ESC 2024 Winter Session 1st Week

Curve Fitting, Decision Theory, & Information Theory with Probability

복습 스터디 4조 심재윤 오동윤 이현우 전제훈 한지희

1. Distributions

Outline : 베이지안적 관점에서의 분포 추정을 알아보기 위해 먼저 흔히 사용되는 베르누이 분포의 Maximum likelihood function을 통해 모수를 추정하는 방식에 대해 알아보자.

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

모수 μ가 주어졌을 때, 베르누이 분포의 값 x가 나올 확률은 다음과 같이 구할 수 있다. 평균은 우리가 기존의 알고 있던 값인

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

가 된다. 우리가 이제 데이터 셋 $D = \{x_1, \ldots, x_N\}$이 있다고 가정해보자. 그 후, 모수 μ에 대한 가능도 함수(likelihood function)을 설정해보면,

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

이 됨을 알 수 있다. 적절한 모수 μ가 설정되어 있다면, 당연하게도 해당 가능도 함수의 값이 크게 나올 것이다. 해당 가능도를 계산해주기 편하게 likelihood function을 log likelihood function으로 변환해보면,

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

로 바꿀 수 있다. 우리는 $lnp(D|\mu)$를 최대화하는게 목표이고, 이를 미분을 통해 구해보면, $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$일 때 최대가 됨을 알 수 있다. ($m$은 $x = 1$인 개수) 고로, 적절한 모수 μ는 $\frac{m}{N}$이다.

2.1 The Beta Distribution

베르누이 분포를 통해 모수 추정하는 방식에 대해 알아보았다. 이에 대한 방식은 만약 데이터 셋이 편향되어 있다면, 좋지 못한 모수를 추정할 확률이 있다. 이에 대해 적절히 보정해주기 위해 베이지안에서는 사전 분포(Prior distribution)을 도입한다. 사전 분포는 모수 μ에 대한 분포로, 사전 분포가 정해지고 난 이후에는 사후 분포(Posterior distribution) $\mu^x(1-\mu)^{1-x}$를 통해 모수에 대한 분포 $p(\mu|x)$를 구하게 된다.

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

만약, 사후분포와 사전분포의 분포 형태가 비슷하다면 더 쉽게 계산을 할 수 있게 된다. 다음과 같은 속성을 conjugacy라고 한다. 베르누이 분포의 켤레 분포는 beta distribution이다.

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

동전 던지기를 해본다고 하자. 동전이 앞면이 나올지 뒷면을 나올지에 대한 확률은 베르누이 분포를 따른다. 동전 던지기의 사후 분포를 계산해보면,

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

사전 분포를 베타 분포로 사용하였더니, 사후 분포가 간단한 식으로 나옴을 알 수 있다. 이것이 conjuagation적 특성이 나타남을 알 수 있다.

2.2 Multinomial Variables
K개의 가능성 중 하나를 선택하는 문제이다. 만약, $x_k = 1$일 때의 확률은 $p(x_k = 1) = \mu_k$이다. 일반적으로 표현해보면,

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

과 같이 되며,

$$E[x|\mu] = \sum_x p(x|\mu)x = (\mu_1, \cdots, \mu_k)^T = \mu$$

$(x_1, \cdots, x_N)$의 데이터 셋 D가 있다고 하자. 그때의 likelihood function은

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

이 식을 보면 likelihood function은 K개의 종류를 가지는 샘플 N개에만 의존한다.

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

이와 같이 모수 정보를 모두 포함하는 식을 충분통계량(sufficient statistics)이라고 한다. 이제 likelihood functino을 이용해서 모수를 추정하자.

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

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

이제 $m_1, \cdots, m_k$를 가지는 결합 분포를 고려해보자. 그렇다면 식을 다음과 같아진다.

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

다음과 같이 다항분포(multinomial distribution)이 됨을 알 수 있다.

