Maximum A Posterior (MAP)

소프트웨어 꼰대 강의

노기섭 교수

(kafa46@cju.ac.kr)

Possible Learning in Bayesian

- **Maximum Likelihood Estimation (MLE)**
  - Same as frequentist!
  - Dataset only!

목표: 오직 Likelihood만 최대화

$$P(\theta|D) = \frac{\boxed{P(D|\theta)} \times P(\theta)}{P(D)} \approx P(D|\theta)$$

- **Maximum A Posterior (MAP)**
  - $P(D)$: 알고(given) 있다고 가정
  - $P(\theta)$: 정규분포라고 가정

Likelihood, prior를 동시에 최대화

$$\boxed{P(\theta|D)} = \frac{\boxed{P(D|\theta)} \times \boxed{P(\theta)}}{P(D)} \approx P(D|\theta) \times P(\theta)$$

- **Bayesian Inference (Variation Inference)**
  - Likelihood, Posterior, Evidence 모두 고려
  - Computing $P(D)$ is intractable
  - Alternatively, using Variational Inference
  - $P(\theta|D)$ 계산이 어렵기 때문에 우리가 알고 있는 함수를 이용하여 잘 모사하도록 접근

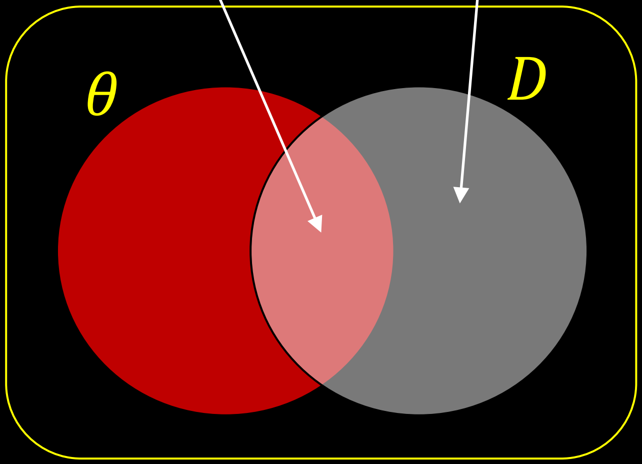
$$\boxed{P(\theta|D)} = \frac{P(D|\theta) \times P(\theta)}{P(D)} \approx \boxed{Q(\theta|\theta')}$$

목표: P를 잘 흉내내는 Q의 파라미터 θ'찾기

Recap: Bayesian Structure

그림으로 쓱 보기~

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)} = \frac{P(D|\theta) \times P(\theta)}{\boxed{P(D|\theta)P(\theta)} + \boxed{P(D|\theta^c)P(\theta^c)}}$$



Graphical Understanding of Bayesian

- **$P(D)$ 를 찾자! 암 검진 사례로 살펴보는 베이지안**
  - 모든 사람 중에서 10% 사람은 암에 걸린다.
  - 암 검진 정확도는 80% 이다.
  - 검진결과 암 판정을 받았다.
  - 당신이 암에 걸렸을 확률은?

