

20기 정규세션

Survival Analysis

TOBIG's 19기 이동준

CONTENTS

01 Introduction to Survival Analysis

02 Function of Survival Analysis

03 CoxPH

04 DeepHIT

05 DRSA

06 Metric

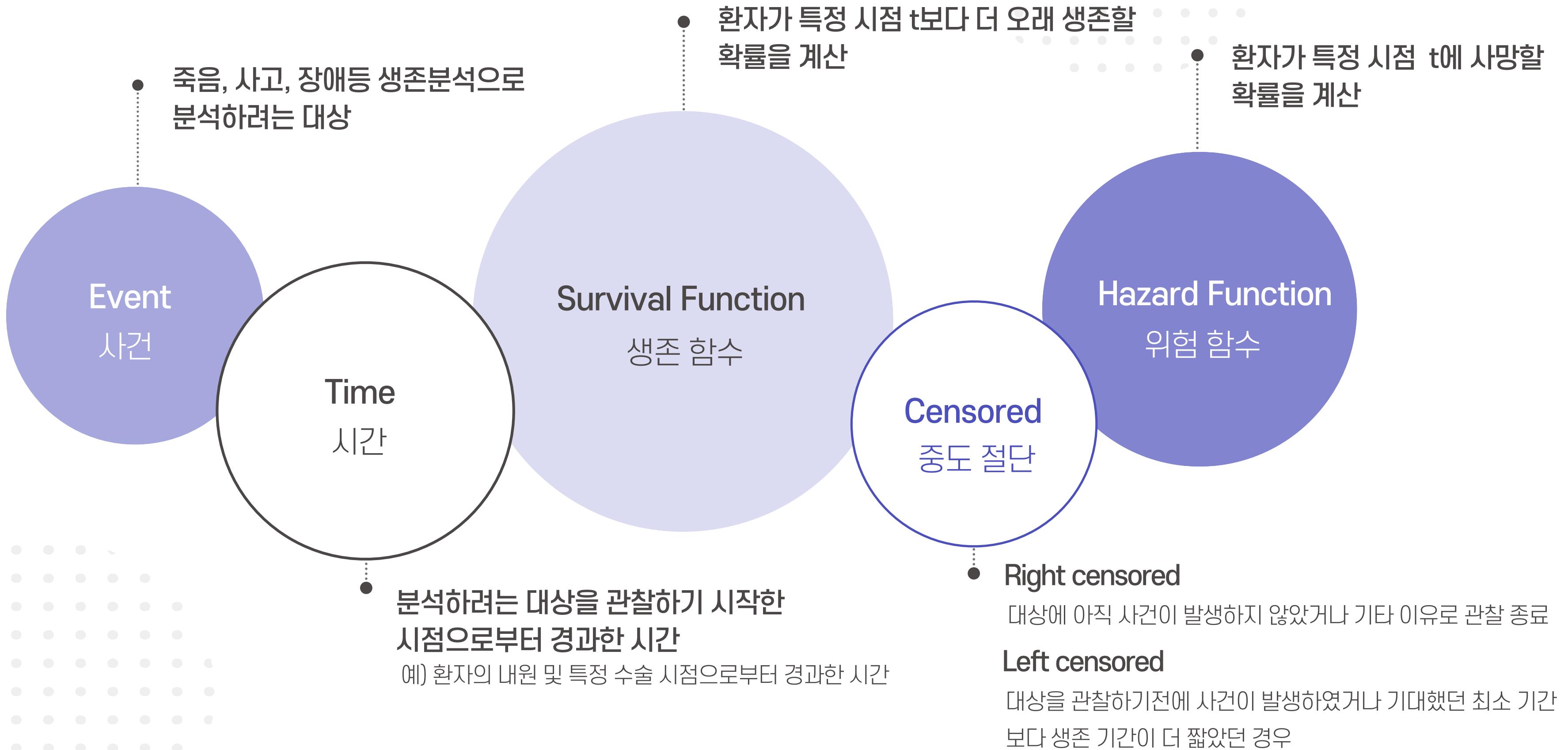
01

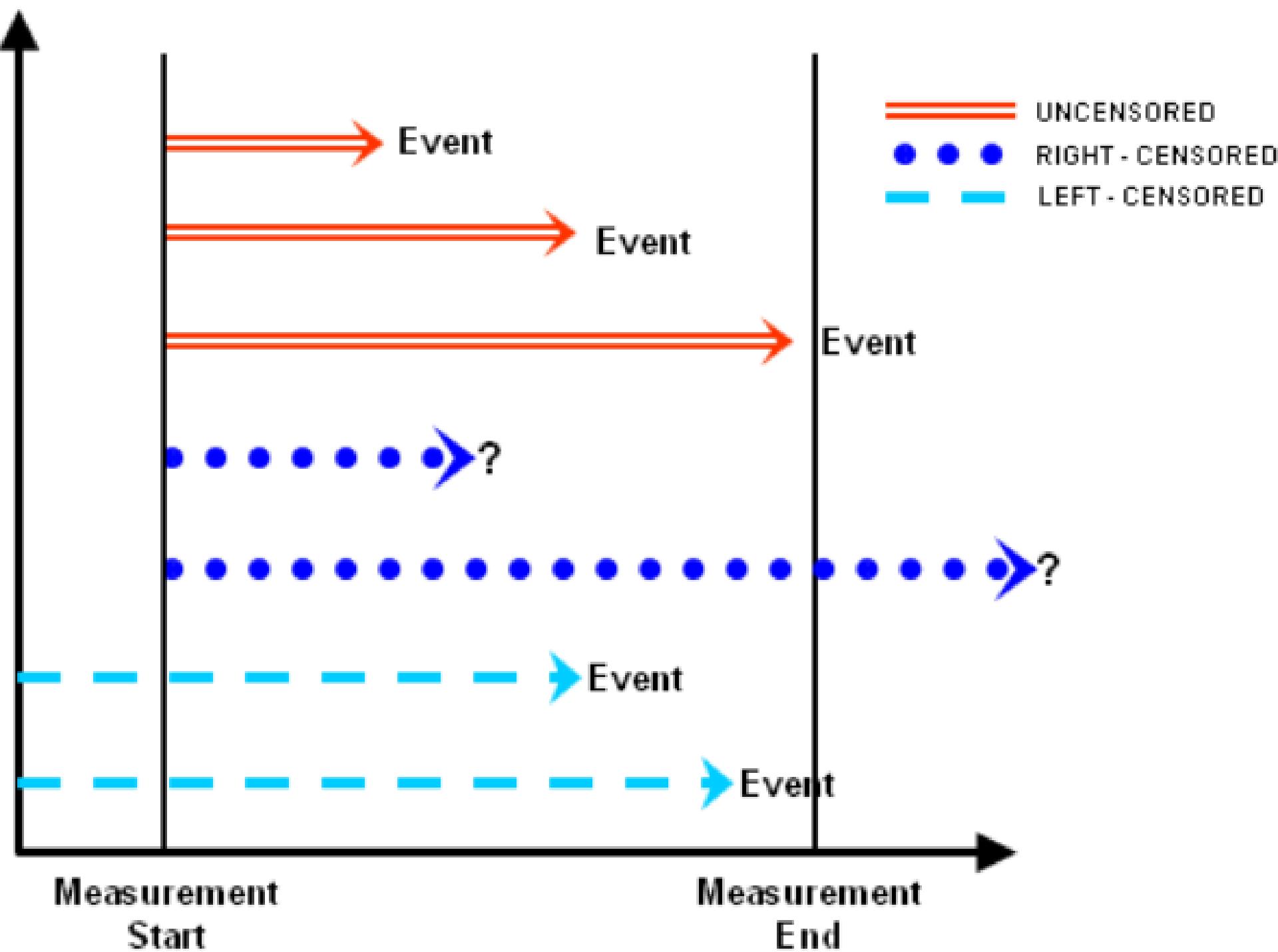
Introduction to Survival Analysis

생존 분석

어떤 사건의 발생 확률을 시간과 함께
생각하는 통계 분석 및 예측 방법

- 기계 공학 Reliability Analysis
- 경제학 Duration Analysis
- 사회학 Event-history Analysis
- 경영학 Customer Defection Analysis
- 의학통계학 Survival Analysis





