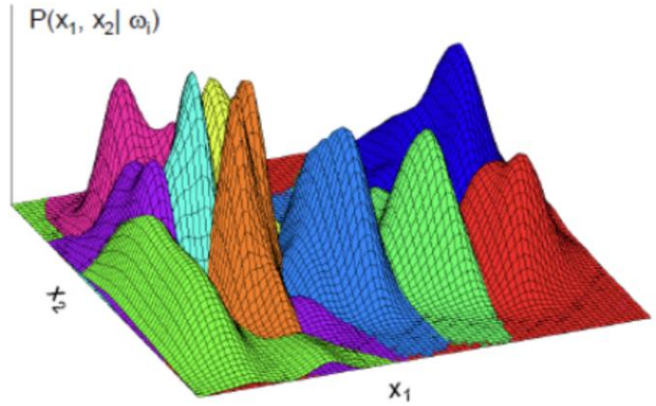


<data>

Density Estimation



<parametric approach>



<nonparametric approach>

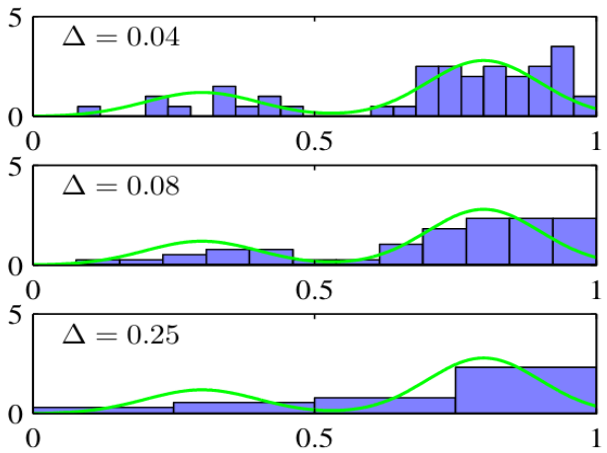
- Nonparametric approach: 분포에 대한 가정을 거의 하지 않고 확률 밀도를 추정한다.

Histogram Methods

- 임의의 확률 분포로부터 추출된 일변량 연속형 확률변수 x 가 있다.
  x 의 확률 밀도를 추정하기 위해 x를 특정 너비 Δ_i 를 가지는 구간(bin)으로 나누고, 특정 구간에 속한 데이터의 개수를 n_i라 하자.
  이 때의 확률밀도 p_i 는 다음과 같이 나타낼 수 있다. (일반적으로, $\Delta_i = \Delta$)

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

- Δ의 값에 따라 추정된 확률 밀도의 smooth한 정도가 달라짐을 볼 수 있다.
  이 때 Δ를 **smoothing parameter**라 한다. 확률 밀도를 잘 추정하기 위해서는 smoothing parameter인 Δ의 값을 잘 정하는 것이 중요하다.

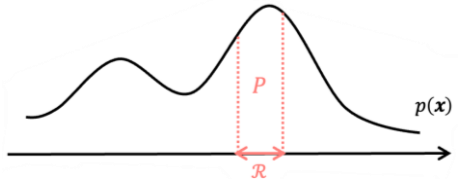


- 그러나 (1) 추정된 밀도가 불연속적이고, (2) 데이터의 차원에 따른 bin의 수가 기하급수적으로 증가(D차원 데이터를 M 개의 bin으로 나누면 전체 bin의 개수는 M^D개)하기 때문에, 히스토그램을 확률 밀도 추정에 활용하는 데 한계가 있다.
- 한 편, 히스토그램을 통해 비모수적 접근법에 대해 얻을 수 있는 두 가지 교훈이 있다.
  (1) **Concept of Locality**: 특정 위치에서의 확률밀도를 추정하기 위해서는 근처의 데이터를 고려해야만 하며, 따라서 (Euclidean) distance measure이 도입되어야 한다.
  (2) **Smoothing parameter**로 적절한 값을 선택해야 한다.

Introduction to Nonparametric Approach

- 관측치 $\boldsymbol{x}$ 가 D차원의 확률 밀도 $p(\boldsymbol{x})$로부터 나왔을 때, $p(\boldsymbol{x})$의 값을 추정하고자 한다. Locality에 따라, $\boldsymbol{x}$ 를 포함하는 어떤 작은 영역 $\mathcal{R}$에 대한 확률 P 는 다음과 같이 나타낼 수 있다.

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}$$



- 총 N개의 관측치 중 K개가 $\mathcal{R}$에 포함될 확률은 이항분포에 의해,

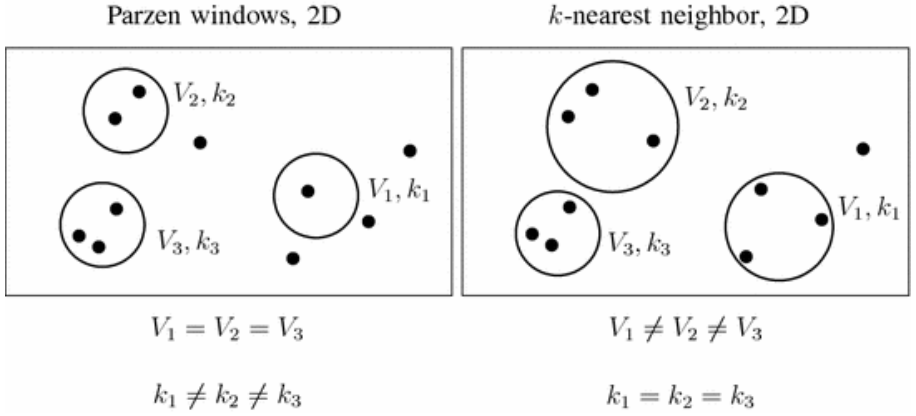
$$Bin(K|N,P) = \frac{N!}{K!(N-K)!} P^K (1-P)^{1-K}$$

- 이 때, $E\left[\frac{K}{N}\right] = P$, $Var\left[\frac{K}{N}\right] = \frac{P(1-P)}{N}$ 이다.

  $N \to \infty$ 이면 $Var\left[\frac{K}{N}\right] \to 0$ 이고 분포가 평균에 몰리므로, $K \approx NP$로 근사할 수 있다.
- 한 편, $\mathcal{R}$이 충분히 작아 $\mathcal{R}$ 내의 $p(\boldsymbol{x})$가 상수가 된다면, $\mathrm{P} \approx p(\boldsymbol{x})V$로 근사할 수 있다. ($V = volume\ of\ \mathcal{R}$)
- 따라서 $p(\boldsymbol{x}) = \frac{K}{NV}$ 이다.

$$p(\mathbf{x}) = \frac{K}{NV}$$

- 위 식에서 K를 고정하고 데이터를 통해 V를 결정하면 K-Nearest-Neighbor approach, V를 고정하고 K를 결정하면 Kernel approach이다.



Kernel Density Estimators

- $\boldsymbol{x}$ 를 중심으로 하는 hypercube $\mathcal{R}$ 에 포함되는 관측치의 개수 K 를 구하기 위해 아래와 같은 함수를 정의하자.

$$k(\mathbf{u}) = \begin{cases} 1, & |u_i| \leq 1/2 \ , i = 1, ..., D, \\ 0, & \text{otherwise} \end{cases}$$

- $k(\boldsymbol{u})$는 커널 함수의 일종인 Parzen window 라 한다. $\mathcal{R}$의 한 변의 길이가 h일 때, $\boldsymbol{x}_n$이 $\mathcal{R}$에 포함되면 $k\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)$ 는 1, 그렇지 않으면 0을 반환하므로, $\mathcal{R}$에 포함되는 관측치의 수 K는 다음과 같다.

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x_n}}{h}\right)$$

- $p(\boldsymbol{x}) = \frac{K}{NV}$ 에 의해, 확률밀도 $p(\boldsymbol{x})$는 다음과 같다. $(h^D = V)$

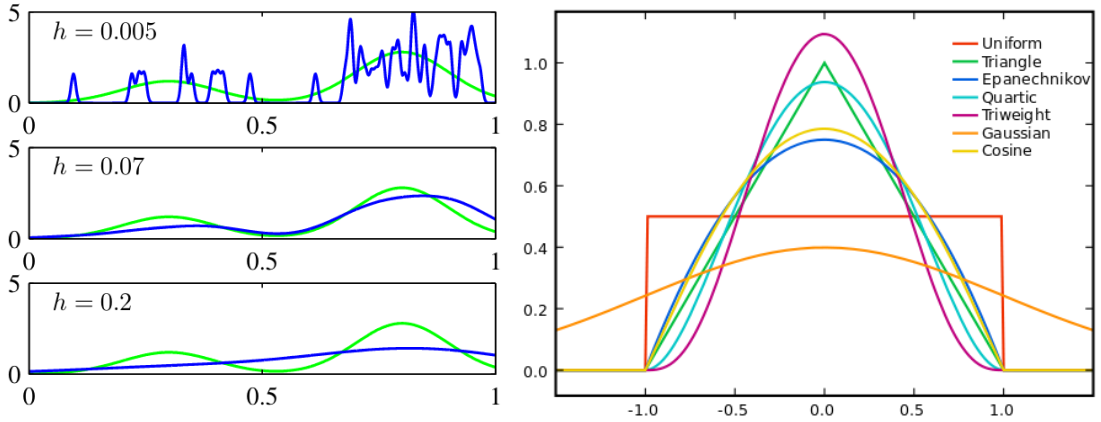
$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$

- Hypercube 경계에서 발생하는 불연속을 해결하기 위해 smoother kernel function을 도입할 수 있다. 주로 사용되는 Gaussian을 적용하면,

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

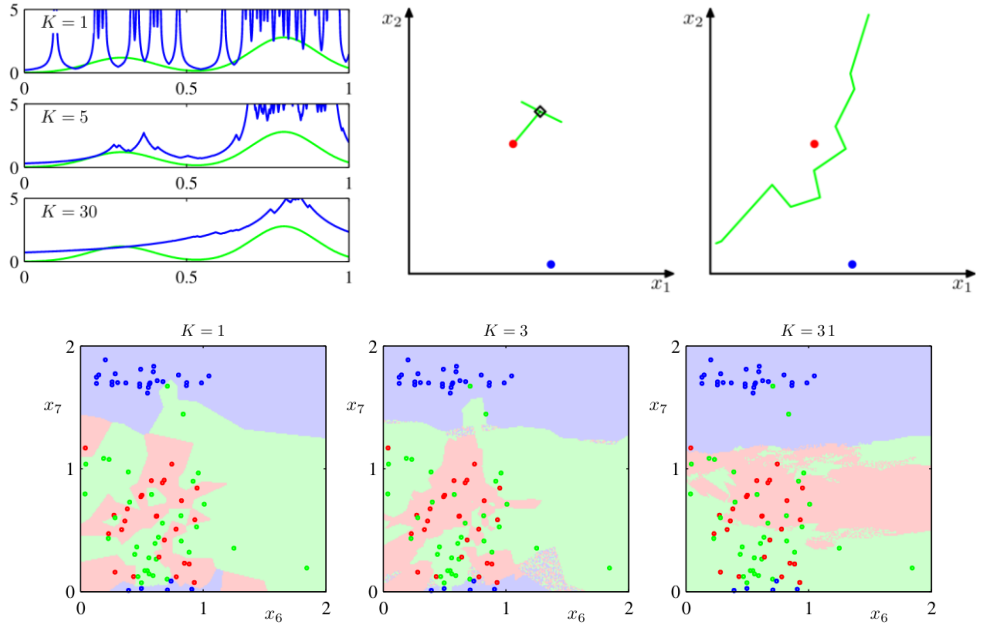
- Smoothing parameter인 h 값에 따라 추정된 분포의 모양이 달라짐을 확인할 수 있다.
- 한 편, $k(\boldsymbol{u})$ 형태의 함수가 아니더라도, 아래 두 조건을 만족한다면 어떤 함수든 커널 함수로 활용할 수 있으며, 주로 활용되는 커널 함수의 형태는 다음과 같다.

$$k(\mathbf{u}) \geq 0,$$
$$\int k(\mathbf{u}) d\mathbf{u} = 1$$



Nearest-Neighbor Methods

- Kernel methods에서 h의 최적값은 공간 내 데이터의 위치에 따라 달라질 수 있다. 이를 해결하기 위해 $p(\boldsymbol{x}) = \frac{K}{NV}$에서 K를 고정하고 V를 결정하는 방안을 고려해보자.
- $\boldsymbol{x}$ 를 중심으로 하는 구를 가정하고, 해당 구 안에 K개의 관측치가 포함될 때까지 반지름을 늘리면서 구의 부피 V를 결정한다. 이를 K-Nearest-Neighbor method라 한다. 이 때 smoothing parameter은 K가 된다.



- KNN을 활용해 분류 과제를 해결하는 과정을 살펴보자. 전체 데이터의 수는 N, 분류 클래스의 종류는 C_k, C_k에 포함되는 데이터의 수는 N_k 일 때, 새로운 데이터 $\boldsymbol{x}$가 속할 클래스를 찾고자 한다.
- $\boldsymbol{x}$를 중심으로 K개의 데이터를 포함하는 구의 부피를 V라 하자. 이 때,

$$p(\mathbf{x}|C_k) = \frac{K_k}{N_k V}$$
$$p(C_k) = \frac{N_k}{N}, p(\mathbf{x}) = \frac{K}{NV}$$

베이즈 정리에 의해, $p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} = \frac{K_k}{K}$

- 즉, 구 안의 데이터 중 가장 많은 데이터가 속한 클래스로 $\boldsymbol{x}$를 할당한다.

- 비모수적 접근은 간단한 계산으로 인한 이점이 있지만, 모든 데이터를 저장하고 있어야 하기 때문에 데이터의 크기가 큰 경우 한계가 발생한다. 따라서 이후 챕터에서는 유연하면서도 데이터의 크기와 무관하게 확률 밀도를 추정하는 방법을 다룰 예정이다.

감사합니다