<보험 data 복제>

$\lambda = X_1 + \cdots + X_p$, $p=100$(차원) → 복제

$Y \sim Pois(\lambda)$ ① Data → $\hat{\beta}$ ② MSE = $\frac{\sum_{i=1}^{n}(y_i - \hat{\lambda}_i)^2}{n}$ good

$\lambda = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p)$

복제: ① 복제 → $\hat{\beta}^*$ (X^*, Y^* 관계) ② MSE = $\frac{\sum_{i=1}^{n}(\hat{\beta}_i^* - \hat{\lambda}_i^*)^2}{n}$ bad

$\hat{\lambda}_i^* = \exp(\hat{\beta}_0^* + \cdots + \hat{\beta}_p^* X_{ip}^*)$ → 가짜 data 얼마나 잘 설명하는지 보려고 X.

$y^* \sim Pois(\lambda)$ 복제한 data, $\hat{\lambda} = \exp(\hat{\beta}_0^* + \hat{\beta}_1^* X_1 + \cdots + \hat{\beta}_p^* X_p)$ ← 가짜 data 성능 얼마나 나쁠까?

비교 → 뭐가 더 좋은지 알수 없음

비교 : Original MSE better 항상 $\hat{\beta}$으로 회귀직선

$\hat{\beta}$ 찾을때 : $\hat{\beta}^* = \underset{\beta^*}{\text{argmin}}$ → Original data X, Y 관계 잘 설명해야 되는데 가짜 data가 Bad.

MSE) $\frac{1}{n}\sum_{i=1}^{n}(y_i - \exp(\hat{\beta}_0^* + \hat{\beta}_1^* X_{i1} + \cdots + \hat{\beta}_p^* X_{ip}))^2$

* ↳ 더 나은 경우 없다. $\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{\sum_{i=1}^{n}(y_i - \lambda_i)^2}{n}$ ↙ 엉뚱한 $\hat{\beta}$(가짜) 가져옴 $\hat{\beta}^*$

• Test용 set에서 X, Y 발생, 복제해보기

Test data: y_{n+1} X_{n+1} ~ y_{n+T} (T개)

$\frac{1}{T}\sum_{i=n+1}^{n+T}(y_i - \exp(\hat{\beta}_0^* + \hat{\beta}_1^* X_{i1} + \cdots + \hat{\beta}_p^* X_{ip}))^2$

= MSE replicated

MSE Original = $\frac{1}{T}\sum_{i=n+1}^{n+T}(y_i - \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}))^2$

Original data → $\hat{\beta}$, 복제 data → $\hat{\beta}^*$

$p=100$

$\vec{X}_i \sim MVN(\vec{0}, \Sigma_p)$, $[\beta_0, \cdots, \beta_p \sim unif(-1,1)]$ → 고정 $\sigma^2=1$ 고정

$y_i \sim N(u_i, \sigma^2)$, $u_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$

$= \beta_0 + \beta_i^* X_i$ for $i = 1, \cdots, n$

$[y_1 \ \vec{X}_1 \ \cdots \ y_n \ \vec{X}_n]$ → VAE → $[y_1^* \ \vec{X}_1^* \ \cdots \ y_n^* \ \vec{X}_n^*]$

MSE = $\frac{\sum_{i=n+1}^{n+T}(y_i - (\hat{\beta}_0 + \hat{\beta}\vec{X}_i))^2}{T}$ ($\hat{\beta}$)

Test MSE (Test data 사용)

$MSE^* = \frac{\sum_{i=n+1}^{n+T}(y_i - (\hat{\beta}_0^* + \hat{\beta}^*\vec{X}_i))^2}{T}$ ($\hat{\beta}^*$)

누가 더 클지 (여러번 반복)

$\frac{\sum_{b=1}^{B} MSE_b}{B} \underset{10}{=}$ 더 좋음 (original 랑비)