Data Description

	δ	T	X									
ID	Death Cause	Survival Time (m)	Lymphnod Node	Age	Gender	Married	...	Benign Tumors	Malignant Tumors	Histology ICD	EOD	
1	2	57	0.4061	53	1	1		0	2	65	0.0080	
2	1	71	0.1382	56	1	1		0	2	64	0	
3	0	135	0.1600	60	1	1	...	0	2	65	0.0620	
4	0	120	0.2195	50	1	0		0	1	65	0	
5	1	29	0.7998	55	1	1		0	2	64	0	
6	2	71	0.7998	55	1	0		0	2	64	0	

02

Function of Survival Analysis

Survival Function

관찰 대상이 특정 시점 t 보다 더 오래 생존할 확률

$$S(t) = \Pr(T > t)$$

Event Density Function

각 시점 t 의 사건 발생 확률

$$f(t) = F'(t) = \frac{d}{dt} F(t)$$

Lifetime Distribution Function

관찰 대상이 특정 시점 t 안에 사망할 확률

$$F(t) = \Pr(T \leq t) = 1 - S(t)$$



Survival Event Density Function

생애분포함수($F(t)$)와 사건밀도함수($f(t)$)로 생존함수를 표현

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(u)du = 1 - F(t)$$

$$s(t) = S'(t) = \frac{d}{dt} \int_t^{\infty} f(u)du = \frac{d}{dt} [1 - F(t)] = -f(t)$$



Hazard Function

관찰 대상이 특정 시간 t 에 사망할 확률

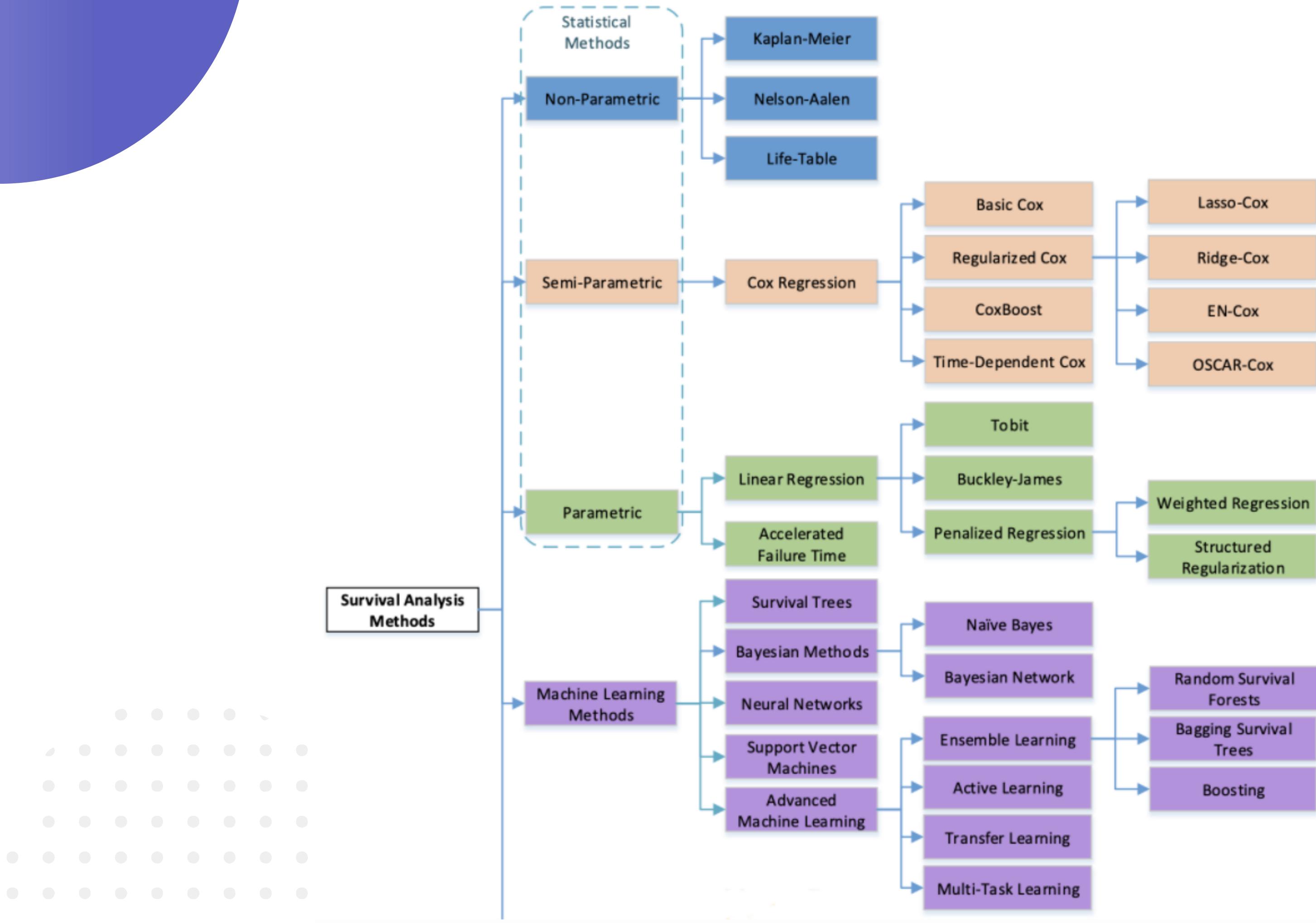
: 대상이 특정 시간까지 생존한 상태에서 정확히 t 시점에서 사망할 확률

