# CS572 Lecture Note

# 9/12

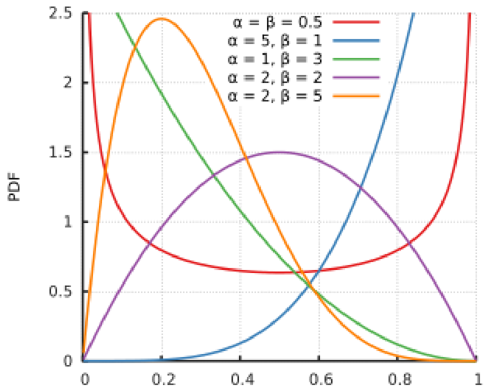
Group 1

Aydar
Seungwoo Yoon
Chansu Park
Shinhyung Kim

Beta Distribution

$$\theta \sim Beta(\alpha_h, \alpha_t)$$

$$P(\theta) = \frac{\Gamma(\alpha_h+\alpha_t)}{\Gamma(\alpha_h)\Gamma(\alpha_t)}\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1} \quad (\ \theta \in [0,1]\ )$$



$$P\left(X[1] = X^h\right) = \int_\theta P\left(X[1] = X^h \middle| \theta\right) P(\theta) d\theta = \frac{\alpha_h}{\alpha_h+\alpha_t}$$

$$P\left(X[2] = X^h | X[1] = X^h\right) = \frac{\alpha_h + 1}{\alpha_h + 1 + \alpha_t}$$

$$P\left(X[M+1] = X^h \middle| D\right) = \frac{\alpha_{h+M_h}}{\alpha_h+\alpha_t+M_h+M_t}$$

$$= \frac{M'}{M+M'}\theta_h' + \frac{M}{M+M'}\frac{M_h}{M} \qquad (M = M_h + M_t,\ M' = \alpha_h + \alpha_t)$$

$$\theta_h' = \frac{\alpha_h}{\alpha_h + \alpha_t}$$

If $M = 0, P = \frac{\alpha_h}{\alpha_h+\alpha_t}$, if $M$ increases, $P \cong \frac{M_h}{M}(MLE)$
-> $\alpha_h, \alpha_t$ indicates our initial belief of the probabilities. So in the first trial, Probability is same with our belief. But as the trial increases, it moves to the actual observation, MLE.

In general, with data $M_h, M_t$, prior is $\beta(\alpha_h, \alpha_t)$,Posterior is $\beta(\alpha_h + M_h, \alpha_t + M_t)$.
-> Conjugate family for binomial (refer the appendix)

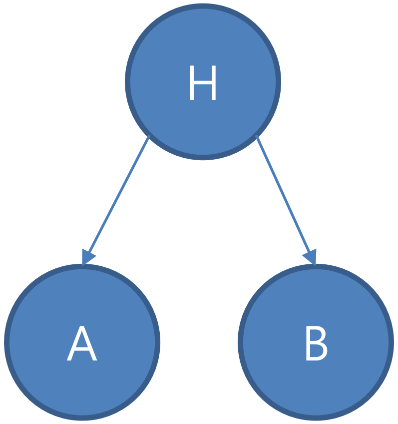
$$argmaxP(D|\theta) = \sum P(D, H = h|\theta)$$

Maximize expected log-likelihood relative to probability distribution over hidden variables.

$$argmaxE_{P(H|D,\theta)}l(\theta; \mathrm{D}, \mathrm{H})$$

Maximize above equation by Expectation Maximization(EM).

Example)




$$D:\ A = 1, B = 1\ 6\ times$$
$$A = 1, B = 0\ 4\ times$$

$$P(A = 1|H = 0) = \frac{M(A=1,H=0)}{M(H=0)}$$

Expectation Maximization

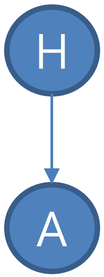
- Pick initial $\theta_0$
- Loop until $\theta$ doesn't change much(converge)

$$E_{i+1} = \sum_{h} P(H = h|D, \theta_i)\,\mathrm{l}(\theta; D, H)$$

$$\theta_{i+1} = argmax_{\theta} E_{i+1}(\theta)$$

Maximizing expected log-likelihood is equivalent to maximizing MLE in a complete dataset, where counts are replaced by expected counts.

Example)