$<$ $\frac{\sum_{b=1}^{B} MSE_b^*}{B} \underset{10}{=}$

Test: $[y_{n+1} \ \vec{X}_{n+1} \ \cdots \ y_{n+T} \ \vec{X}_{n+T}]$

⇒ Iteration $b=1, \cdots, B$ ∴ β 항상 바뀜

* 차원 ↑ Y 예측력↓ 차원(p)↓ Y 예측력↑

100차원

ex. iteration 10번(p=10) 계산 $\overline{MSE_b}$ (평균), $\overline{MSE_*}$ → $\frac{\overline{MSE_b}(p)}{\overline{MSE_b^*}(p)}$ 비율) 10, 100차원 다름 ∴ 비교하기 수월

p=10 20 ... 100 시도 → 어느순간부터 예측력 나빠지는지

결론

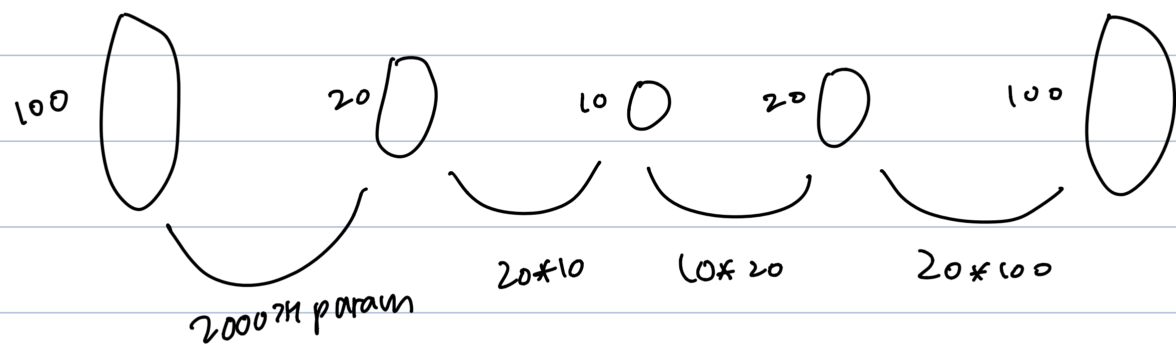
문제: GAN, AE, VAE ⇒ 복제 잘 안되서. Y 예측에서어려워하니까 MSE 안좋음.

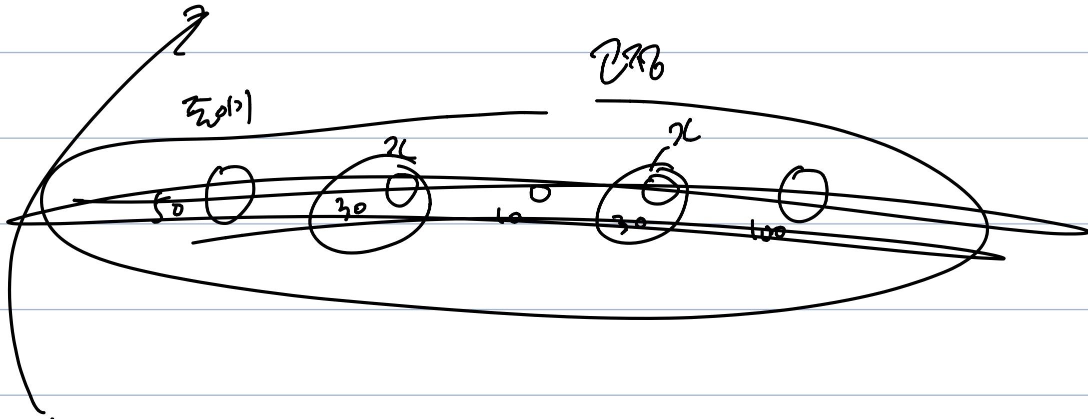
↳ 발전 data 10~100개 정리해 보기 (h/w)

B=10 X 차원 10, 20 ... 100 까지

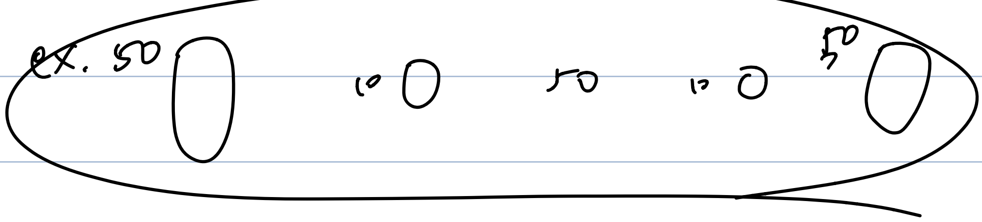
1개 data=10000개 정도

<AE>





data 10만개로 하기! 100 20 10 20 100 정도로 해보기



param 개수 신경써서 맞추기

10:1 비율로 줄어들게 하기

답주) 발표...? (여유) discrete GAN model 어떻게 구현 하는지

(지연우 논문 참고)

+ 보고서 읽기, 분석