Probability

https://youtu.be/AXB6r-hjsig

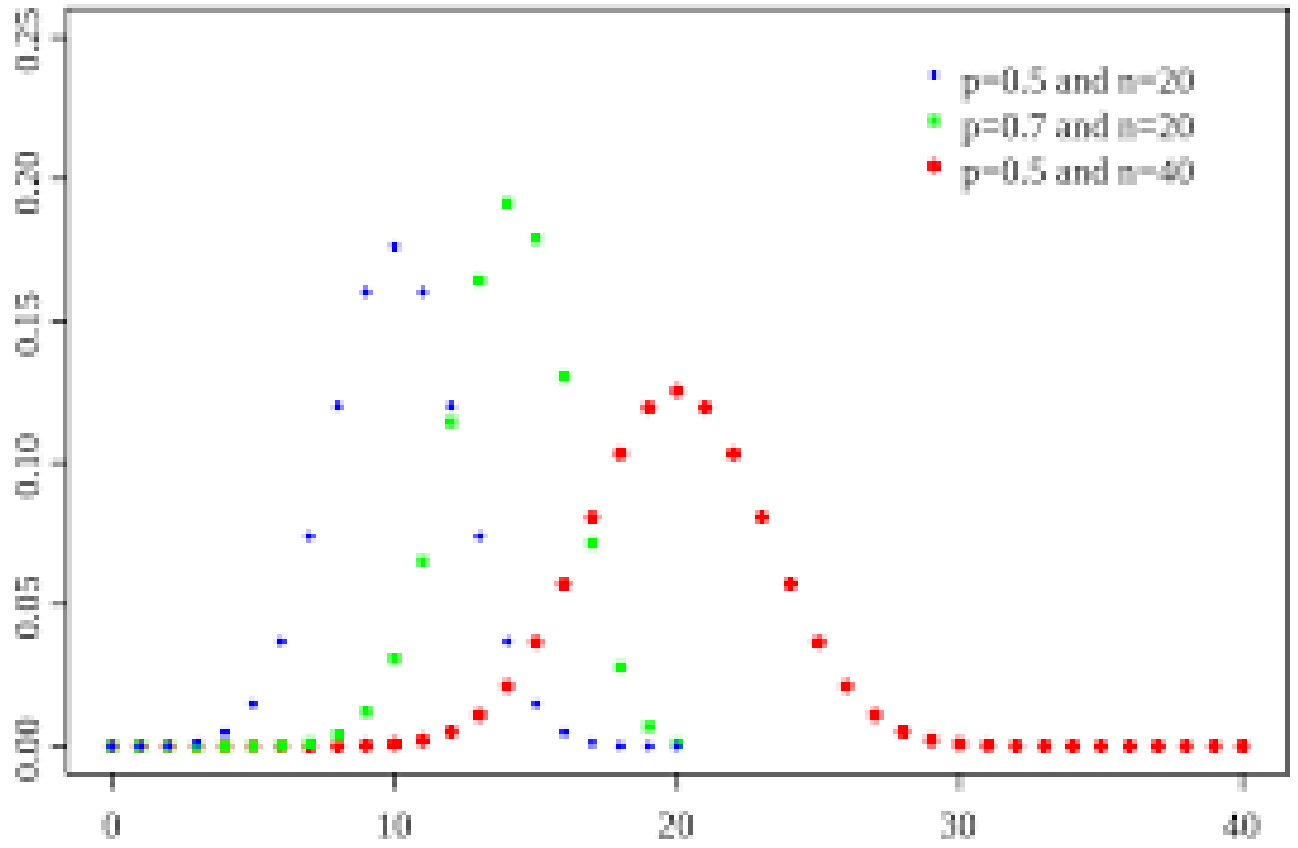
50 : 50

? : ?