$\theta$: $P(H=1)$
$P(A=1|H=0)$
$P(A=1|H=1)$

$D$: $A=0; M(A=0)$
$A=1; M(A=1)$

$\theta_0$; $P(H=1)=0.4$

| $H$ | $P(A=1\|H)$ |
|---|---|
| 0 | 0.25 |
| 1 | 0.54 |

$$P(H|A,\theta_0)=\frac{P(A|H)P(H)}{P(A)}=\frac{P(A|H)P(H)}{\sum_H P(A|H)P(H)}$$

$$P(H=1|A=0)=\frac{0.46*0.4}{0.46*0.4+0.75*0.6}=a$$

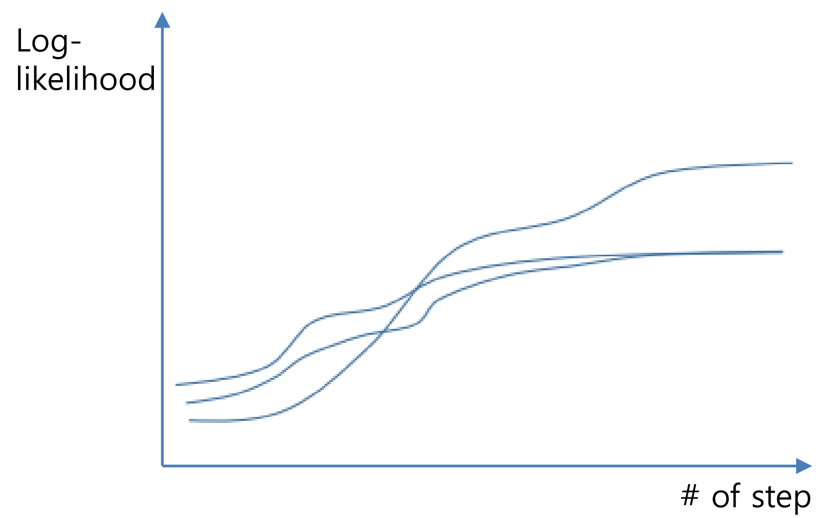
$$P(H=0|A=0)=1-a$$

| $A$ | $H$ | $Expected\ count$ |
|---|---|---|
| 0 | 0 | $M(A=0)(1-a) \quad = v$ |
| 0 | 1 | $M(A=0)a \quad = w$ |
| 1 | 0 | $M(A=1)(1-b) \quad = x$ |
| 1 | 1 | $M(A=1)b \quad = y$ |

$$\theta_1: \quad P(H=1) = \frac{w+y}{v+w+x+y}$$

$$P(A=1|H=0) = \frac{x}{v+x}$$

(EM requires initial guess)




General phenomena of EM
- There are some plateaus
- Monotonic improvement is guaranteed

KAIST

$D = \{y_1, y_2, \dots, y_m\}$

Want to divide D into 2 groups

$group1 \sim N(\mu_1, \sigma_1^2)$
$group2 \sim N(\mu_2, \sigma_2^2)$

$P(group1\ is\ taken) = \pi$

Problem: We have to fine parameters that we don’t know

# Additional Materials

## Beta distribution

**Beta distribution** is a family of continuous probability distribution defined on the interval [0,1] parametrized by two positive shape parameters, denoted by $\alpha$ and $\beta$, that appear as exponents of the random variable and control the shape of the distribution.

Beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success.

Beta distribution is the conjugate prior probability distribution for Bernoulli, binomial, negative binomial and geometric distributions.

## Conjugate Prior

In Bayesian probability theory, if the posterior distributions $P(\theta|x)$ are the same probability distribution family as the prior probability distribution $P(\theta)$, the prior and posterior are then called **conjugate distributions**, and the prior is called a **conjugate prior**.

Conjugate priors may give intuition, by more transparently showing how a likelihood function updates a prior distribution.

## Expectation Maximization

**Expectation-Maximization algorithm(EM)** is an iterative method to find maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables.

The EM iteration alternates between performing an expectation step, and maximization step.

**Expectation step** creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters.
**Maximization step** computes parameters maximizing the expected log-likelihood found on the Expectation step.

# Expectation Maximization

Given the statistical which generates a set $X$ of observed data, a set of unobserved latent data $Z$, and a vector of unknown parameters $\theta$, along with a likelihood function $L(\theta; X, Z) = P(X, Z|\theta)$, the MLE of the unknown parameters is determined by maximizing the marginal likelihood of the observed data,

$$L(\theta; X) = P(X|\theta) = \int P(X, Z|\theta)dZ$$

The EM seeks to find the MLE of the marginal likelihood by iteratively applying these two steps:

Expectation step: Define $Q\left(\theta\middle|\theta^{(t)}\right)$ as the expected value of the log likelihood function of $\theta$, w.r.t. the current conditional distribution of $Z$ given $X$ and the current estimates of the parameters $\theta^{(t)}$:

$$Q\left(\theta\middle|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[logL(\theta; X, Z)]$$

Maximization step: Find the parameters that maximize this quantity:

$$\theta^{(t+1)} = argmax_{\theta}Q(\theta|\theta^{(t)})$$

# Appendix

Proof for posterior:

$$P(\theta|D) \propto P(D|\theta)p(\theta)$$

$$= \binom{M_h + M_t}{M_h} \theta^{M_h} (1-\theta)^{M_t} \frac{\Gamma(\alpha_h + \alpha_t)}{\Gamma(\alpha_h)\Gamma(\alpha_t)} \theta^{\alpha_h - 1} (1-\theta)^{\alpha_t - 1}$$

$$\propto \theta^{M_h + \alpha_h - 1} (1-\theta)^{M_t + \alpha_t - 1}$$

$$\sim \beta(M_h + \alpha_h, M_t + \alpha_t) \blacksquare.$$

# References

https://en.wikipedia.org/wiki/Beta_distribution
https://en.wikipedia.org/wiki/Conjugate_prior
https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm