Deep Recurrent Survival Analysis

Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang,
Weinan Zhang, Lin Qiu, Yong Yu

Apex Data & Knowledge Management Lab

Shanghai Jiao Tong University

Table of Contents

Background

- Time-to-event data analysis
  - The *probability* of the event over time.
  - May have different meanings in different areas.

Area	Time	Event	Event Probability
Medicine Research	Survival time	Disease	Survival rate
Information System	Duration time	Next visit	Visiting rate
Second-price Auction	Bid price	Winning the auction	Losing rate

Survival Analysis (SA)

- Survival Analysis
  - To analyze the *expected duration* of time until one or more events happen.

Task of SA

- Given the feature of the sample, forecast
  - the probability of event *happening* at each time: $\mathrm{p}(z)$
  - the probability of event *happened* at that time: $\mathrm{W}(t)$
  - the probability of event *not happened* at the time: $S(t)$

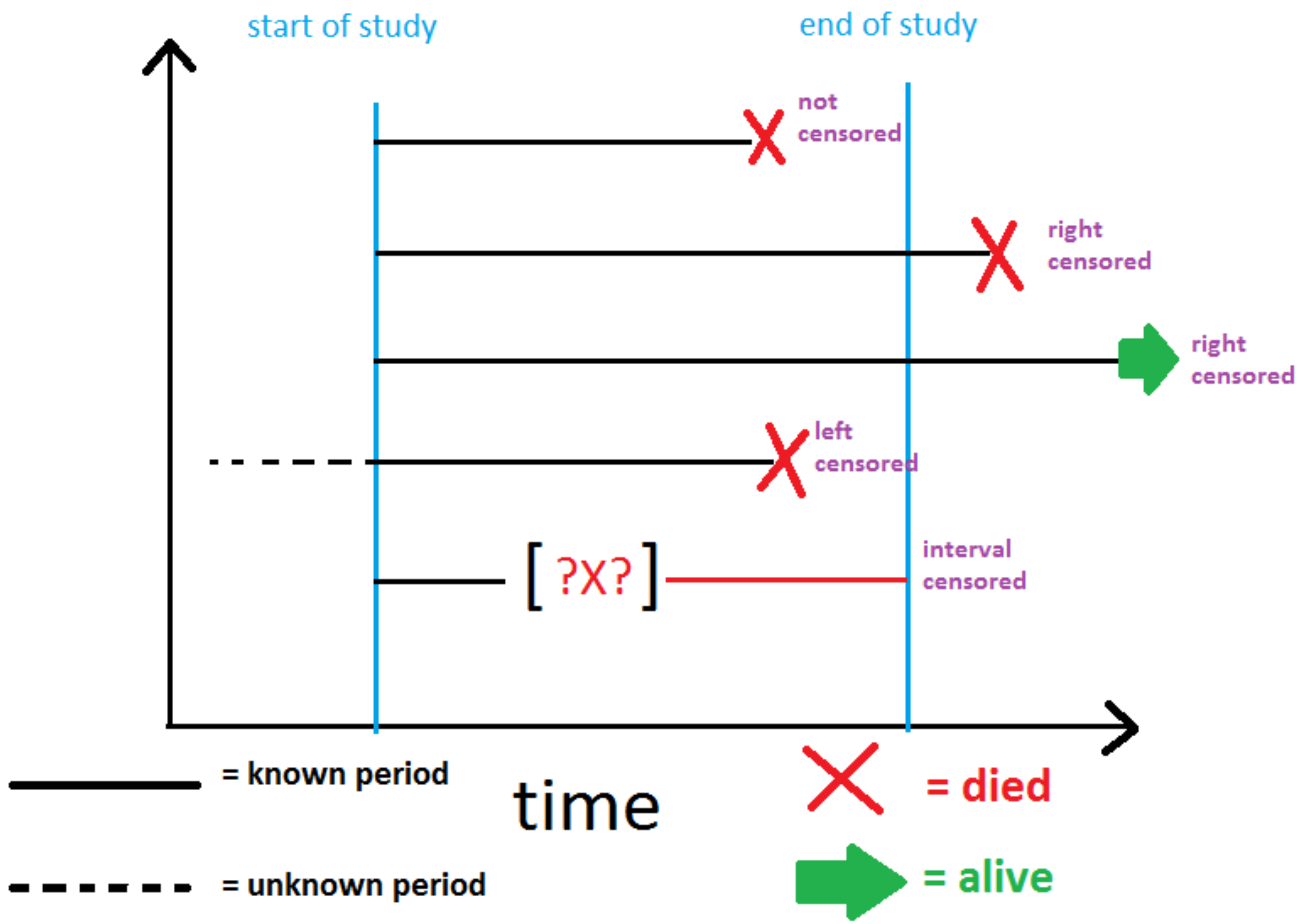
- 2 goals
  - Probability density function (P.D.F.) of the event prob. over time.
  - Cumulative distribution function (C.D.F.) of the event *at the time*.

