기계학습

13강 모델조합(2), 확률분포(1)

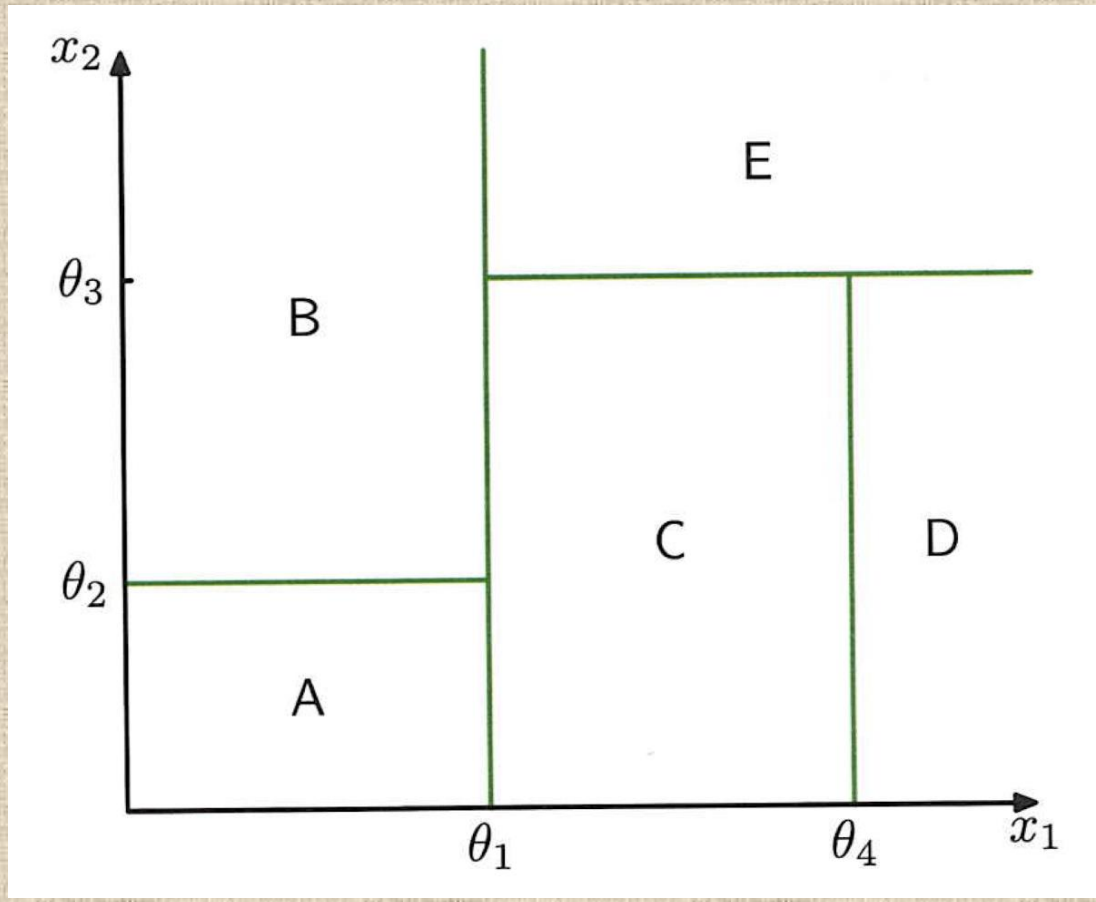
장필훈 교수

학습목차

1. 모델조합(2) : tree, mixture
2. 확률분포(1)

01
모델조합(2)

1-1 트리기반모델



Bishop. Fig.14.5

- 구역 내에는 타겟변수를 예측하기 위한 개별 모델이 존재한다.