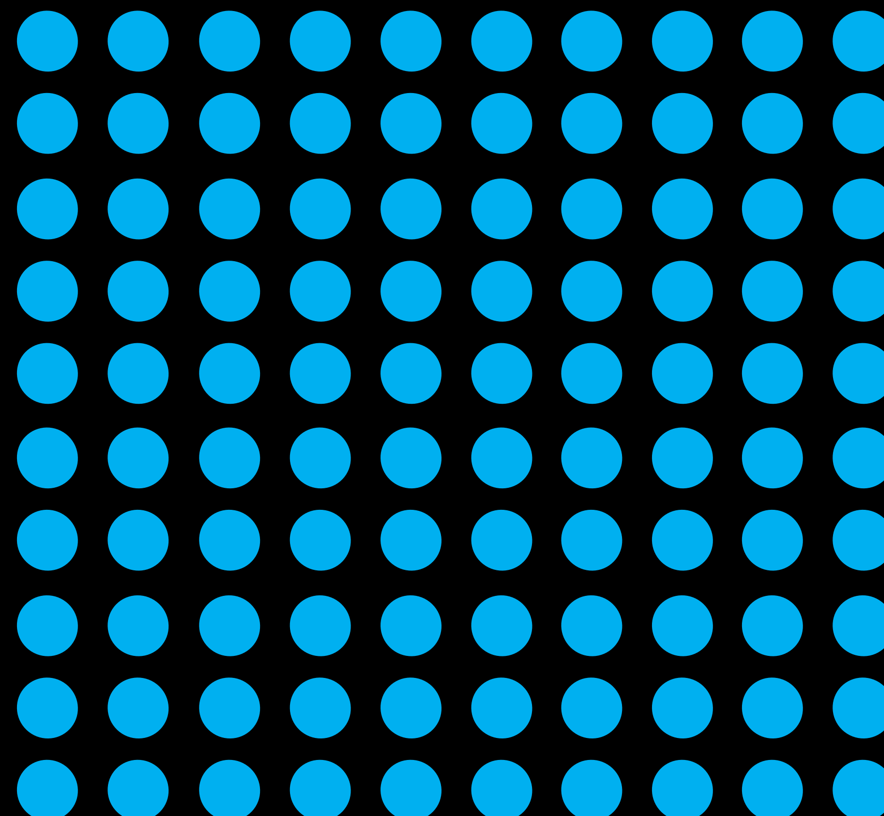
Belief 업데이트가
어떻게 되는거임?

Bayesian update?

난 80 확률로 암에 걸렸다?
아니면 다른 확률?

Example Step 01.

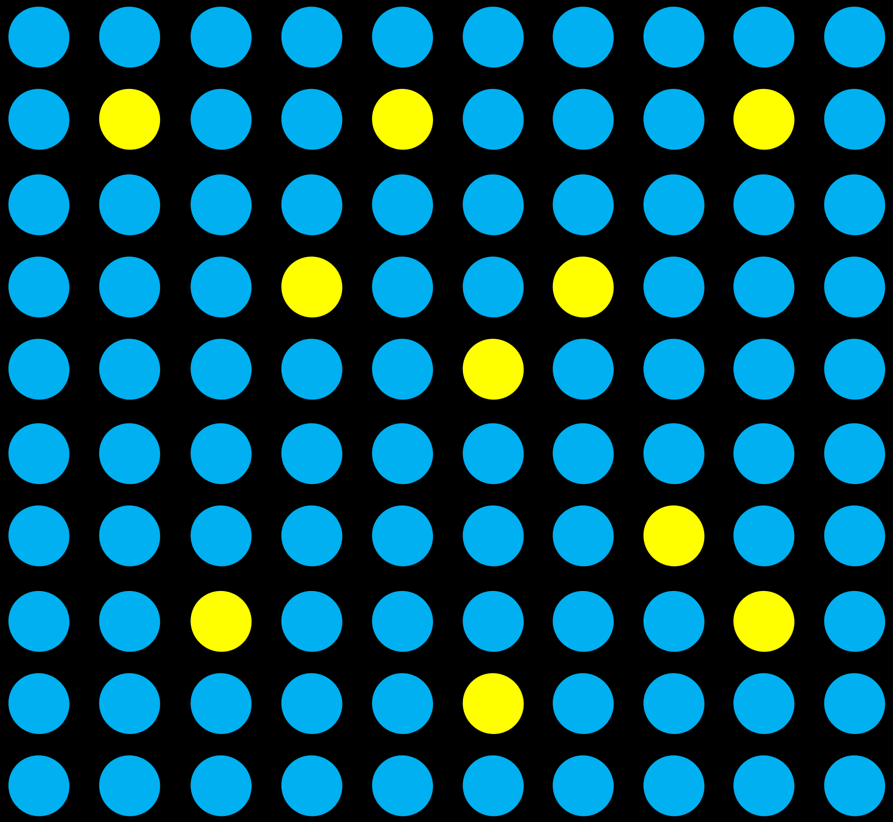
100명 샘플링



Example Step 02.

