PRML과 부수적인 자료들을 공부하여 간단히 정리하였습니다.

이번 장은 주어진 데이터를 이용하여 Distribution을 만드는 것을 배울 것이다. density estimation을 하는 것이다. 이에 대한 방법으로 크게 parametric, nonparmetric 방법으로 나눌 수 있다. 추가로 몇가지 중요한 분포들에 대해 살펴볼 것이다.

2.1 Binary Variables

동전던지기와 같이 random variable이 딱 두가지의 값을 가지는 경우 ($x \in \{0, 1\}$) 에 대해 살펴보자.

- Bernoulli distribution
  - $x = 1$의 확률을 $p(x = 1/\mu) = \mu$ 라고 하자. ($0 \leq \mu \leq 1$)
  - $E[x] = \mu, Var[x] = \mu(1 - \mu)$
  - parameter를 MLE로 추정하면 $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$

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

MLE의 문제점을 여기서 볼 수 있다. 만약에 동전을 3번 던져서 모두 앞면이 나왔다고 하자. 이 data를 기반으로 동전이 앞면이 나올 확률을 MLE로 추정한다면 1일 것이다. 이처럼 극단적으로 overfitting이 되는 경우가 생길 수 있다. 이에 대한 해결책으로는 더 많은 data나 bayesian 접근법이 있을 것이다.

- Binomial distribution
  - N번 중 μ의 확률로 사건이 x개 발생한 경우 (Bernoulli trial이 N번 발생)
  - $E[x] = N\mu, Var[x] = N\mu(1 - \mu)$

$$Bin(x/N, \mu) = \binom{N}{x} \mu^x(1 - \mu)^{N-x}$$

2.1.1 The beta distribution

위의 분포를 보고 bayesian의 접근방식을 생각해보자. parameter μ에 대한 prior를 만들어보자. *conjugacy* (prior와 posterior가 같은 분포를 갖는) 의 성질을 이용하면 Beta dist를 생각할 수 있다. prior도 beta이고 posterior도 beta dist의 모습을 보이도록 만들어준다.

- conjugacy를 이용하면 계산, 해석적인 측면에서 상당히 유용하다.

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

- $E[\mu] = \frac{a}{a+b}, Var[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$

Binomial likelihood function과 Beta prior를 곱하여 posterior dist of μ 를 만들면

$$p(\mu/x, l, a, b) \propto \mu^{x+a-1}(1 - \mu)^{N-x+b-1}$$

합이 1이 되게 constant를 만들지 않아도 posterior가 beta distribution임을 파악할 수 있다.

- posterior에서 a와 b는 각각 x=1, x=0 인 data의 수와 같은 의미(역할)임을 알 수 있다.
- 우리는 prior를 beta로 이용했고 posterior가 beta로 나왔다. 그렇다면 나온 posterior를 다시 prior로 이용할 수 있을 것이다. 이처럼 sequential한 접근이 가능해진다.

우리의 목표는 predict이므로 predictive distribution을 구해보자.

$$p(x = 1/D) = \int_0^1 p(x = 1, \mu/D)d\mu$$

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

여기서 posterior dist의 평균을 구하면

$$p(x=1/D)=\frac{x+a}{x+a+N-x+b}$$

이고 데이터의 수가 많아지면 posterior mean은 MLE와 같아진다. 또한, uncertainty도 줄어들며 likelihood function의 모양과 가까워진다. 물론, 반대로 prior의 정보가 강하다면 prior와 비슷해진다. prior가 강하거나 data수가 많아 likelihood가 강해지면 uncertainty가 줄면서 posterior distribution의 모양이 뾰족해진다.

수리통계학에서 배웠던 공식을 이용하여 살펴보면

- $E_\theta[\theta]=E_D[E_\theta[\theta/D]]$
  - D에 대해 averaged over된 posterior mean of θ = prior mean of θ
- $V_\theta[\theta]=E_D[V_\theta[\theta/D]]+V_D[E_\theta[\theta/D]]$
  - 평균적으로 posterior variance of θ가 prior variance보다 더 작다.

2.2 Multinomial Variables

이번에는 확률변수가 2가지의 값을 갖는게 아닌 K개의 값을 갖는 경우를 살펴보자. 이를 위해 우리는 vector로 확률변수를 표현한다. 예를 들어,

- 주사위를 던졌더니 3이란 수가 나왔다.
  - $\mathbf{x}=(0,0,1,0,0,0)^T$ 이렇게 표현한다.
- 각 원소 x_k들의 합은 1이다.

$x_k=1$인 확률을 parameter μ_k로 표현하면, $\mathbf{x}$의 distribution은

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

- $\sum\mu_k=1, 0\le\mu_k\le1$

이다. expectation은

$$E[\mathbf{x}/\boldsymbol{\mu}]=\sum_{\mathbf{x}}p(\mathbf{x}/\boldsymbol{\mu})\mathbf{x}=(\mu_1,...,\mu_K)^T=\boldsymbol{\mu}$$

으로 구할 수 있다. 그렇다면 이제 likelihood function을 구해보자.

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

- $m_k=\sum_n x_{nk}$: 전체 data에서 k값을 가지는 data 갯수

이 likelihood function을 이용하여 parameter $\boldsymbol{\mu}$를 구해보자.

- constraint $\sum_{k=1}^{K}\mu_k=1$ 에서 log likelihood 를 최대화 해야 한다.
- Lagrange multiplier λ를 이용하여 아래 식을 최대화하면 된다. (Lagrange method)

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

- μ_k에 대해 미분하면 $\mu_k=-m_k/\lambda$ 이고 constraint때문에 $\lambda=-N$ 이라는 것을 파악할 수 있다. 따라서 MLE는

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

- Multinomial distribution
  - $\sum\mu_k=1, 0\le\mu_k\le1$

- $\sum m = N$

$$Multi(m_1, m_2, ..., m_K/\mu, N) = \frac{N!}{m_1!m_2!...m_K!}\prod_{k=1}^{K}\mu_k^{m_k}$$

2.2.1 The Dirichlet distribution

$$Dir(\boldsymbol{\mu}/\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$$

- $\sum \mu_k = 1, 0 \leq \mu_k \leq 1$
- Multinomial의 conjugate prior
- K = 2이면 beta 분포이다. Binomial의 일반화된 분포가 Multinomial이듯 Beta의 일반화된 분포가 Dirichlet 분포라고 할 수 있다.
- posterior : $p(\boldsymbol{\mu}/D, \boldsymbol{\alpha}) \propto p(D/\boldsymbol{\mu})p(\boldsymbol{\mu}/\boldsymbol{\alpha}) \propto \prod_{k=1}^{K}\mu_k^{\alpha_k+m_k-1}$
  - 이전에 봤듯이 prior의 a_k는 data에서 'observation of $x_k = 1$' 의 갯수와 같은 의미(역할)이라고 할 수 있다.