1-1 트리기반모델

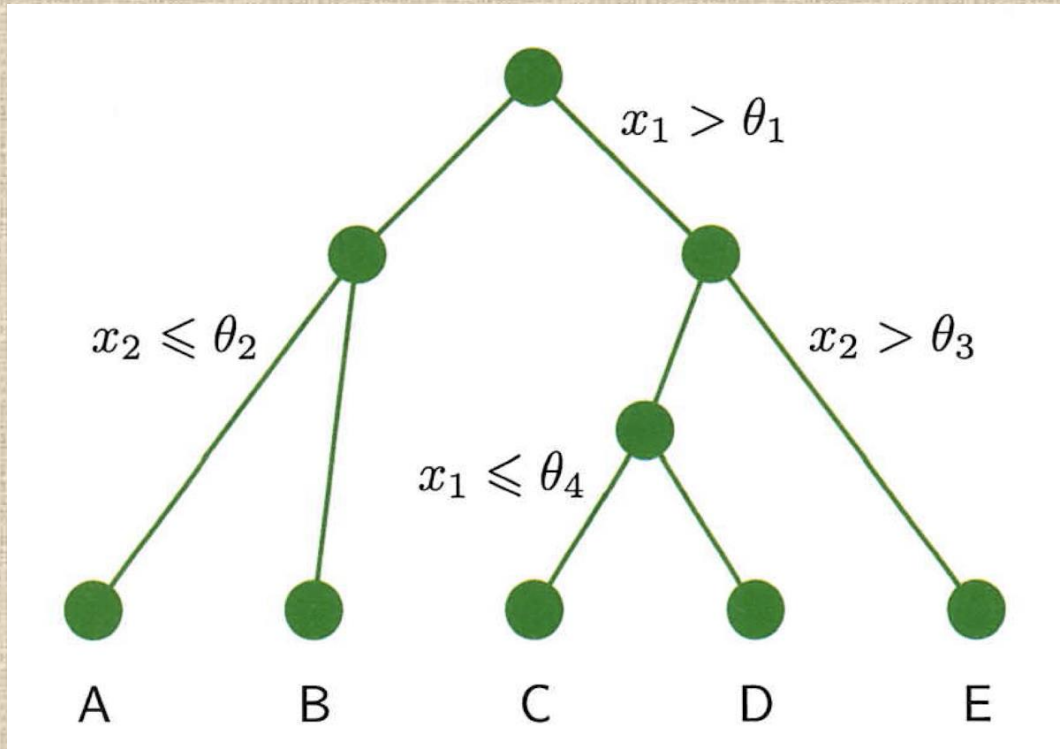
- 구역 내 개별모델의 예시
  - 회귀: 단순상수를 출력으로.
  - 분류: 구역 내 모든 데이터포인트를 특정 클래스로.
- 결과를 해석하기 쉽다는 것이 트리기반모델의 가장 큰 장점
  - 예: 스무고개

1-1 트리기반모델

- 학습할 때 고려할 문제
  - 트리의 구조는?
  - 분할의 기준은?
  - 각 구역 내에서 할당해야 하는 값(혹은 클래스)은?

1-1 트리기반모델

- 회귀 예제



Bishop. Fig.14.6

입력: $\mathbf{x} = (x_1, \ldots, x_D)^T$

출력: 라벨 t_n

제곱합 오류함수

(따라서 구역내 평균이 최적)

1-1 트리기반모델

- 회귀 예제(계속)
  - 어떤 구조를 사용하는 것이 오류함수를 최소화 할 것인가 하는 문제는 보통 계산적으로 실행이 불가능하다.
  - greedy알고리즘 사용해서 해결
    - 루트에서 시작, 분할해 나감
    - 언제 분할을 멈출 것인가?

1-1 트리기반모델

- 분할을 언제 멈출 것인가
  - 당장은 오류를 감소시키는 분할이 없지만, 몇단계 후에 오류를 많이 감소시킬 수 있는 분할이 있음.
  - 따라서, 데이터포인트 개수만큼 잎을 가지는 가지를 만들고, 가지치기 하는 방법을 사용한다.
  - 잔차오류와 복잡도의 trade-off