100명 샘플링

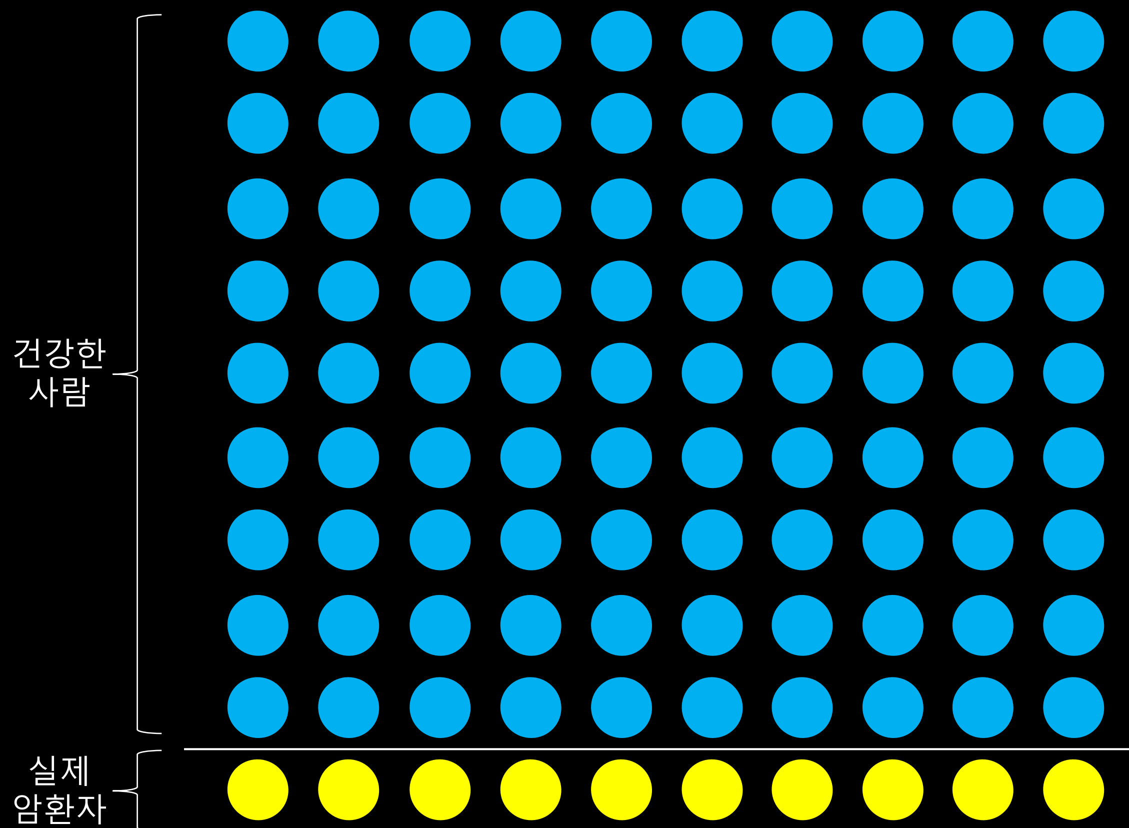
랜덤하게 10%는
암에 걸림



100명 샘플링

랜덤하게 10%는
암에 걸림

아래쪽으로
이쁘게 정리하면…