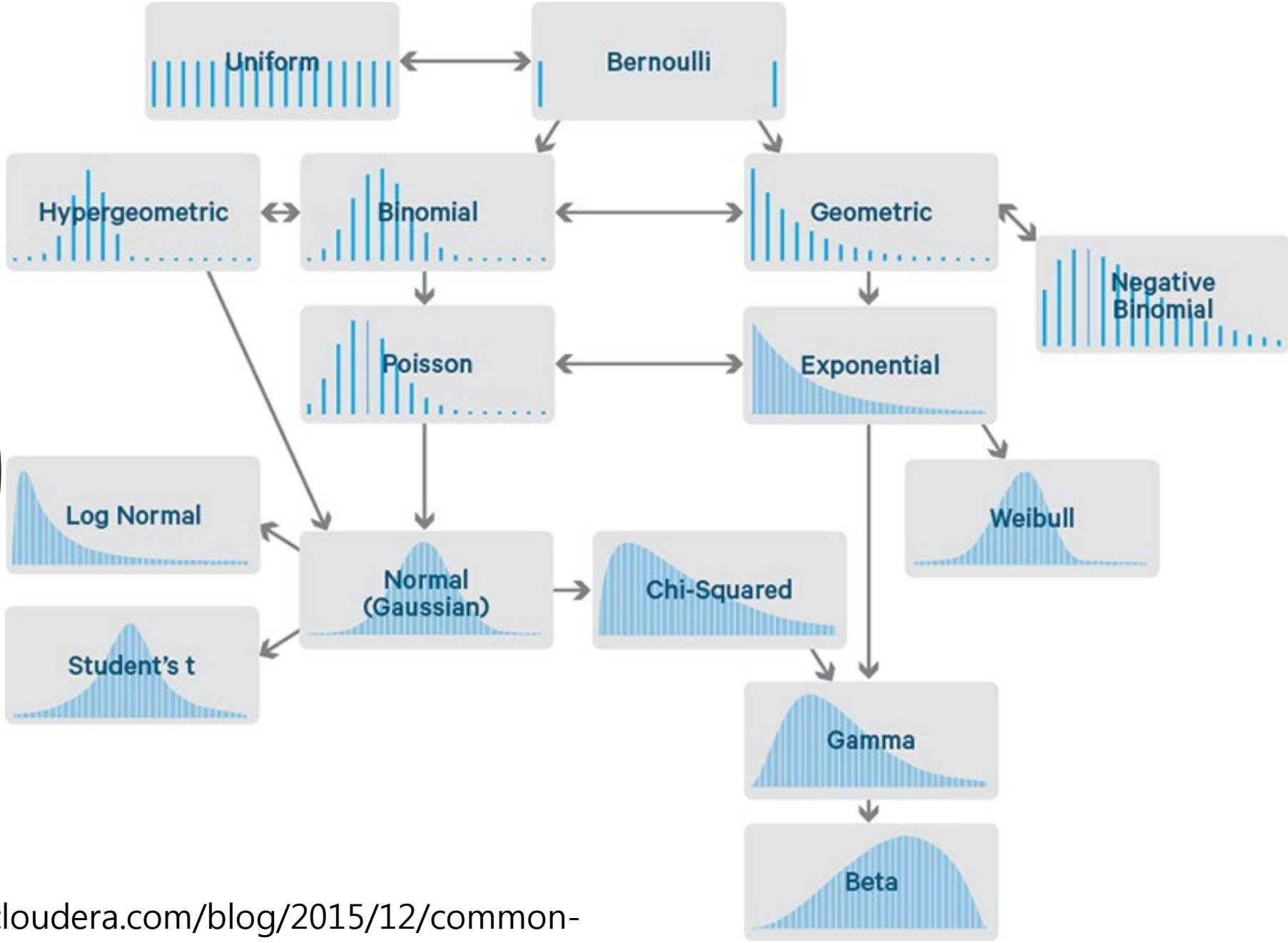
머리가 나올 확률 $\mathrm{p}(H) = ?$

Binomial Distribution

- 불연속된 값들의 분포
- Yes / No Question
- Bernoulli trial
- 상호 독립적인 사건 (i.i.d)
- Notation : B(n, p)
- Mean : np
- Variance : np(1 – p)