1-1 트리기반모델

- 가지치기
  - 영역 R_τ에 대한 최적예측은

$$y_\tau = \frac{1}{N_\tau} \sum_{\mathbf{x}_n \in \mathcal{R}_\tau} t_n$$

  - 제곱오류에 대한 기여도는

$$Q_\tau(T) = \sum_{\mathbf{x}_n \in \mathcal{R}_\tau} \{t_n - y_\tau\}^2$$

1-1 트리기반모델

- 가지치기(계속)
  - 가지치기 기준값은

$$C(T) = \sum_{\tau=1}^{|T|} Q_\tau(T) + \lambda|T|$$

  - λ값은 교차검증으로 결정

1-1 트리기반모델

- 분류문제
  - 성능척도가 다르다.(이외에는 동일)
  - 크로스 엔트로피, 지니 인덱스.

$$Q_\tau(T) = -\sum_{k=1}^{K} p_{\tau k} \ln p_{\tau k} \qquad Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k}(1 - p_{\tau k})$$

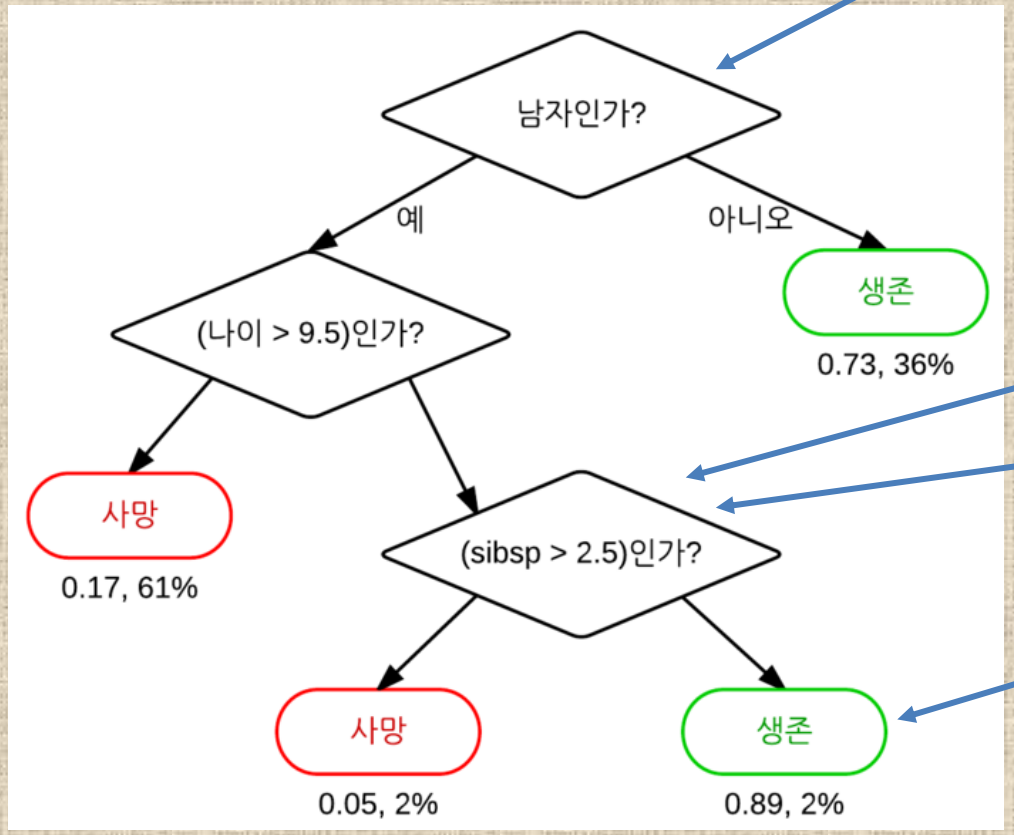
구역 $\mathcal{R}_\tau$에서 클래스 k에 할당된 비율이 $p_{\tau k}$

1-1 트리기반모델

- 단점
  - 분할이 공간축에 평행 (예:대각선이 최적분할이면?)
  - 강한분할. 입력공간의 구역은 단 하나의 결과에 대응
    - 역도 성립
  - 조각별로 상수인 예측값 = 불연속