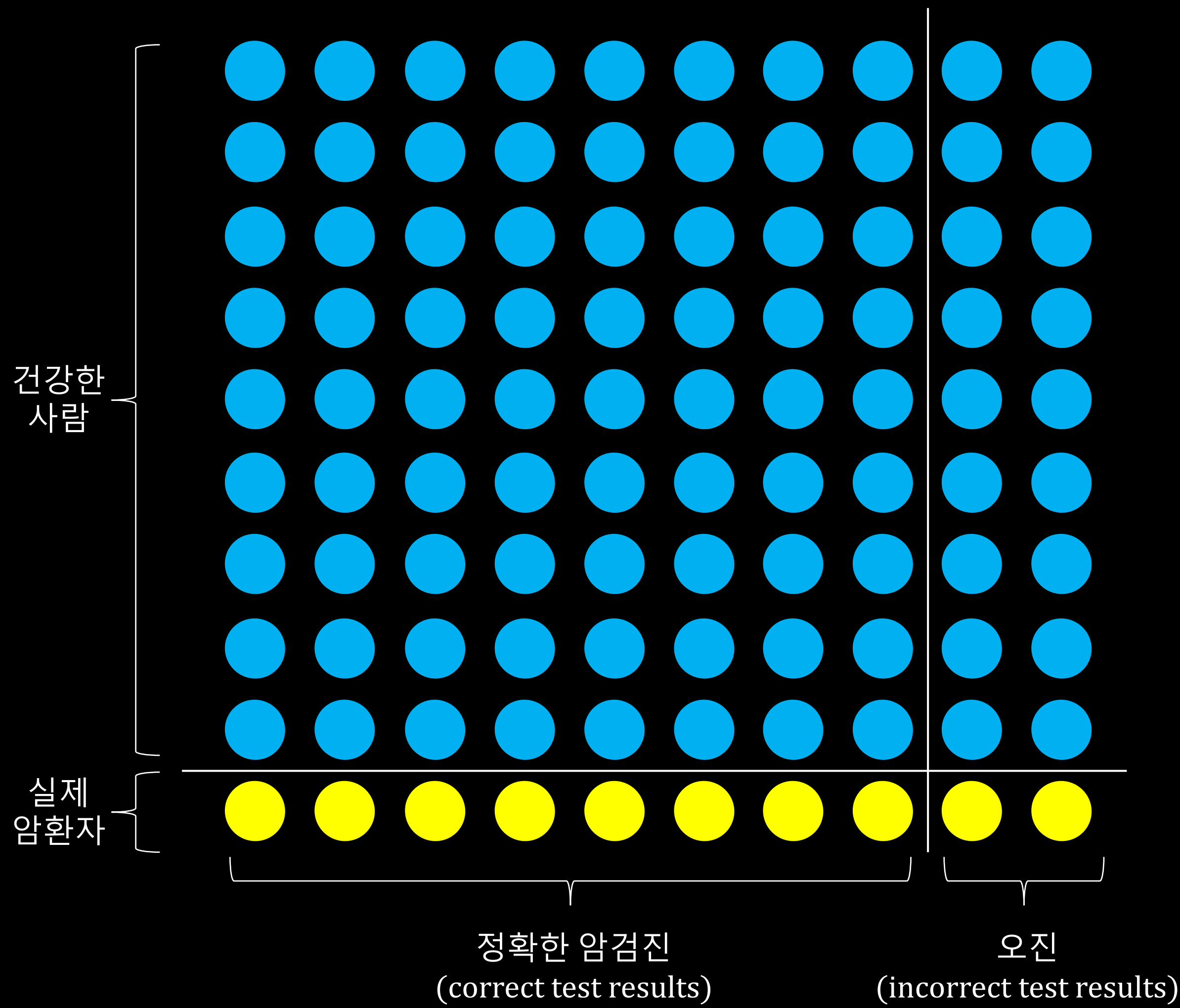


Example Step 03.

100명 샘플링

랜덤하게 10%는
암에 걸림

검진 결과 20%는
오진
(건강한데 암으로 판정)

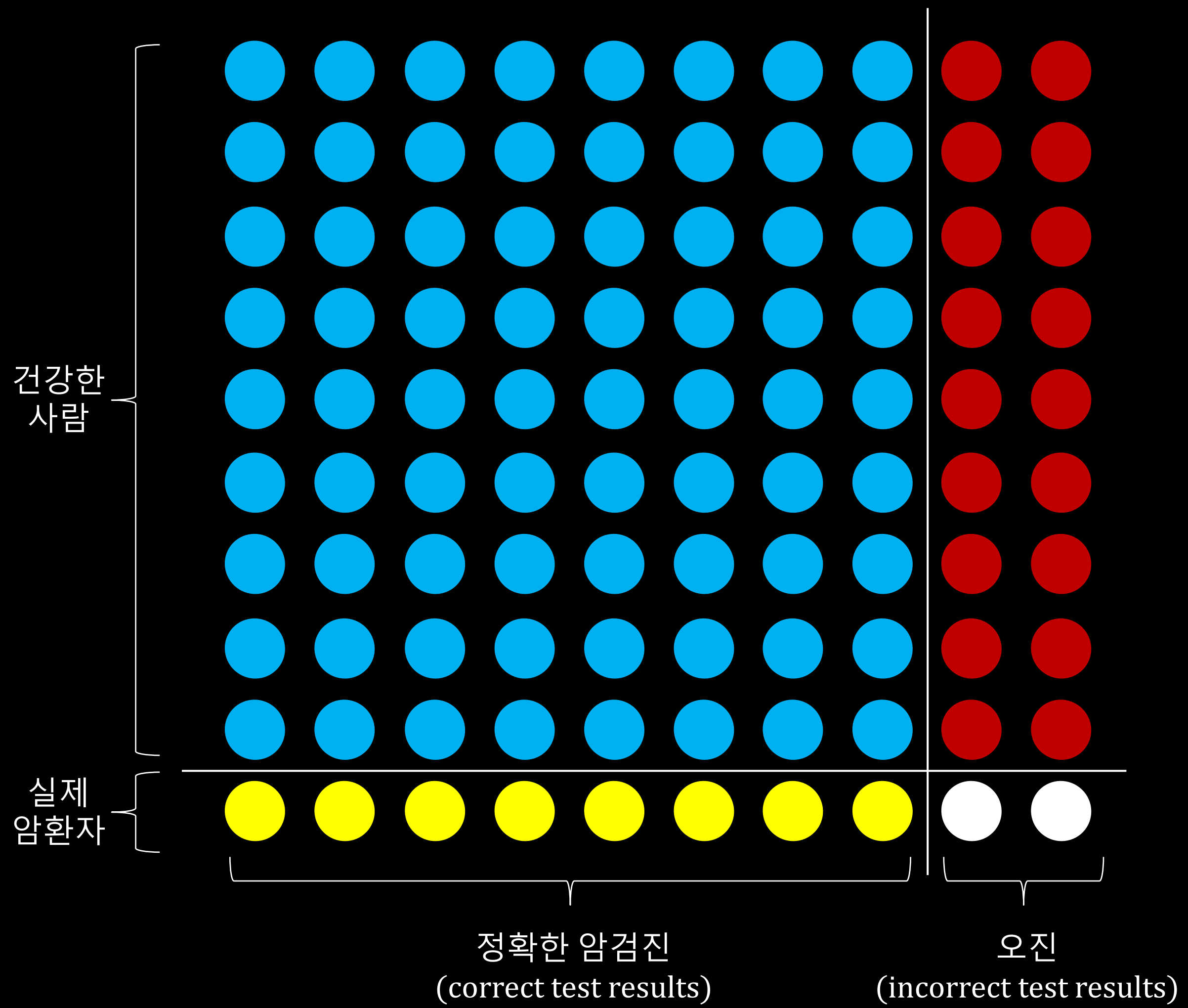


Example Step 04.

100명 샘플링

랜덤하게 10%는
암에 걸림

검진 결과 20%는
오진
(건강한데 암으로 판정)



Example Step 05.

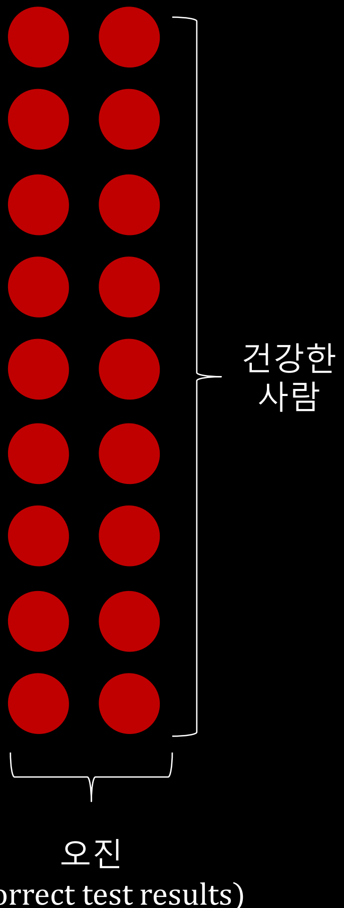
Bayesian belief update via Evidence

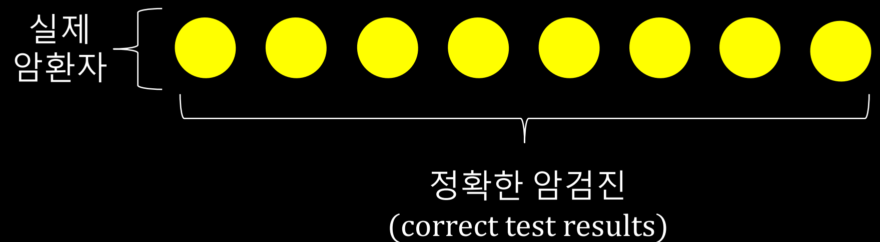
10% ⇨ 31%

당신이 암에 걸렸을 확률은?

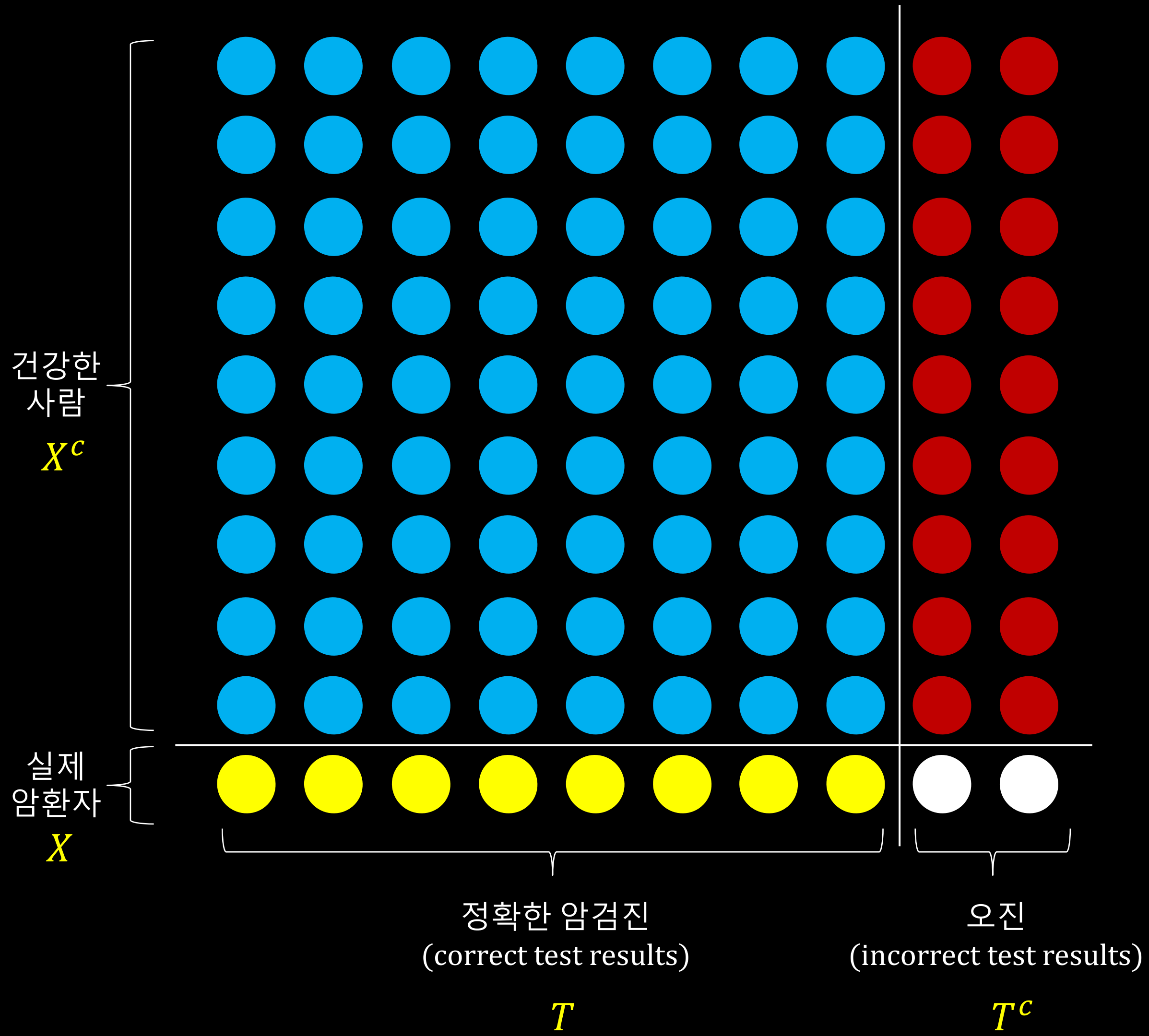
$$\frac{\text{실제로 암 걸린 인원}}{\text{암 판정 받은 전체 인원}} = \frac{8}{8 + 18} = \frac{8}{26} \cong 30.1\ \%$$

$$= \frac{\text{암인데 암으로 진단}}{\text{암인데 암으로 진단} + \text{건강한데 암으로 진단}}$$

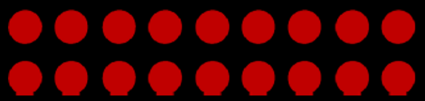




Example Step 06.