2.3 The Dirichlet Distribution
다항 분포에 대한 켤레분포로 디리슈레(Dirichlet)분포가 사용된다.

$$Dir(\mu|a) = \frac{\Gamma(a_0)}{\Gamma(a_1)\cdots\Gamma(a_K)} \prod_{k=1}^{K} \mu_k^{a_k-1}$$

likelihood function을 보면,

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

Dirichlet 분포 또한 사전분포로 사용하면 사후분포 또한 Dirichlet 분포가 됨을 알 수 있다. 이를 계산해보면,

$$p(\mu|D,a) = Dir(\mu|a+m) = \frac{\Gamma(a_0+N)}{\Gamma(a_1+m_1)\cdots\Gamma(a_K+m_K)} \prod_{k=1}^{K} \mu_k^{a_k+m_k-1}$$

다음과 같이 된다.

2. The Gaussian Distribution

일변량 가우시안 분포의 식은 다음과 같다.

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

다변량 가우시안 분포의 식은 다음과 같다.

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

(μ는 D 차원의 평균 벡터이며, $\sum$은 $D \times D$의 convariance 행렬이다.)
가우시안 분포는 $p(x_a, x_b)$가 가우시안 분포를 따르며 $p(x_a|x_b)$ 또한 가우시안 분포를 따르는 특징있으며, $p(x_a)$ 또한 가우시안 분포를 따른다. 해당 특징을 이용하여 가우시안 분포의 likelihood function을 구해보자.

$$lnp(X|\mu,\sum) = -\frac{ND}{2}ln(2\pi) - \frac{N}{2}ln|\sum| - \frac{1}{2}\sum^{N}(x_n-\mu)^T \sum^{-1}(x_n-\mu)$$

해당 likelihood function이 최대가 되는 μ와 $\sum$은

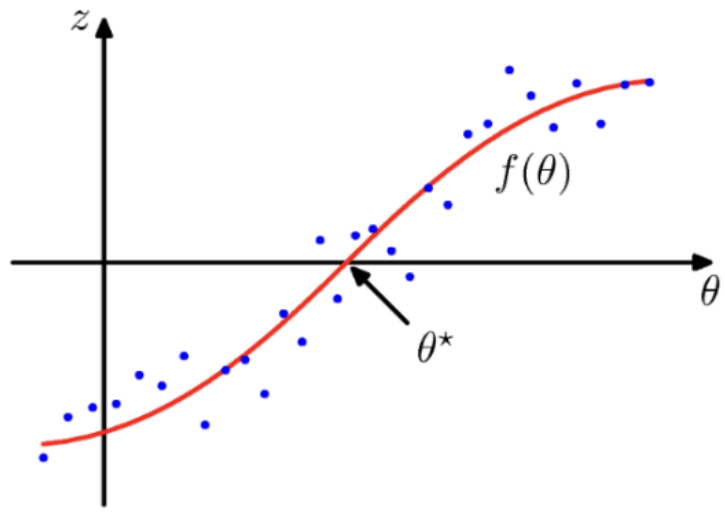
$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \sum_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})(x_n-\mu_{ML})^T$$

가 된다.

2.1 Sequential estimation
관찰 데이터의 집합이 큰 경우에는 어떻게 할까? 데이터의 집합이 크다면, 데이터 셋 전체를 한번에 계산하려고 하면 계산량이 많아져 오래 걸리게 된다. 이를 해결하고자 sequential estimation이 사용된다.

$$\mu_{ML}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}x_N + \frac{1}{N}\sum_{n=1}^{N-1} x_n = \mu_{ML}^{(n-1)} + \frac{1}{N}(x_N - \mu_{ML}^{(N-1)})$$

2.1.1 Robbins Monro Algorithm



결합 분포 $p(\theta, z)$가 있으며, θ가 주어졌을 때, $f(\theta) = E[z|\theta] = \int zp(z|\theta)dz$라고 하자. 이때, $f(\theta^*) = 0$을 만족하는 θ^*를 찾는 것이 목표가 된다.
앞서 언급한 sequential estimation을 통해 하나씩 업데이트 하여 추정할 수 있다.

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

a_N은 다음과 같은 조건을 만족해야 한다.
1) $\lim_{N\to} a_N = 0$: θ가 특정 값에 수렴
2) $\sum_{N=1} a_N = \infty$: θ^*를 찾기도 전에 임의의 값에 수렴하지 않도록
3) $\sum_{N=1} a_N^2 <$ 축적되는 노이즈는 유한하도록

2.2 Bayesian inference for the Gaussian

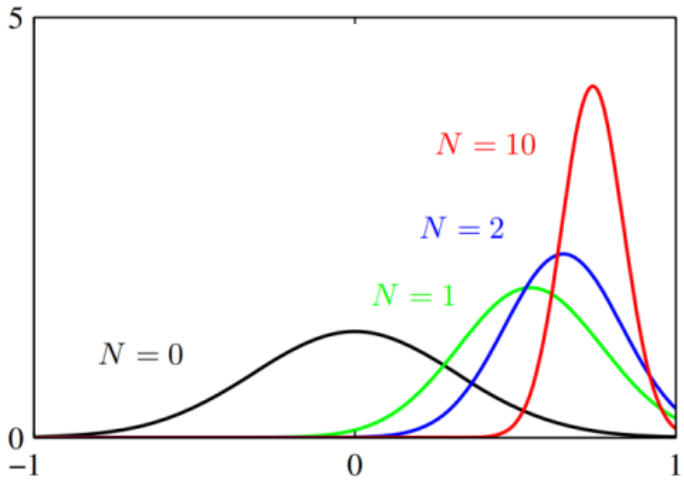
1) 분산(σ^2)를 알고 있을 때 평균값(μ)의 추론
N개의 관찰 데이터 $X = (x_1, \cdots, x_N)^T$가 주어졌을 때 likelihood function는 다음과 같다.

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

이때 likelihood function은 μ에 대한 이차 함수이므로, 이때 μ에 대한 사전 확률 함수를 가우시안 분포를 따르게끔 고른다면 사후 확률 분포 또한 가우시안 분포를 따르게 된다.
따라서 사전 확률 분포를 $p(\mu) = N(\mu|\mu_0, \sigma_0^2)$로 잡아주면, 사후 확률 분포는 $p(\mu|X) \propto p(X|\mu)p(\mu)$에 따라

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

$$(\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}, \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2})0$$



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

— $N = 0$이면 $\mu_n = \mu_0$

— $N \to$이면 $\mu_N \to \mu_{ML}$

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

— $N = 0$이면 $\sigma_N = \sigma_0$

— $N \to$이면 $\sigma_N \to 0$

D차원에 대한 평균값의 추론은 sequential update formula를 활용해보면,

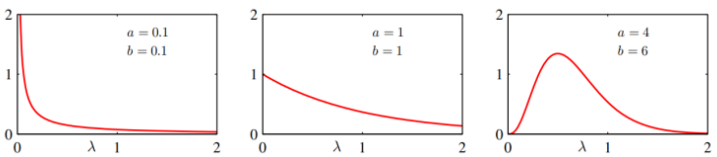
$$p(\mu|D) \propto [p(\mu)\prod_{n=1}^{N-1} p(x_n|\mu)]p(x_N|\mu)$$

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)을 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.

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

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



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

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

사전 확률 함수로 감마 분포 $Gam(\lambda|a_0, b_0)$를 도입했으므로 여기에 가능도 함수를 곱해 사후 확률 분포를 추론하면 마찬가지로 감마 분포 $Gam(\lambda|a_N, b_N)$를 따르게 된다. 이때, a_N, b_N은 아래와 같다.

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

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.

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

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.

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

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

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

대입하여 전개해보면, $\mu_0 = \frac{c}{\beta}$, $a = \frac{(1+\beta)}{2}$, $b = d - \frac{c^2}{2\beta}$이 된다. 이러한 분포의 형태를 normal-gamma 또는 Gaussian-gamma distribution이라고 한다.