$$h(t) = \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t)$$



03

Cox Proportional Hazard Model



Cox Proportional Hazard Model

CoxPH 모델은 각 환자의 covariate를 고려하여 생존함수를 추정하는 방법

1. 위험함수를 기저위험함수(baseline hazard)와 매개변수로 나누어 정의함.
2. exponential function은 hazard function의 non-negative한 성질을 만족하기 위해 쓰임.
3. X는 흡연여부, 성별 등 환자의 정보, beta는 모델이 학습해야하는 계수

$$h(t|X = x) = h_0(t)\exp(X^T \beta)$$

Cox Proportional Hazard Model

$$h(t|X = x) = h_0(t) \exp(X^T \beta)$$

$$\frac{h(t|X_1)}{h(t|X_2)} = \frac{h_0(t) \exp(X_1^T \beta)}{h_0(t) \exp(X_2^T \beta)} = \exp[(X_1 - X_2)^T \beta]$$

가정 1. 생존함수가 Exponential Function을 따름

가정 2. 두 군의 위험비가 연구기간동안 일정하게 유지된다는 proportional hazard 가정



Partial likelihood Function

$$L(\beta) = \prod_{i:\delta_i=1} \frac{\exp(X_i^T \beta)}{\sum_{j:y_j \geq y_i} \exp(X_j^T \beta)}$$

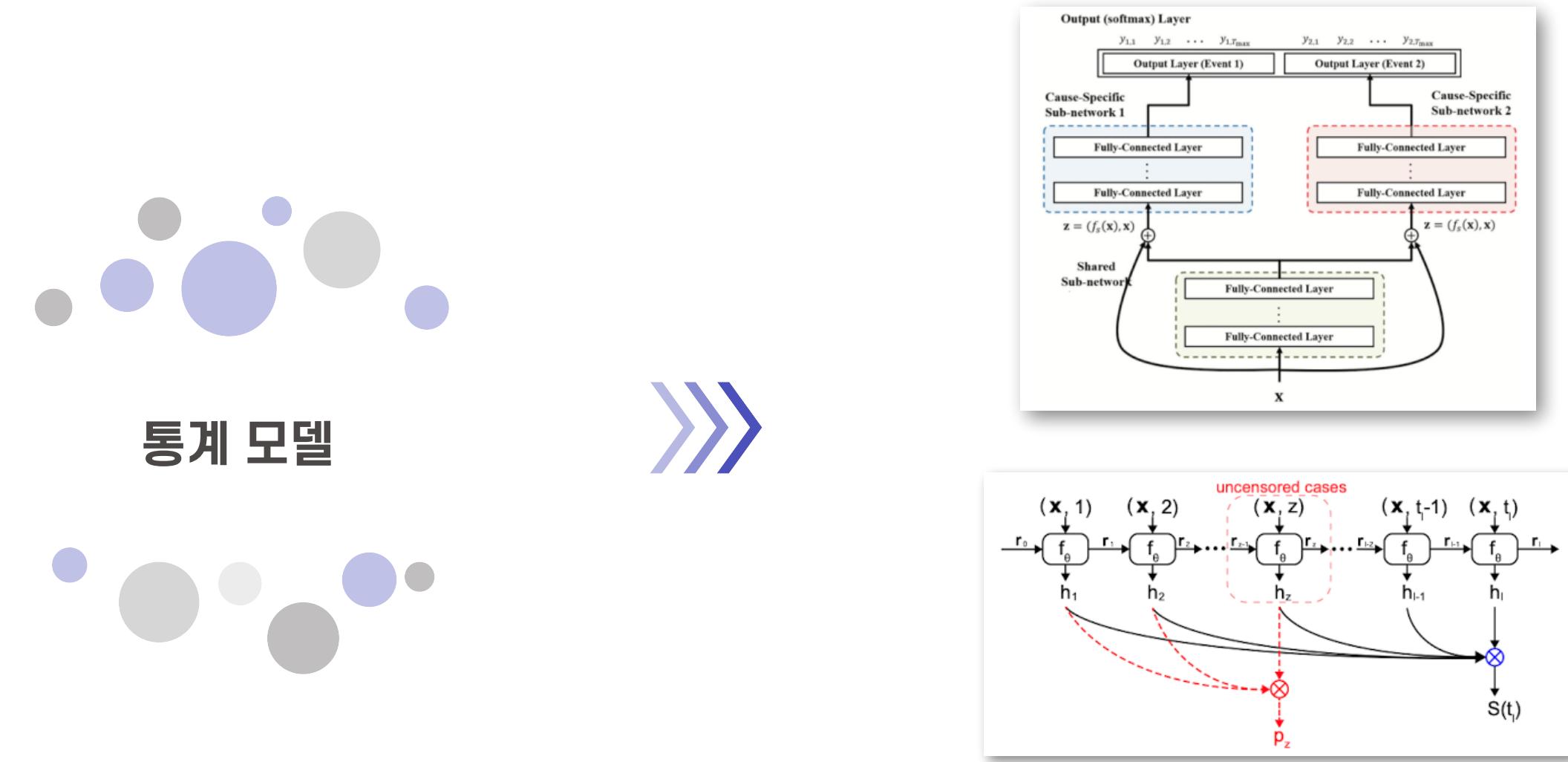
$$l(\beta) = \sum_{i:\delta_i=1} \left(X_i^T \beta - \log \left[\sum_{j:y_j \geq y_i} \exp(X_j^T \beta) \right] \right)$$