Distribution
확률분포



https://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

머리가 나올 확률 $\mathrm{p}(H) = \frac{2}{5}$

$p(H) = \theta$

$p(H) = 1 - \theta$

p(H) = θ p(H) = 1 - θ

$$p(HTHTT) = \theta\,(1-\theta)\,\theta\,(1-\theta)\,(1-\theta)$$

Data = 

$p(H) = \theta$

$p(Data \mid \theta) = \theta^{a_H}(1-\theta)^{a_T}$

가정 : θ 가 실제 압정을 던졌을때 머리가 나오는 확률이다.

1. 목소리를 크게 낸다.
2. 아득바득 우겨본다.
3. 압정을 계속 던져본다
4. 최적화 되어있는 θ 를 찾아본다

MLE(Maximum Likelihood Estimation)

관측된 데이터가 최대화 되는 θ 를 찾는 방법

$$\hat{\theta} = argmax_\theta P(D|\theta) = argmax_\theta \theta^{a_H}(1-\theta)^{a_T}$$

최대,최소 문제 해결에는 미분 → 곱 연산으로 되어있는 부분 → 로그 함수(단조 증가)

$$\hat{\theta} = argmax_\theta lnP(D|\theta) = argmax_\theta \ln\{\theta^{a_H}(1-\theta)^{a_T}\} = argmax_\theta\{a_H ln\theta + a_T \ln(1-\theta)\}$$

$$\frac{d}{d\theta}\{a_H ln\theta + a_T \ln(1-\theta)\} = 0$$

$$\frac{a_H}{\theta} - \frac{a_T}{1-\theta} = 0$$

$$\frac{a_H}{\theta} = \frac{a_T}{1-\theta}$$

$\theta = \frac{a_H}{a_H+a_T}$ $\frac{\text{압정 머리가 나오는 수}}{\text{전체 경우의 수}}$ → MLE 관점에서 바라본 최적화된 $\hat{\theta} = \frac{a_H}{a_H+a_T}$

$$p(H) = \frac{1}{2}$$

$$N = a_H + a_T\,, \qquad \hat{\theta} = \frac{a_H}{a_H + a_T}\ (MLE), \qquad \theta^*\ (true\ parameter)$$