1-2 decision tree

- 결정트리



https://ko.wikipedia.org/wiki/결정_트리_학습법

root node

꼭 Y/N일 필요 없고
제한이 없다.(숫자, 클래스..)

internal node

leaf node

1-2 decision tree

- 기준은 보통 크로스엔트로피, 지니인덱스로 잡는다.
- $G = 1 - \sum_n p_n^2$ (p: 사건이 발생할 확률,Y/N이라면 n=2)
  - 불순도를 나타내기 때문에 낮을수록 좋음
  - 상위노드의 impurity는 하위 노드의 weighted sum

1-3 선형회귀의 혼합

- 각 모델의 결과에 확률적 해석을 부여
- K개의 선형회귀모델을 고려했을 때, 혼합분포는

$$p(t|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(t|\mathbf{w}_k^T \phi, \beta^{-1})$$

관측데이터집합 $\{\phi_n, t_n\}$이 주어지면 로그 가능도 함수는,

$$\ln p(\mathbf{t}|\theta) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(t_n|\mathbf{w}_k^T \phi_n, \beta^{-1}\right) \right)$$

1-3 선형회귀의 혼합

- EM으로 가능도 함수 최대화 한다.

n:data index,
k:class index

  - 어떤 성분이 데이터포인트를 생성했는지 $\mathbf{z}_{nk} = \{0,1\}$

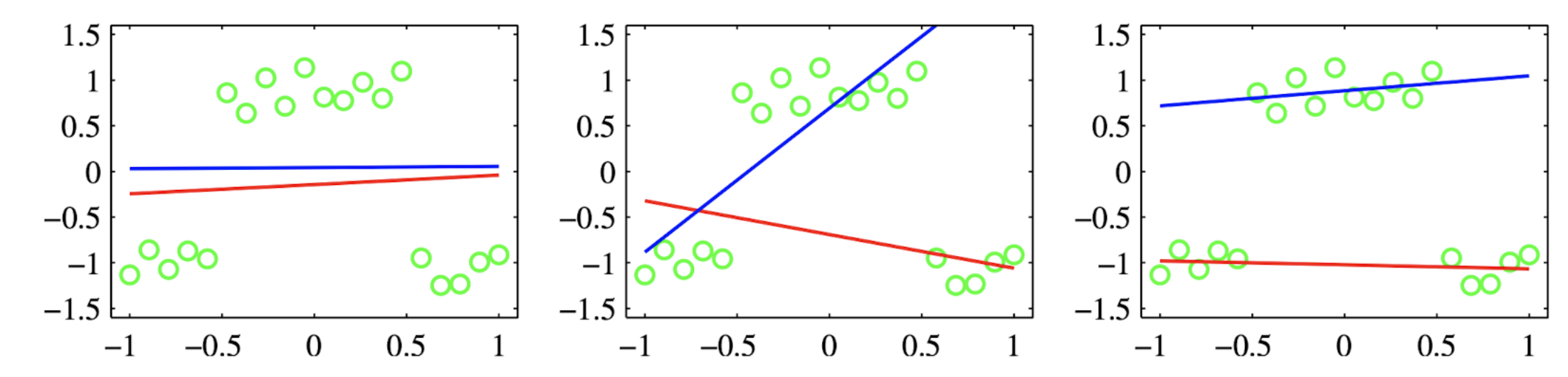
$$\ln p(\mathbf{t}, \mathbf{Z}|\theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln\{\pi_k \mathcal{N}(t_n|\mathbf{w}_k^T \phi_n, \beta^{-1})\}$$

  - E단계는 아래식, M은 위 식 최대화.

$$\mathbb{E}[z_{nk}] = p(k|\phi_n, \theta^{\text{old}}) = \frac{\pi_k \mathcal{N}(t_n|\mathbf{w}_k^T \phi_n, \beta^{-1})}{\sum_j \pi_j \mathcal{N}(t_n|\mathbf{w}_j^T \phi_n, \beta^{-1})}$$

1-3 선형회귀의 혼합

- EM(계속)
  - 제약조건: $\sum_k \pi_k = 1$
  - $\mathbf{w}, \beta$에 대해서도 같은 과정을 거침



Bishop. Fig.14.8

1-4 로지스틱 및 기타의 혼합

- 로지스틱K개의 확률적 혼합. 타겟변수에 대한 조건부분포=

$$p(t|\phi,\theta) = \sum_{k=1}^{K} \pi_k y_k^t [1-y_k]^{1-t}$$

ref. Bernoulli Regression

- 조건부 분포만 알면, EM으로 시도해 볼 수 있다.
- 기타모델을 혼합하는 것도 가능하다. 방법은 매우 다양.