1. Censored된 Instance는 포함시키지 않는다.
2. 다만, ordering 관점에서 uncensored < censored의 경우 비교가 가능하기 때문에 이를 부분적으로 포함한다.



04

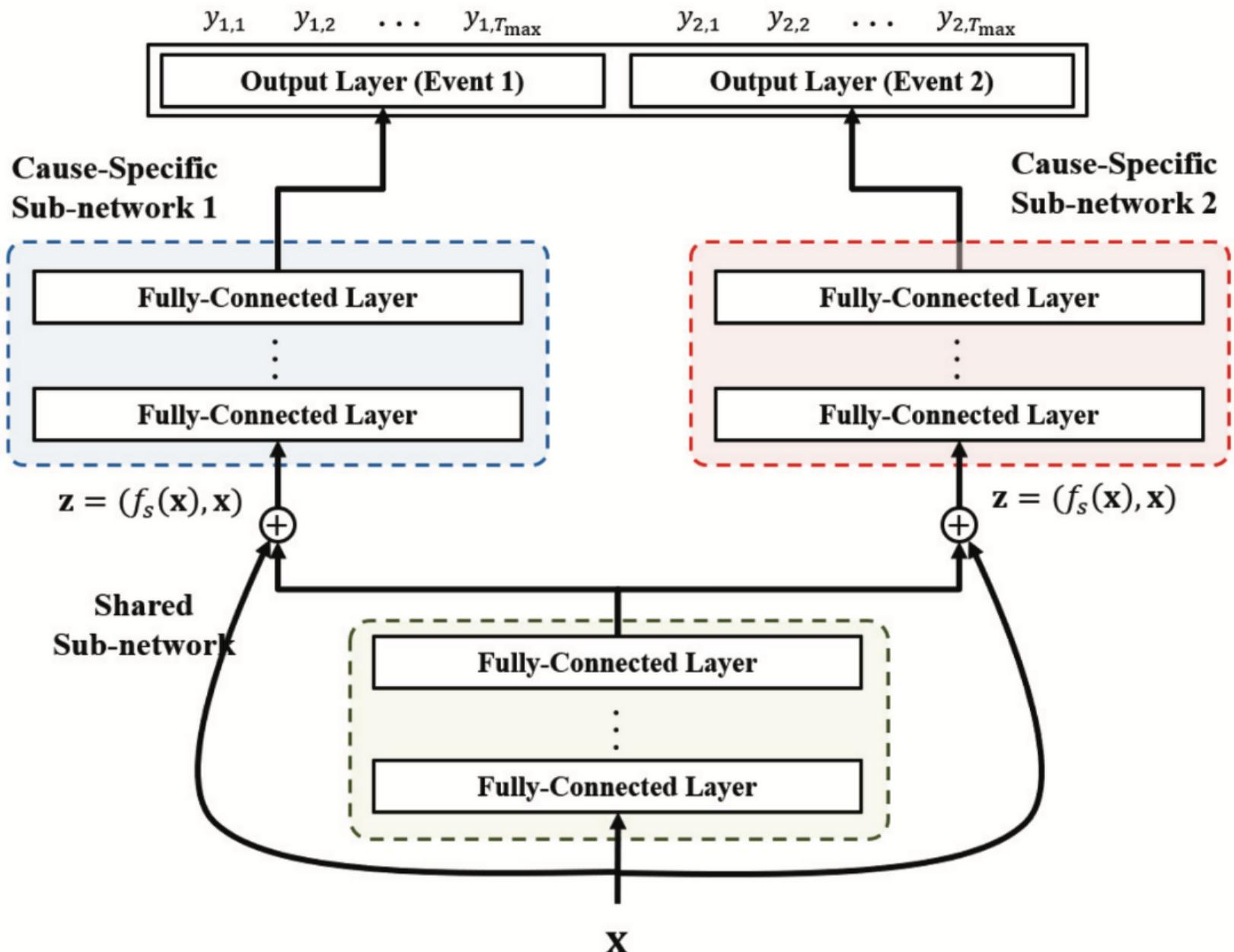
DeepHit



- 통계학 모델 기반의 생존 분석 방법은 현실과 괴리가 존재하는 가정들이 존재
- event-time의 hazard ratio가 시간과 무관하게 일정함

- 딥러닝을 활용한 생존 분석 연구 활발
- DeepHit (top), DRSA (bottom)

Output (softmax) Layer



Shared Sub-network

- x 로 표현되는 한 샘플의 feature가 MLP Layer 통과
- MLP를 통과하고 나온 output에 샘플의 원본 X 더함[Residual]

Cause-Specific Sub-networks

- Shared Sub-network를 통과하고 나온 벡터 z 가 input으로 입력
- 대상이 되는 Event의 개수만큼 Cause-specific sub network를 구성

Output Layer

- Event별 Output Layer 벡터들을 모두 이어붙이고 softmax 함수 통과
- $y_{i,j}$ 는 Event i 가 j 시점에 발생할 확률의 추정값 $P(\text{event}, \text{time}|x)$

Problem Formulation

ID	Death Cause	Survival Time (m)	Lymphnode	Age	Gender	Married	...	Benign Tumors	Malignant Tumors	Histology ICD	EOD
1	2	57	0.4061	53	1	1		0	2	65	0.0080
2	1	71	0.1382	56	1	1		0	2	64	0
3	0	135	0.1600	60	1	1	...	0	2	65	0.0620
4	0	120	0.2195	50	1	0		0	1	65	0
5	1	29	0.7998	55	1	1		0	2	64	0
6	2	71	0.7998	55	1	0		0	2	64	0

Loss - Cumulative Incidence Function

$$\begin{aligned} F_{k^*}(t^* | \mathbf{x}^*) &= P(s \leq t^*, k = k^* | \mathbf{x} = \mathbf{x}^*) \\ &= \sum_{s^*=0}^{t^*} P(s = s^*, k = k^* | \mathbf{x} = \mathbf{x}^*) \end{aligned}$$

$$\hat{F}_{k^*}(s^* | \mathbf{x}^*) = \sum_{m=0}^{s^*} y_{k,m}^*$$

Loss1 - Cumulative Incidence Function

$$\mathcal{L}_1 = - \sum_{i=1}^N \left[\mathbf{1}(k^{(i)} \neq \emptyset) \cdot \log \left(y_{k^{(i)}, s^{(i)}}^{(i)} \right) + \mathbf{1}(k^{(i)} = \emptyset) \cdot \log \left(1 - \sum_{k=1}^K \hat{F}_{k^{(i)}}(s^{(i)} | \mathbf{x}^{(i)}) \right) \right]$$

- 1 : Indicator function으로 x가 True일때 1, False일 때 0을 반환
- 앞부분 : i번째 샘플이 이벤트 k가 발생한 경우 $\log(y)$ 를 반환
 - 로그값은 1에 가까울 수록 0에 가까운 음수를 반환하며, Loss에 -가 붙어있으므로, 결국 output layer를 통과하고 나온 softmax 확률값이 클수록 loss 함수가 작아짐 => 즉 이벤트가 발생하는 시간을 잘 맞출수록 loss 함수 감소
- 뒷부분 : i번째 샘플이 이벤트 k가 발생하지 않은 경우(censoring), s 시점까지 어떤 이벤트도 발생하지 않을 $\log(\text{확률값})$ 을 의미 => censoring 데이터에 대하여, 관측 시점 이전까지 아무 이벤트도 발생하지 않는 것을 예측하도록 설계된 loss

Loss2 - Ranking

$$\mathcal{L}_2 = \sum_{k=1}^K \alpha_k \cdot \sum_{i \neq j} A_{k,i,j} \cdot \eta \left(\hat{F}_{k^{(i)}}(s^{(i)} | \mathbf{x}^{(i)}), \hat{F}_{k^{(j)}}(s^{(i)} | \mathbf{x}^{(j)}) \right)$$

where $A_{k,i,j} \triangleq \mathbf{1}(k^{(i)} = k, s^{(i)} < s^{(j)})$, $\eta(x, y) = \exp\left(\frac{-(x-y)}{\sigma}\right)$

- Event 발생 시점이 다른 두 샘플로부터 Event 발생 순서를 맞히는 Loss
- α 는 Event 별로의 가중치 : 여러 Event가 존재하는 competing risk의 상황에서 사용하기 위함
- A 는 Event가 발생한 두 i, j 번째 샘플에 대하여, i 번째 샘플이 j 번째 샘플보다 Event k 가 먼저 발생했을 때 1을 반환하는 함수
 - 계산과정에서 permutation이 아닌 combination을 찾기 위함
- $\eta(x, y)$ 는 두 input x, y 에 대하여 $x-y$ 값이 크면 클수록 작아지는 함수
- 결론적으로 해당 loss function은 i 번째 샘플이 j 번째 샘플보다 Event k 가 먼저 발생했을 때에 1을 반환하고, 두 샘플의 CIF추정치의 차이가 클수록 작아지는 양상 => 즉, 모든 샘플 쌍들의 순서를 맞히는데 CIF 추정치의 차이가 최대한 커지도록 설계