$= 전체인원 \times P(X) \times P(T|X)$
$= 100 \times 0.1 \times 0.8 = 8$명



$= 전체인원 \times P(X^c) \times P(T|X^c)$
$= 100 \times 0.9 \times 0.2 = 18$명

주목할 점....

검진(test) 종류에 따라

$P(T|X)$ 및 $P(T|X^c)$ 값이 다르다.

Reaching to Bayesian & Computable Evidence

$= \text{전체인원} \times P(X) \times P(T|X)$
$= 100 \times 0.1 \times 0.8 = 8\text{명}$

$= \text{전체인원} \times P(X^c) \times P(T|X^c)$
$= 100 \times 0.9 \times 0.2 = 18\text{명}$

참고:
검진(test) 종류에 따라
$P(T|X)$ 및 $P(T|X^c)$ 값이
다르다.
→ 그때 그때 Befief 업데이트!

$$P(X|T) = \frac{\cancel{\text{전체인원}} \times P(X) \times P(T|X)}{\cancel{\text{전체인원}} \times P(X) \times P(T|X) + \cancel{\text{전체인원}} \times P(X^c) \times P(T|X^c)}$$

$$P(X|T) = \frac{P(X)P(T|X)}{P(X)P(T|X) + P(X^c)P(T|X^c)} = \frac{P(T|X)P(X)}{\boxed{P(T|X)P(X) + P(T|X^c)P(X^c)}} = \frac{P(T|X)P(X)}{\boxed{P(T)}}$$

어쨌든 구할 수 있다.

$$P(H|E) = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(H^c)P(H|E^c)}$$

순서 바꾸고
간략히 정리

$$= \frac{P(E|H)P(H)}{P(E|H)P(H) + P(H|E^c)P(H^c)} = \frac{P(E|H)P(H)}{P(E)}$$

딥러닝에 맞게
바꿔주기

$$P(\theta|D) = \frac{P(D|\theta)P(D)}{\boxed{P(D)}}$$

딥러닝에서는...
가능한 θ 종류가 너무 많다.
θ^c 공간이 너무나 크다. ㅠ
하지만, 어쨌든 구할 수는 있다....

어떤 값으로
존재한다는 것이 중요!!

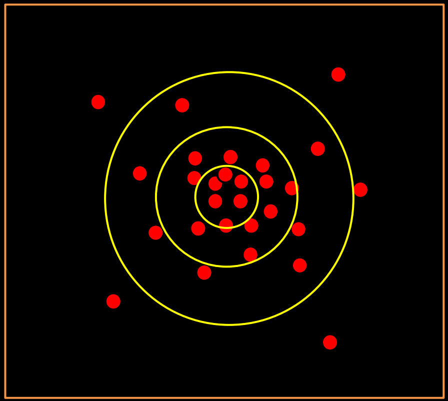
본격적으로 MAP 시작!

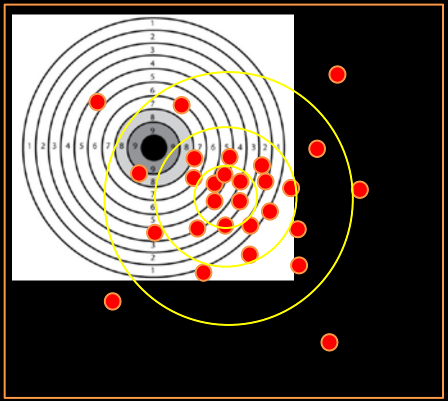
- **MLE (Maximum Likelihood Estimation) → Likelihood 최대화**
- **그렇다면 MAP는?**
  - Prior 까지 같이 최대화
  - Posterior를 최대화 == Likelihood와 Prior를 동시에 최대화

$$P(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta}) \times P(\boldsymbol{\theta})}{P(\boldsymbol{D})}$$

  - 사전 지식 $P(\theta)$를 알고 있다면 좋은 점???
    - Overfitting 감소

만약 대충이라도....
과녁의 위치를 약간이라도 알았다면?





Find Prior Objective

■ **어떻게 Prior에 적용하나요?**

$$\boldsymbol{P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)}}$$

$\boldsymbol{P(D)}$는 안다고 가정 ➔ $\eta = \dfrac{1}{P(D)}$

$$\boldsymbol{P(\theta|D) = \eta \times P(D|\theta) \times P(\theta)}$$

"θ에 Prior를 부여한다(걸어준다)" 라고 표현합니다.

- Prior $P(\theta)$는 어떤 확률 분포를 가지고 있을 것임
  - 정규분포라고 가정 (참고: 다른 분포라고 가정해도 됨)

$$P(\theta) \sim N(\mu_\theta, \sigma_\theta^2) = \prod_{i=1}^{k} \frac{1}{\sigma_\theta \sqrt{2\pi}} e^{-\frac{(\mu_\theta - \theta_i)^2}{2\sigma_\theta^2}}$$

$$\ln P(\theta) = \ln \prod_{i=1}^{k} \frac{1}{\sigma_\theta \sqrt{2\pi}} e^{-\frac{(\mu_\theta - \theta_i)^2}{2\sigma_\theta^2}} = -\sum_{i=1}^{k} \left(\ln \sigma\sqrt{2\pi} + \frac{(\mu_\theta - \theta_i)^2}{2\sigma_\theta^2} \right)$$

$$\underset{\theta}{\operatorname{argmax}} \ln P(\theta) = \underset{\theta}{\operatorname{argmax}} -(\mu_\theta - \theta_i)^2$$

$$\underset{\theta}{\operatorname{argmin}} \ln P(\theta) = \underset{\theta}{\operatorname{argmin}} (\mu_\theta - \theta_i)^2$$

Find MAP Objective

$$P(\boldsymbol{\theta}|\boldsymbol{D}) = \boldsymbol{\alpha} \times P(\boldsymbol{D}|\boldsymbol{\theta}) \times P(\boldsymbol{\theta})$$

$$\ln P(\boldsymbol{\theta}|\boldsymbol{D}) = \ln(\boldsymbol{\alpha} \times P(\boldsymbol{D}|\boldsymbol{\theta}) \times P(\boldsymbol{\theta}))$$

$$\ln P(\boldsymbol{\theta}|\boldsymbol{D}) = \ln \boldsymbol{\alpha} + \ln P(\boldsymbol{D}|\boldsymbol{\theta}) + \ln P(\boldsymbol{\theta})$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln P(\boldsymbol{\theta}|\boldsymbol{D}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln \boldsymbol{\alpha} + \ln P(\boldsymbol{D}|\boldsymbol{\theta}) + \ln P(\boldsymbol{\theta})$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln P(\boldsymbol{\theta}|\boldsymbol{D}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ (\ln P(\boldsymbol{D}|\boldsymbol{\theta}) + \ln P(\boldsymbol{\theta}))$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln P(\boldsymbol{\theta}|\boldsymbol{D}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{k} \frac{(\mu_\theta - \theta_i)^2}{2\sigma_\theta^2} \right)$$

$$\underset{\theta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{i=1}^{k} (\mu_\theta - \theta_i)^2 \right) \Rightarrow \underset{\theta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \|\theta\|_2^2 \right)$$

Finally, We Check Results…

$$\underset{\theta}{\operatorname{argmin}}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \boxed{\alpha\|\theta\|_2^2}\right)$$

요거를... Penalty 또는
Regularization term 이라고
부릅니다 ^^.

간략히 표현 (지저분하니까...)

$$\underset{\theta}{\operatorname{argmin}}\left(\left(Y - \hat{Y}\right)^2 + \alpha\|\theta\|_2^2\right)$$

	MLE	MAP
Goal	Find Optimal Parameter (θ) Set	
Objective: Minimize	$\left(Y - \hat{Y}\right)^2$	$\left(Y - \hat{Y}\right)^2 + \alpha\|\theta\|_2^2$

수고하셨습니다 ..^ ^..