Error Bound Function

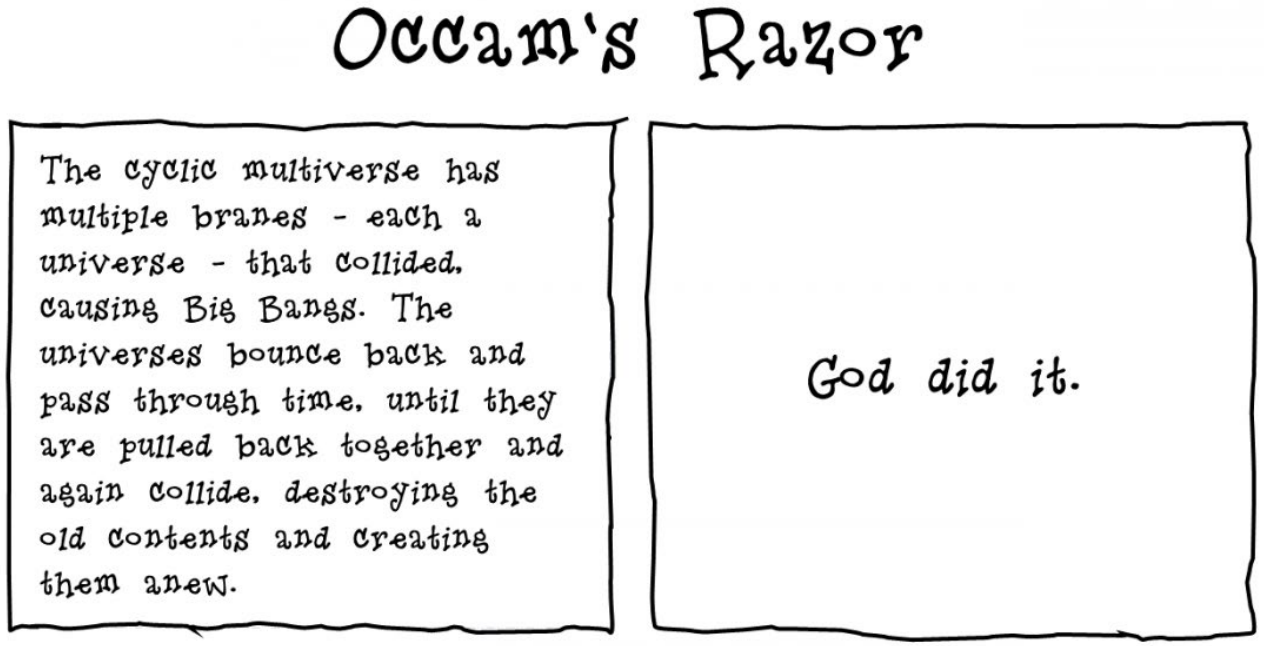
- Hoeffding's inequality

$$\boldsymbol{p\left(\left|\hat{\theta} - \theta^*\right| \geq \varepsilon\right) \leq 2e^{-2N\varepsilon^2}}$$

Probably Approximate Correct (PAC) Learning

- Probably (5% case)

- Approximately ($\varepsilon = 0.1$)



단순성의 원리

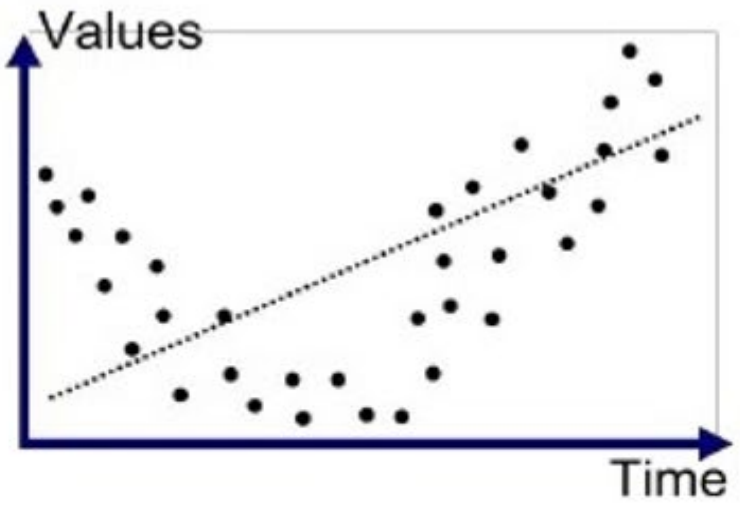
- 복잡하면 이해하기 어렵다

딥러닝에서 표현하는 특징이 너무 많아지면 Overfitting에 빠지기 쉽다.

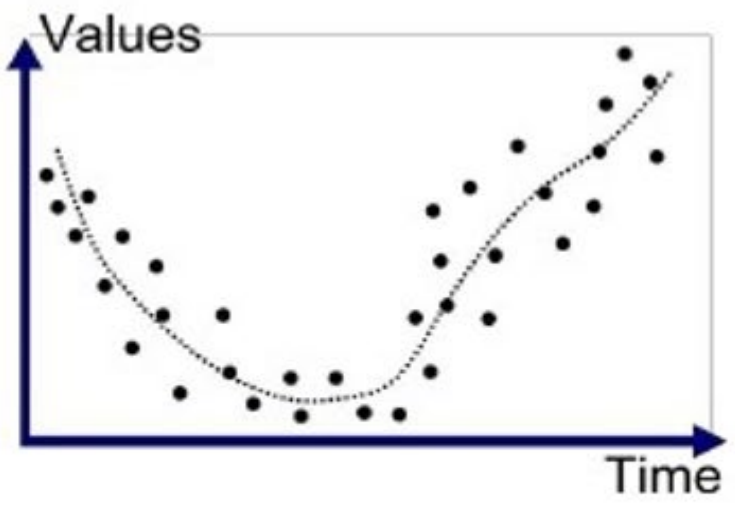
Overfitting의 해결책

- Feature 수를 줄이는 방법

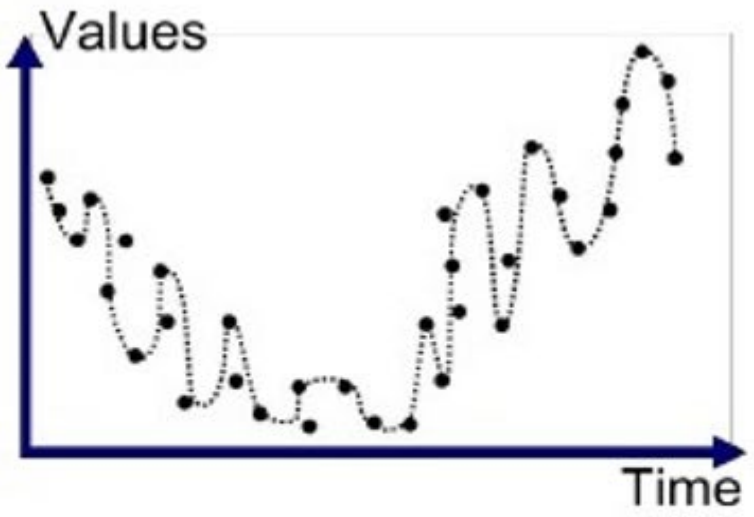
$$p\left(\left|\hat{\theta} - \theta^*\right| \geq \varepsilon\right) \leq 2e^{-2N\varepsilon^2}$$



Underfitted



Good Fit/Robust



Overfitted

MAP(Maximum a Posteriori Estimation)

$$Head\ or\ Tail = 50 : 50$$

$$p(\theta \mid D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{Likelihood \ * Prior\ Knowledge}{Normalizing\ Constant}$$

$$p(\theta \mid D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{Likelihood \ * Prior\ Knowledge}{Normalizing\ Constant}$$

$$p(\theta \mid D) = \theta^{a_H}(1-\theta)^{a_T}$$

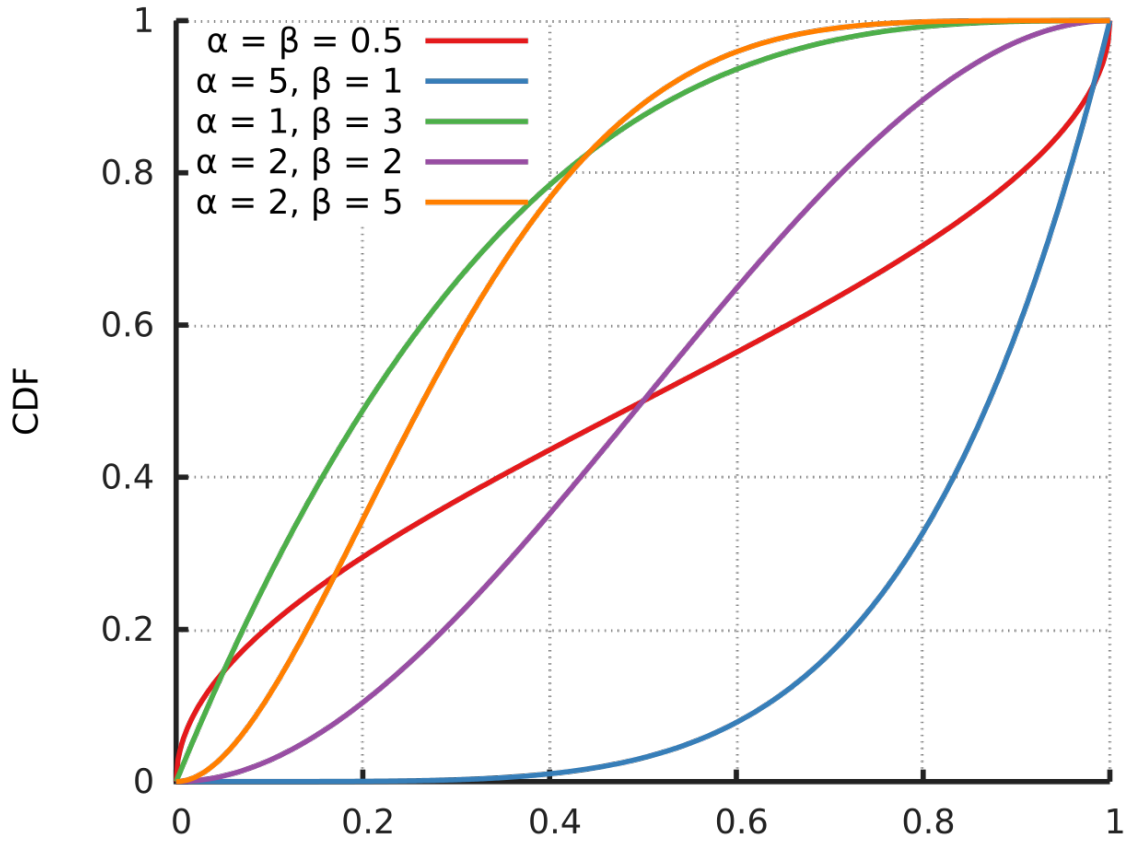
$p(\theta)$ *is the part of the prior k*

$$p(\theta \mid D) \propto P(D|\theta)P(\theta) \propto \theta^{a_H}(1-\theta)^{a_T}P(\theta)$$

$$P(\theta) = ????$$

Beta Distribution

- 두개의 매개변수로 표현
- [0,1] 구간에서 정의되는 연속 확률 분포
- Notation : $Beta(\alpha, \beta)$
- Mean : $\frac{\alpha}{\alpha+\beta}$
- Variance : $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



$P(\theta)$ 를 beta distribution으로 표현하면

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad \Gamma(\alpha) = (\alpha-1)!$$

$$p(\theta \mid D) \propto P(D|\theta)P(\theta) \propto \theta^{a_H}(1-\theta)^{a_T}P(\theta)$$

$$\propto \theta^{a_H}(1-\theta)^{a_T}\ \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$$

MLE(관점)

- θ로 부터 $\hat{\theta} = argmax_\theta P(D|\theta)$

$$P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$$

$$\hat{\theta} = \frac{a_H}{a_H + a_T}$$

MAP(관점)

- θ로 부터 $\hat{\theta} = argmax_\theta P(\theta|D)$

$$p(\theta \mid D) \propto \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$$

$$\hat{\theta} = \frac{a_H + \alpha - 1}{a_H + \alpha - 1 + a_T + \beta - 1}$$

빈도론자(Frequentist)

- θ 는 알려지지 않은 고정된 파라미터
- Maximum Likelihood 가 대표적인 추정치(estimator)를 최대로 만드는(argmax) θ 를 찾아야한다.
- ML분야에서는 주로 log-likelihood를 사용한다.
- 통계적으로 모델의 정확도를 평가하기 위한 방법으로 Bootstrap 기법을 사용한다

베이지언(Bayesian)

- 파라미터 θ를 랜덤 변수로 간주하여 확률 분포에 사용한다.
- 여기에서 가지고 있는 정보 D는 고정된다
- MLE에서 압정을 3번 던져서 모두 머리가 나온경우 θ의 값이 1로 고정되지만, 베이지언 방식에서는 사전 확률로 인해 이 값이 보정된다.

조건부 확률

- 사건 B가 발생했을때 사건 A가 발생할 확률

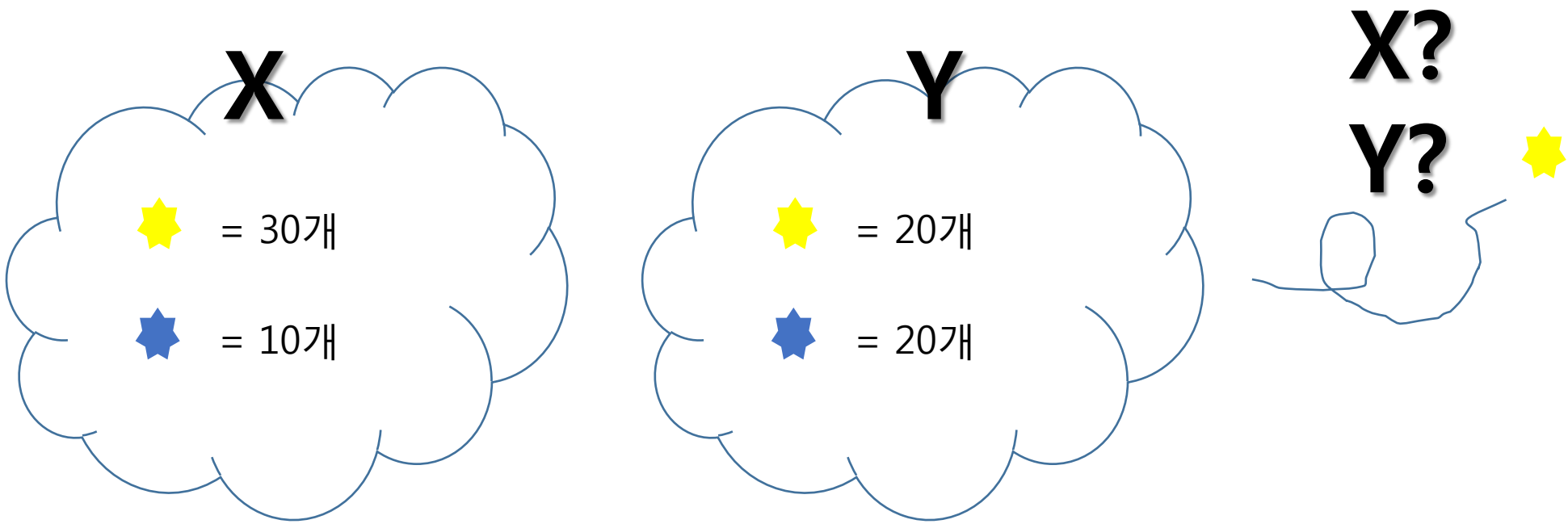
$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