02
확률분포(1)

2-1 확률분포

- 관찰집합이 주어졌을 때 그 분포를 모델링 하는것이 목표
  - $\mathbf{x}_1, \ldots, \mathbf{x}_n$이 주어졌을 때 확률변수 $\mathbf{x}$의 확률분포 $p(\mathbf{x})$
  - =밀도추정문제
  - 가능성은 무한하기 때문에, 정확한 답을 찾을수는 없다
    - 결국 모델선택의 문제

      (여러 모델을 알고 있으면 그중에 선택하면 된다.)

2-2 베르누이분포

- 이진확률변수 $x \in \{0,1\}$
- $x = 1$일 확률 $p(x = 1|\mu) = \mu, 0 \leq \mu \leq 1$
- $p(x = 0|\mu) = 1 - \mu$
- 다음과 같이 적는다.

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

- ‘베르누이분포’

2-2 베르누이분포

- 평균

$$\mathbb{E}[x] = p_{x=1} \cdot 1 + p_{x=0} \cdot 0 = \mu \cdot 1 + (1-\mu) \cdot 0 = \mu$$

- 분산

$$\mathbb{E}[x^2] = p_{x=1} \cdot 1^2 + p_{x=0} \cdot 0^2 = \mu = \mathbb{E}[x]$$

$$\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \mu - \mu^2 = \mu(1-\mu)$$

2-2 베르누이분포

- 확률변수 x의 관측값 데이터 집합 $\mathcal{D} = \{x_1, \ldots, x_N\}$이 주어지면 i.i.d를 가정했을 때 가능도 함수는,

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

- 따라서 로그가능도 함수는,

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

2-2 베르누이분포

- 가능도 함수를 μ에 대해 미분. 최대가능도추정값=

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- '표본평균'
- $x = 1$인 표본 개수를 m이라고 하면, $\mu_{ML} = \frac{m}{N}$

2-2 베르누이분포

- N이 작을 때 왜곡이 쉽다
  - 동전 세번 던졌을 때 모두 앞면이면,
    앞면일 확률 $\mu_{ML} = 1$
  - 최대가능도만 따르면, '앞으로 계속 앞면이다'
  - 베이지안적인 관점이 필요함
    - μ에 대해서도 사전분포 가정하는 방법 : 베타분포