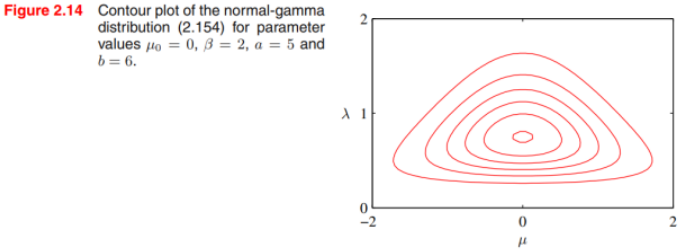


Figure 2.14 Contour plot of the normal-gamma distribution (2.154) for parameter values $\mu_0 = 0$, $\beta = 2$, $a = 5$ and $b = 6$.

Conjuage prior로는 아래의 Wishart distribution을 도입한다. (이때 ν는 자유도, B는 정규화 상수)

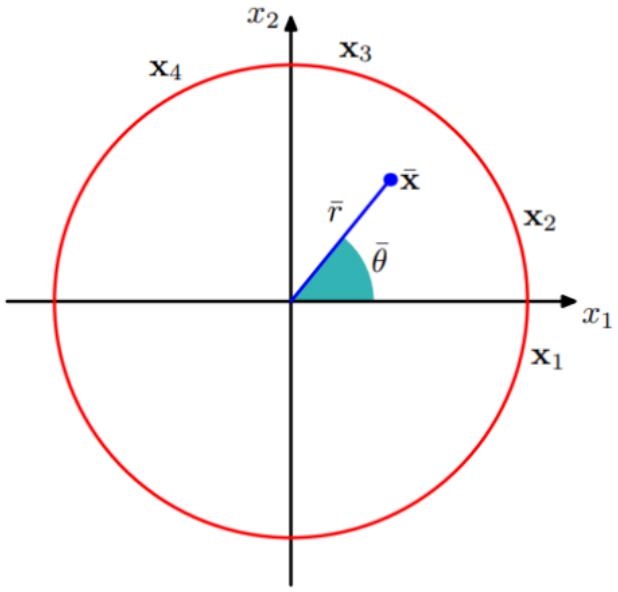
$$\mathcal{W}(\Lambda|W, \nu) = B|\Lambda|^{(\nu-D-1)/2}\exp(-\frac{1}{2}Tr(W^{-1}\Lambda))$$

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

$$p(\mu, \Lambda|\mu_0, \beta, W, \nu) = N(\mu|\mu_0, (\beta\Lambda)^{-1})\mathcal{W}(\Lambda|W, \nu)$$

2.3 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, 이에 대해 해당한다. 가령 특정 위치에서의 바람의 방향이라던가 하루 단위, 또는 연간 단위의 주기를 갖는 모델을 의미한다.



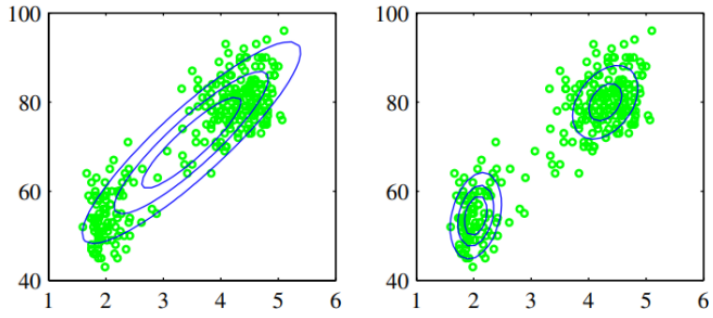
주기성 변수의 관찰 데이터를 $D = \theta_1, \theta_2, \cdots, \theta_n$이라고 하고, 단위 원 내의 한 점으로 나타낸다. 이때 이 점들에 대한 평균 값은 $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$이며, 대입해서 전개해보면 $\bar{x} = (\bar{r}cos\bar{\theta}, \bar{r}sin\bar{\theta} = (\frac{1}{N}\sum_{n=1}^{N} cos\theta_n, \frac{1}{N}\sum_{n=1}^{N} sin\theta_n)$

$$\therefore \bar{\theta} = tan^{-1}(\frac{\sum sin\theta_n}{\sum cos\theta_n})$$

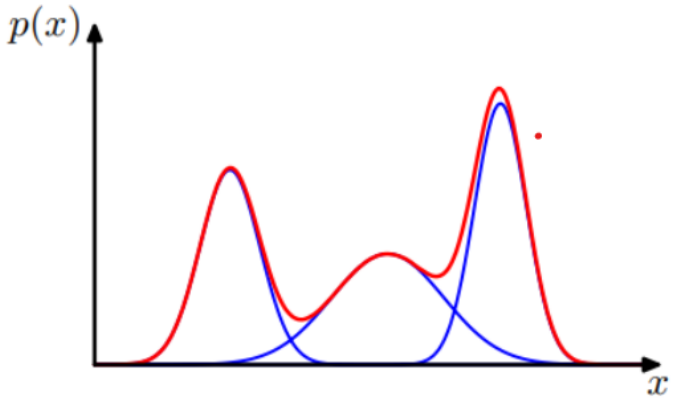
이 결과는 주기성 변수에 대한 MLE 결과와도 일치한다.

2.4 Mixtures of Gaussians

현실적으로 가우시안 분포만을 적용하기 어려운 경우가 다수 존재한다. 다음 그림은 옐로스톤 국립공원에 있는 간헐천의 화산 폭발 데이터로, 크게 두 개의 지배적인 집단을 형성하고 있다.



이에 대한 해결 방안으로 혼합 가우시안 모델(Mixture Gaussian Distribution)이 있다. 혼합 가우시안 모델은 선형 결합을 통해 매우 복잡한 밀도의 표현이 가능하며, 충분한 개수의 결합으로 선형결합의 계수 뿐만 아니라 평균과 공분산을 조절하여 거의 대부분의 연속 밀도를 임의의 정확도로 근사 가능하다.



K개의 중첩 형태는 다음과 같다.

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sum_k)$$

이때, $N(x|x_k, \sum_k)$는 component이며, π_k는 mixing coefficients이다. 그 후, Marginal density를 전개하면,

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

위 식과 비교하면 $p(k) = \pi_k$, $p(x|k) = N(x|x_k, \sum_k)$으로 고려할 수 있고, Posterior probability $p(k|x)$의 값을 유도해낼 수 있다. (이 값은 responsibility라고 불린다.)

3. The Exponential Family

지수족에 속하는 분포는 다음과 같은 형태의 일반화된 형태로 표현 가능하다.

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

정규 분포, 감마 분포, 베타 분포, 베르누이 분포, 포아송 분포 이외에도 다양한 분포들이 지수족에 속한다.

3.1 Maximum Likelihood and Sufficient Statistics

$$-\nabla lng(\eta_{ML}) = \frac{1}{N}\sum_{n=1}^{N} u(x_n)$$

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

3.2 Conjugate Priors

사후 확률 분포가 사전 확률 분포와 같은 분포족(family)에 속하게 되면, 이 때의 사전 확률 분포를 conjugate prior라고 한다.

3.3 Noninformative Prior

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 Priors를 사용한다. Noninformative prior distribution의 가장 쉬운 방법은 prior를 상수로 설정하는 것이다.

$$p(\lambda) = const$$

세션에서 이럴 경우 발생하는 문제에 대해서 살펴보고, 상수 값을 가지는 사전 분포는 알맞은 상황에만 사용해야 한다는 결론을 얻었다. 사전 분포가 상수 값을 가지는 예시에 대해서 살펴보자.

$$p(x|\mu) = f(x - \mu)$$

μ는 location parameter라고 부른다. 위의 밀도 함수에 $\hat{x} = x + c$를 대입할 경우, $\hat{\mu} = \mu + c$가 만족한다.

$$p(\hat{x}|\hat{\mu}) = f(\hat{x} - \hat{\mu})$$

이런 경우, translation invariance가 만족한다고 한다. 이에 대해서, 사전 분포를 적용했을 때에서 translation invariance를 만족해야한다. 즉, $A \leq \mu \leq B$의 범위에서 $A - c \leq \mu \leq B - c$로 바뀌어도 같은 밀도 함수 값을 가져야 한다.

$$\int_A^B p(\mu)d\mu = \in t_{A-c}^{B-c} p(\mu)d\mu = \int_A^B p(\mu - c)d\mu$$

위의 식이 성립하기 위해서는 $p(\mu - c) = p(\mu)$이어야 한다. 즉, $p(\mu) = const$이다.