05

DRSA

Survival Function

$$\begin{aligned}
 S(t|x^i; \theta) &= \Pr(t < z|x^i; \theta) \\
 &= \Pr(z \notin V_1, z \notin V_2, \dots, z \notin V_{l^i}|x^i; \theta) \\
 &= \Pr(z \notin V_1|x^i; \theta) \cdot \Pr(z \notin V_2|z \notin V_1, x^i; \theta) \cdots \\
 &\quad \cdot \Pr(z \notin V_{l^i}|z \notin V_1, \dots, z \notin V_{l^{i-1}}, x^i; \theta) \\
 &= \prod_{l:l \leq l^i} \left[1 - \Pr(z \in V_l | z > t_{l-1}, x^i; \theta) \right] \\
 &= \prod_{l:l \leq l^i} (1 - h_l^i),
 \end{aligned}$$

$$S(\tau|x) = \Pr(T > \tau | X = x) = \prod_{t \leq \tau} (1 - h(t|x))$$

Event Density Function

$$\begin{aligned}
 p_T(\tau|x) &= \Pr(T = \tau | X = x) \\
 p_C(\tau|x) &= \Pr(C = \tau | X = x)
 \end{aligned}$$

Hazard Function

$$\begin{aligned}
 h(\tau|x) &= \frac{p_T(\tau|x)}{S(\tau-1|x)} \\
 p_T(\tau|x) &= h(\tau|x) \cdot S(\tau-1|x) = h(\tau|x) \cdot \prod_{t \leq \tau-1} (1 - h(t|x))
 \end{aligned}$$

06

Metric



C-index (Corcordance index)

- 대상의 정확한 생존시간을 평가하지 않고 대신 여러 대상의 생존시간을 상대적으로 비교
- 사망순서를 잘 예측하는지 판단

Time dependent C-index (Performance Metric)

- C-index를 개량한 지표로 전체시간에 대한 평가가 아닌 각 대상마다의 사건 발생 시간을 기준으로 탐색 범위를 제한하여 시간적 요소를 반영
- 1.0에 가까울수록 정확히 예측, 0.5에 가까울수록 무작위로 예측
- 사건 발생 확률을 얼마나 잘 반영하는지 뿐만 아니라 생존시간도 같이고 려할 수 있는 평가지표

$$\hat{R}(\tau^i|x^i) = 1 - \prod_{i=0}^{\tau} (1 - h(\tau^i|x^i))$$

$$c(\tau) = P(\hat{R}(t^i|x^i) > \hat{R}(t^j|x^j) | \tau^i < \tau^j, \tau^i \leq t)$$

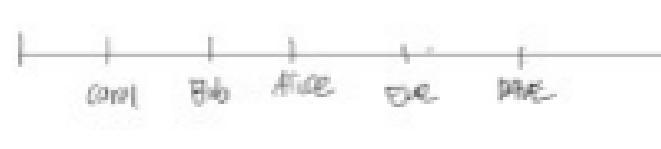


C-index : 1.0
수행 가능한 조건 개수 count time = 1 조건 1개

C-index 22.12-07-04



C-index : 0.0



C-index : $\frac{6}{10} = 0.6$

- (Alice, Bob)
- (Bob, Carol)
- (Carol, DAVE)
- (Dave, Eve)
- (Alice, Carol)
- (Bob, DAVE)
- (Carol, Eve)
- (Alice, DAVE)
- (Bob, Eve)
- (Alice, Eve)



C-index = 9.5

- (Alice, Bob)
- (Bob, Carol)
- (Carol, DAVE)
- (DAVE, Eve)
- (Alice, Carol)
- (Bob, DAVE)
- (Carol, Eve)
- (Alice, DAVE)
- (Bob, Eve)
- (Alice, Eve)

0.5*



Carol의 경우 가능한 조건은 3개인 경우 3개의 조건 + 2개
수행 가능한 조건이므로 C-index : $\frac{5}{7}$

- (Alice, Bob)
- ~~(Bob, Carol)~~
- ~~(Alice, Carol)~~
- (Bob, DAVE)
- (Carol, DAVE)
- (Carol, Eve)
- (Alice, DAVE)
- (Bob, Eve)
- (Alice, Eve)

즉 수행 가능한 조건은 5개

과제

1. Deep Recurrent Survival Analysis를 읽고 deephit과 어떤 차이가 있는지
A4 1/2 분량으로 적기
 - a. 힌트) hazard function과 cumulative incidence function
2. 오늘 배웠던 기술로 적용해볼만한 분야 혹은 데이터가 있다면 어떤 것일지 자신이
기존의 관심있던 도메인에 적용해서 생각해보기 (A4 1/2 분량)

**THANK
YOU**