- 2 relationships between the three prob. functions
  - Event Rate: $W(t) = \int_0^t p(z)dz$
  - Survival Rate: $\mathrm{S}(t) = \int_t^\infty p(z)dz = 1 - W(t)$

Challenges in SA

- No ground truth
  - For the **form** of the event probability distribution
  - For the **value** of the event probability
- Sparsity
  - Event is sparse, rare to happen
- Censorship
  - Some clues are censored (without the true event time)

Censorship



http://www.karlin.mff.cuni.cz/~pesta/NMFM404/survival.html

APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Censorship (cont.)

- For the *censored* samples:
- Observing time t
- True event time z is unknown
- Only knows that
  - $\boxed{\text{Right censored: } t < z}$
  - Left censored: $t > z$
  - Interval censored: $z \in [t_1, t_2]$

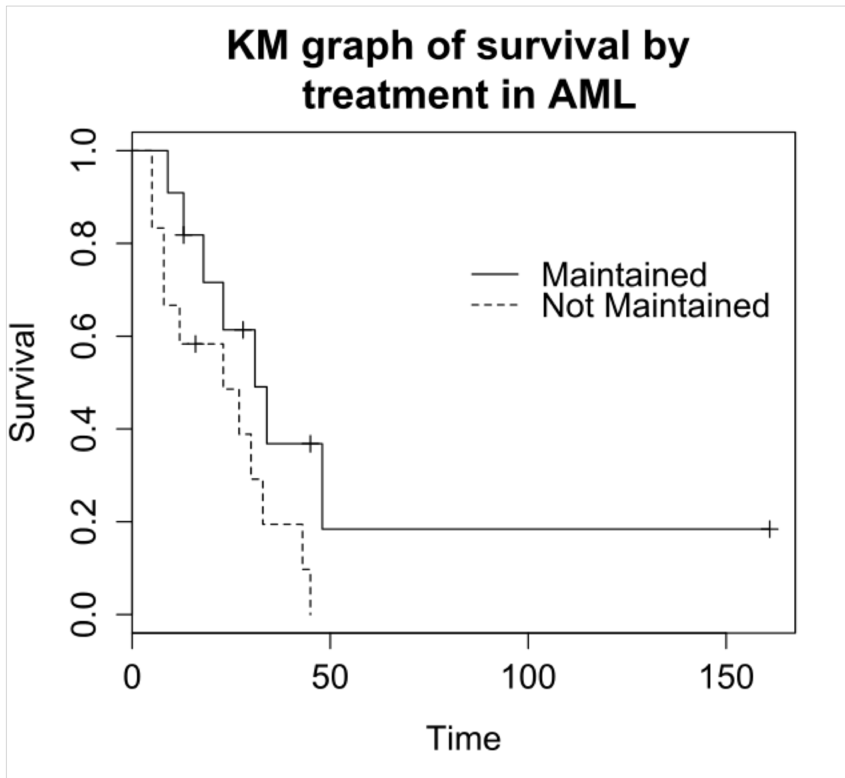
APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Task Formulation

- Data format
  - $\{(\boldsymbol{x}, t, z)\}_1^N$
  - $\boldsymbol{x}$: sample feature
  - t: observing time
  - z: true event time
    - z is known for uncensored data ($t > z$);
    - z is unknown for censored data ($t < z$).
- Input:
  - Sample features $\boldsymbol{x}$
- Output
  - P.D.F. of event probability $p_z(z)$
  - C.D.F. of event rate $W(t)$ & survival rate $S(t) = 1 - W(t)$

Existing Methods

- Statistical methods
  - Kaplan-Meier method
  - Coarse-grained, counting-based, low generalization



Kaplan and Meier 1958.

Existing Methods (cont.)