4. Nonparametric Methods

세션에서는 frequentist 입장에서의 Nonparametric method를 배웠다.

1. 히스토그램
2. Kernel density estimation
3. Neareast neighbour density estimation

4.1. 히스토그램

히스토그램의 확률 값은 아래와 같다.

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

Δ를 어떻게 설정하냐에 따라 히스토그램이 달라진다. 따라서, Δ를 어떻게 설정하는 지가 중요한 문제가 된다.

4.2. Kernel estimation

$p(x)$로부터 추출된 벡터 x가 특정한 영역인 R에 들어갈 확률은 다음과 같다.

$$P = \int_R p(x')dx'$$

R이 너무 작은 영역이어서, $p(x)$가 바뀌지 않는다. (상수이다) 라고 가정한다면 P는 다음과 같이 나타낼 수 있다.

$$P = \int_R p(x')dx' \simeq p(x)V$$

이때, V는 영역 R의 부피이다. 위에서 얻은 P의 식을 합쳐보면 다음과 같은 식을 얻을 수 있다.

$$P = \int_R p(x')dx' \simeq p(x)V = \frac{k}{N}$$

$$p(x) = \frac{k}{NV}$$

주로 N은 고정되어 있기 때문에, 적절한 V를 찾으면 $p(x)$를 잘 추정할 수 있다. 이때, V는 이론적으로 다음과 같은 두 가지 조건을 만족해야 한다.

1. 충분한 샘플이 영역 R에 속할 수 있을만큼 R이 충분히 커야한다.
2. 영역 R에 포함된 $p(x)$가 상수 값이 될 수 있도록 R이 충분히 작아야 한다.

V를 고정시키고 k를 결정하는 방법은 Kernel density estimation이고, k를 고정시키고 V를 결정하는 방법은 Nearest neighbour density estimation이다.

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

위와 같은 커널 함수를 이용하여, 중심에 포인트 x가 존재하는 매우 작은 크기의 hyper cube인 영역 R에 속하는 샘플의 개수를 구할 수 있다.($V = h^d$)

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k(\frac{x - x_n}{h})$$

이때, 가우시안 함수를 사용하여 smoothing 해준 결과는 다음과 같다.

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

이를 통해, 어떠한 분포를 가정하지 않고 확률 밀도 $p(x)$를 표현할 수 있다.

4.3 Nearest neighbour density estimation

Nearest neighbour density estimation은 x를 중심으로 하는 구를 사용하며, 구가 k개의 샘플을 포함할 때까지 반지름을 늘린다. 그 후, 정확히 k개의 샘플이 포함된 구의 부피를 V라고 하고, 이때의 확률 밀도 $p(x)$를 추정한다. Nearest neighbour method를 활용해 분류 문제를 해결하는 과정을 살펴보자.
N는전체 데이터의 크기이고, C_k는 분류되는 클래스의 종류, N_k는 특정 클래스 C_k에 포함되는 데이터의 크기라고 하자. 10개의 공을 갖고 있을 때 빨간공 2개, 파란공 8개를 갖고 있다고 가정해보자. 즉, $N = 10$, $C_1 = Red$, $N_1 = 2$, $C_2 = Blue$, $N_2 = 8$과 같이 표현할 수 있다.
입력 데이터 x를 중심으로 샘플 데이터 중 K개가 속하는 구를 구한다. 이때, 구의 부피는 V이고, 구 안에 존재하는 샘플 데이터의 클래스를 확인하고 개수를 구한다. 클래스 k의 샘플 수를 K_k라고 하면,

$$p(x|C_k) = \frac{K_k}{N_k V}$$

각 클래스에 대한 prior는 다음과 같다.

$$p(C_k) = \frac{N_k}{N}$$

베이즈 정리를 이용해 두 식을 결합하면 다음과 같은 결과를 얻을 수 있다.

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} = \frac{K_k}{K}$$

이를 통해, 각 클래스에 속할 확률은 간단하게 $\frac{K_k}{K}$ 임을 알 수 있다.