베이즈 법칙

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

독립사건

$$A \; and \; B : P(A \cap B) = P(A)P(B)$$

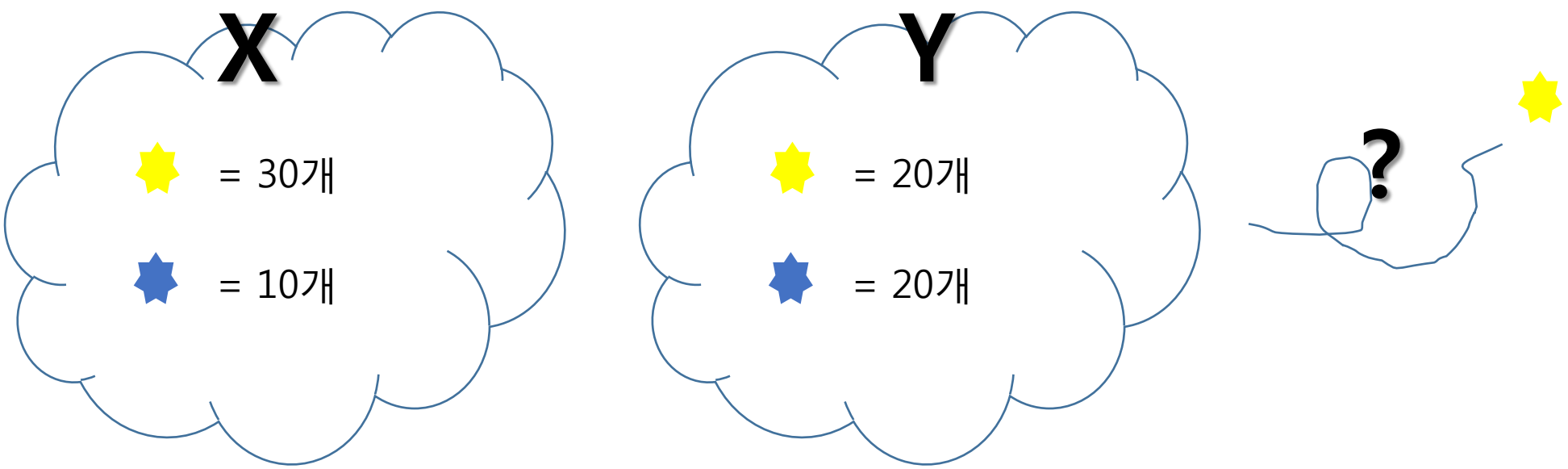

X
= 30개
= 10개
Y
= 20개
= 20개
X?
Y?

Bayesian Probability

$$p(Y \mid X) = \frac{p(X,Y)}{p(X)} \Rightarrow p(X,Y) = p(Y \mid X) \times p(X)$$

$$p(X \mid Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(Y|X) \times p(X)}{p(Y)}$$

$$\therefore p(X|Y) = \frac{p(Y|X) \times p(X)}{p(Y)}$$



$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{Likelyhood * Prior\ Knowledge}{Constant} = Posterior = \frac{\frac{3}{4} * \frac{1}{2}}{\frac{5}{8}} = \frac{\frac{3}{8}}{\frac{5}{8}} = \frac{3}{5} = 0.6$$

P(B) = 어떤 사탕을 골랐던지 상관없이 X를 골랐을 확률. 문제에서는 상자을 임의로 선택한 것이므로 0.5라고 가정할 수 있습니다. 이를 데이터를 보기 전의 가설의 확률, 즉 **사전확률**입니다.

P(A|B) = X에서 노란색 사탕이 나올 확률. 3/4입니다. 이를 데이터가 가설에 포함될 확률, 즉 **우도**입니다.

P(A) = 노란 사탕을 고를 확률입니다. X, Y에 50개 노란 사탕과 30개의 파란 사탕이 들어있으므로 P(A)는 5/8이 됩니다. 이를 어떤 가설에든 포함되는 데이터의 비율, 즉 **한정상수**입니다.

P(B|A) = 노란색 사탕이 X박스에서 나왔을 확률. 우리가 알고 싶은 확률입니다. 이를 데이터를 확인한 이후의 가설 확률, 즉 **사후확률**입니다.

Term Frequency - Inverse Document Frequency(TF-IDF)

TF-IDF는 TF X IDF 연산 결과

TF : (단어 빈도, Term Frequency)
IDF : (역문서 빈도, Inverse Document Frequency)

DF : (문서 빈도, Document Frequency)

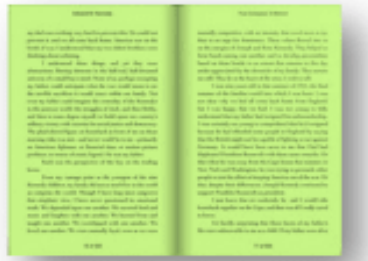
TF 표현식

Boolean Frequency : $tf(t,d) = 0\ or\ 1$
Log scale Frequency : $tf(t,d) = \log(f(t,d) + 1)$

IDF 표현식

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}| + 1}$$

Measuring similarity



1	0	0	0	5	3	0	0	1	0	0	0	0

1*3
+
5*2
= **13**

3	0	0	0	2	0	0	1	0	1	0	0	0