2-3 이항분포

- $x = 1$인 개수 $m(\leq N)$에 대한 분포

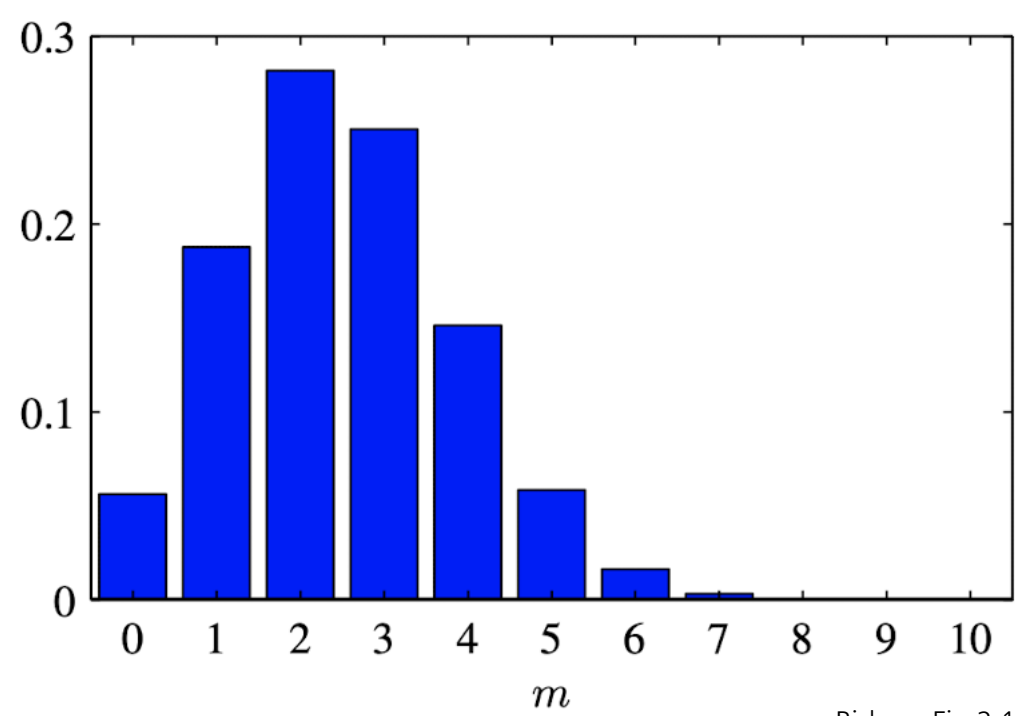
$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

Binomial distribution

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!\,m!}$$

2-3 이항분포

- $N = 10, \mu = 0.25$



Bishop. Fig.2.1

2-3 이항분포

- $\mathbb{E}[m] = \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu)$

$$\mathbb{E}[m] = \mathbb{E}[x_1 + x_2 + \cdots + x_N] = \mu + \mu + \cdots + \mu = N\mu$$

- $\mathrm{Var}[m] = \mathbb{E}[m^2] - \mathbb{E}[m]^2$

sum of variances of independent trials:

$$\mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \mathrm{Var}(X_j)$$

2-3 이항분포

- 따라서,

$$\mathrm{Var}[m] = \mathrm{Var}[x_1 + x_2 + \cdots + x_N] = \mathrm{Var}\left(\sum_i^N x_i\right) = \sum_i^N \mathrm{Var}(x_i)$$

$\mathrm{Var}(x_i) = \mu(1-\mu)$이므로,

$$\mathrm{Var}[m] = \sum_i^N \mu(1-\mu) \ = N\mu(1-\mu)$$

2-3 이항분포

- sum of variances of independent trials

$$\mathrm{Var}(X+Y) = Cov(X+Y, X+Y)$$

$$= E\big((X+Y)^2\big) - E(X+Y)E(X+Y)$$

$$= E(X^2 + 2XY + Y^2) - \big(E(X) + E(Y)\big)^2$$

$$= E(X^2) - E^2(X) + E(Y^2) - E^2(Y)$$

$$= \mathrm{Var}(X) + \mathrm{Var}(Y)$$

2-3 베타분포

- 매개변수 μ에 대한 사전분포 $p(\mu)$

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

베타함수 : 정규화역할

$$\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}du$$

$$= \lim_{m\to\infty} \frac{m^x m!}{x(x+1)(x+2)\cdots(x+m)}$$

Euler form

- a, b는 초매개변수(hyperparameter)

2-3 베타분포

- 베타함수

$$\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

- 기댓값

$$E(x) = \int_0^1 x \cdot \frac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a,b)}dx = \frac{1}{\mathrm{B}(a,b)}\int_0^1 x^{a}(1-x)^{b-1}dx$$

$$= \frac{\mathrm{B}(a+1,b)}{\mathrm{B}(a,b)} = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

2-3 베타분포

- 기댓값(계속)

integration by parts

$$\int_0^\infty u^x e^{-u} du = [-u^x e^{-u}]_0^\infty + x\int_0^\infty u^{x-1} e^{-u} du = x\Gamma(x)$$

$\Gamma(x+1) = x\Gamma(x)$이므로,

$$\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a}{a+b}$$

2-3 베타분포

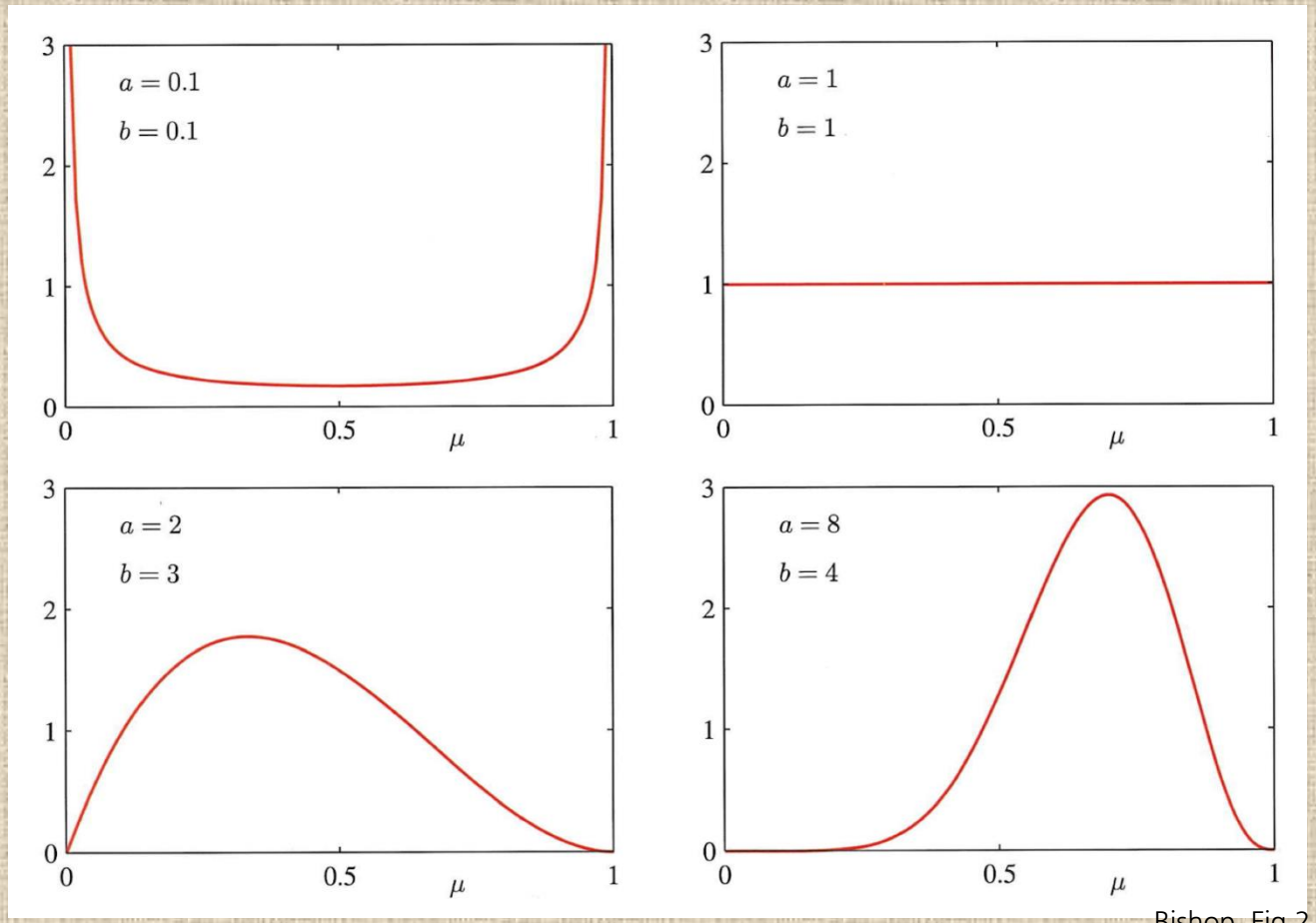
- 분산

$$\mathrm{Var}(x) = \frac{1}{\mathrm{B}(a,b)} \int_0^1 x^{a+1}(1-x)^{b-1} dx - \frac{a^2}{(a+b)^2}$$

$$= \frac{\mathrm{B}(a+2,b)}{\mathrm{B}(a,b)} - \frac{a^2}{(a+b)^2} = \frac{ab}{(a+b)^2(a+b+1)}$$

앞(33p)에서와 같이 Γ로 변환하고 비슷하게 전개

2-3 베타분포



Bishop. Fig.2.2

https://en.wikipedia.org/wiki/Beta_distribution

$a + b$ 커질수록 날카로워진다.

분산식에서 확인 가능

2-3 베타분포

- 이제 이항분포의 가능도 함수에 베타함수(사전분포)를 곱해서 사후분포를 볼 수 있다.

$$P(m|\mu) = \binom{M}{m} \mu^m (1-\mu)^{M-m}$$

$$P(\mu) = \frac{\mu^{a-1}(1-\mu)^{b-1}}{\mathrm{B}(a,b)}$$

$$P(\mu|m) \propto P(m|\mu)P(\mu) = \binom{M}{m} \frac{\mu^{m+a-1}(1-\mu)^{M-m+b-1}}{\mathrm{B}(a,b)}$$

2-3 베타분포

conjugate prior of binomial distribution

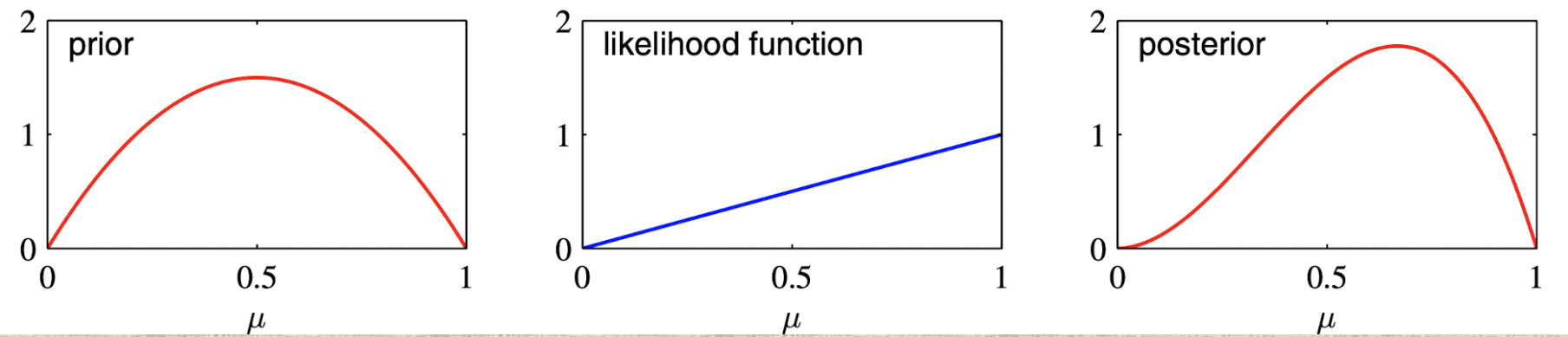
- μ에 관해 함수적 종속성이 있다.(사전,사후분포간)
- 새로 구한 사후분포도 또다른 베타함수이고, p32의 베타분포 식과 비교하면 정규화계수를 쉽게 구할 수 있다.

  결과적으로, 다음과 같은 베타분포가 된다.

$$\frac{\Gamma(M+a+b)}{\Gamma(m+a)\Gamma(M-m+b)}\mu^{m+a-1}(1-\mu)^{M-m+b-1}$$

2-3 베타분포

- 새로 구한 식을 사전분포와 비교해보면,
  - a가 m만큼, b가 $M-m$만큼 증가한 것으로 볼 수 있다.
- 추가관측이 이루어지면 사후분포가 다시 사전분포가 된다.
  - 관측이 이루어질때마다 그에 따라 a, b를 업데이트.



베이지안식 해결과정 예시

Bishop. Fig.2.3

다음시간

14강

- 확률분포(2)