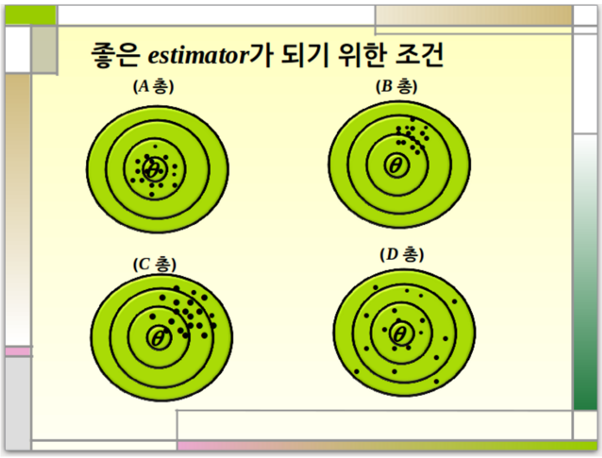
Bayesian (2)

Estimator (추정량)

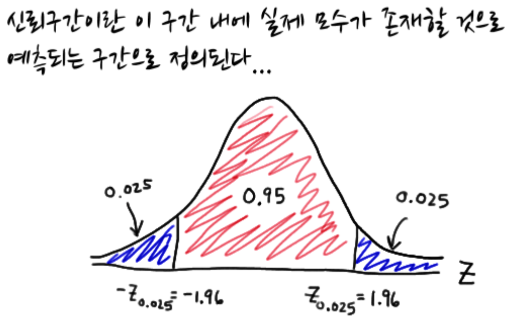
: 관찰된 표본에서 모집단을 추론하는 방식



Statistical Estimation (통계적 추정)

: 표본을 통해서 모집단의 값을 확률적으로 추론하는 과정

- Point Estimation (점 추정): 특정 값을 예측하는 것
- Interval Estimation (구간 추정): 일정한 구간을 예측하는 것. Ex) 신뢰구간



Parameter Estimator (파라미터 추정)

Bayesian Estimate (베이즈 추정)

: 주어진 데이터를 기반으로 모수 μ의 조건부확률분포 p를 계산하는 직업. 베이즈 정리 사용

$$p(\mu \mid x_1, \ldots, x_N) = \frac{p(x_1, \ldots, x_N \mid \mu) \cdot p(\mu)}{p(x_1, \ldots, x_N)} \propto p(x_1, \ldots, x_N \mid \mu) \cdot p(\mu)$$

- $p(\mu)$ = 모수의 prior 분포
- $p(\mu \mid x_1, \ldots, x_N)$ = 모수의 posterior 분포
- $p(x_1, \ldots, x_N \mid \mu)$ = 모수의 likelihood 분포

모수 분포 표현하는 방법

1) **parametric** (모수적) 방법: 다른 확률분포를 이용해 추정된 모수의 분포 나타냄. Ex) 베르누이 분포, 카테고리분포, 정규분포

2) **non-parametric** (비모수적) 방법: 모수의 분포와 동일한 분포를 가지는 표본 집합을 만들어 히스토그램 등으로 분포 표현. Ex) MC (몬테카를로) 방법

MAP (Maximum A Posterior)

: 구하고자 하는 대상을 posterior를 사용해서 (=prior 알고 있을 때) 추정할 때

- prior를 반영하기 때문에 output에 대해 특정 제약조건을 걸고 싶은 경우 사용

Weight Decay Term

Weight decay = regularization

- L2 regularization 적용 = 파라미터의 prior를 가우시안 분포로 걸어주는 것

MLE (Maximum Likelihood Estimation)

: likelihood가 최대가 되는 파라미터를 찾는 것. log를 취해서 계산

⇒ 예측값과 실제 값의 차이의 제곱. = L2 Loss

- 딥러닝에서 L2 Loss를 사용한다 = 주어진 데이터로부터 likelihood를 최대화 하겠다
- 분류 문제에서는 베르누이 분포를 이용하여 cross entropy error 유도 (link)

Minimize Cross-Entropy

Entropy = 기댓값

Cross entropy = 두 확률 분포 사이의 차이를 측정하는 지표

$$H(p,q) = H(p) + D_{KL}(p||q) = -\sum_{i=0}^{n} p(x_i)\log(q(x_i))$$

- true distribution p의 entropy + p, q의 KL divergence

KL divergence = “놀라움”

- 두 확률분포가 가까웠다는 가정 하에, 가깝지 않다면 “놀라운 일” 이고 KL divergence는 높은 값을 갖게 됨.
- 분류문제에서 cross-entropy를 최소화, 즉 KL divergence를 최소화하면서 NN 학습

Latent Variable

: 실제 관측이 되지 않았지만 관측된 데이터에 상호 영향을 미칠 것이라 판단되는 변수

EM Algorithm (기댓값 최대화 알고리즘)

: Latent variable을 사용하여 MLE 추정량 구하는 방법

- 관측 데이터 X, latent variable Z, 추정해야할 파라미터 θ
- 주어진 추정값 θ^*를 업데이트 해나가는 iterative 알고리즘임

알고리즘)

단계 1) 파라미터 초기값 $\theta^{(0)}$을 설정한다.

단계 2) $\theta^{(t)}$에 대하여 다음을 계산한다.

$$\boxed{\begin{aligned} Q(\theta|\theta^{(t)}) &= \int_z \{\log f_\theta(x,z)\}\, f_{Z|X,\theta^{(t)}}(z)dz \\ &= E_{Z|X,\theta^{(t)}}\{\log f_\theta(X,Z)\} \end{aligned}}$$

이 단계는 조건부 기대값을 구하는 과정이라 볼 수 있어서 Expectation Step이라고 한다.

단계 3) $Q(\theta|\theta^{(t)})$를 최대화하는 $\theta^{(t+1)}$을 찾는다. 즉,

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)})$$

이 단계는 Q를 최대화한다는 의미에서 Maximization Step이라고 한다. 이제 알고리즘 명칭이 왜 EM 인지 알았을 것이다.

단계 4) 수렴할 때까지 단계 2) ~ 단계 3)을 반복한다.