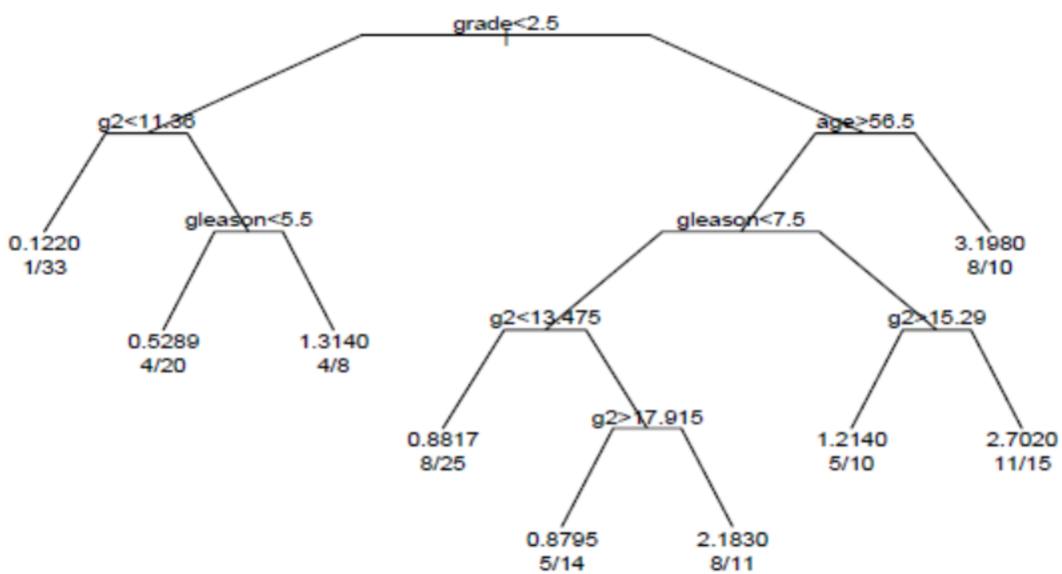
- Statistical methods
  - Cox proportional hazard (CPH) model
  - Hazard function
    - The probability of event occurring at time t *given not occurred before*.
    - $\lambda(t|x) = \lambda_0(t)e^{\beta x}$
    - The base hazard function has some assumptions, e.g., Weibull distribution.
    - Drawback: not flexible in practice.

Cox 1992; Zhang and Lu 2007.

APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Existing Methods (cont.)

- Machine learning methods
  - Survival tree model
  - Drawback:
    - based on segmented data
    - coarse-grained



Wang et al. 2016.

Existing Methods (cont.)

- Deep learning method
  - DeepSurv[1]
    - bases on CPH method using deep learning as enhanced feature extraction.
  - DeepHit[2]
    - directly predicts $p(z)$ at each time
    - calculates $S(t)$ by summing $p(z)$ over $[1, t]$

1. Katzman et al. 2018; 2. Lee et al. 2018.

Cons of the Existing Methods

- Statistical methods
  - Counting-based statistics, loss of generality
    - Kaplan-Meier
  - Specific form of the probability distribution
    - CPH, Lasso-cox
- Machine learning methods
  - Based on segmented data, too coarse-grained
    - Survival Trees
  - Assumption of the specific form of distribution
    - DeepSurv

- No consideration about sequential patterns over time!

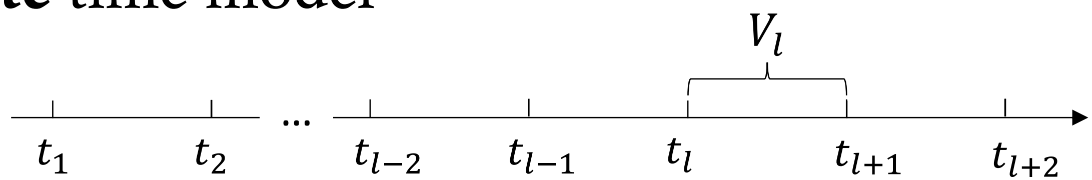
Deep Recurrent Survival Analysis (DRSA)

- No assumption about distributional forms
- Captures sequential patterns in the feature-time space
- First work ever, utilizes auto-regressive model for SA
- Handling censorship with unbiased learning
- Significant improvement against both stat. methods and ML methods

APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Our method

- **Discrete** time model



  - $z \in V_l$ means event occurs at time l
  - $z \notin V_l$ means event *not* occurs at time l
- **Hazard** rate function, means the event probability at that time *given not happened before*.
- $h_l = \Pr(z \in V_l | z > t_{l-1}, \boldsymbol{x}; \boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}, t_l | \boldsymbol{r}_{l-1})$
- Use the recurrent cell f_θ to model cond. probability h_l
  - r_{l-1} is the transmitted information through time
  - x^i, t_l are the input to the unit

Relationships among Probability Functions

- $S(t_l|\boldsymbol{x};\boldsymbol{\theta})$

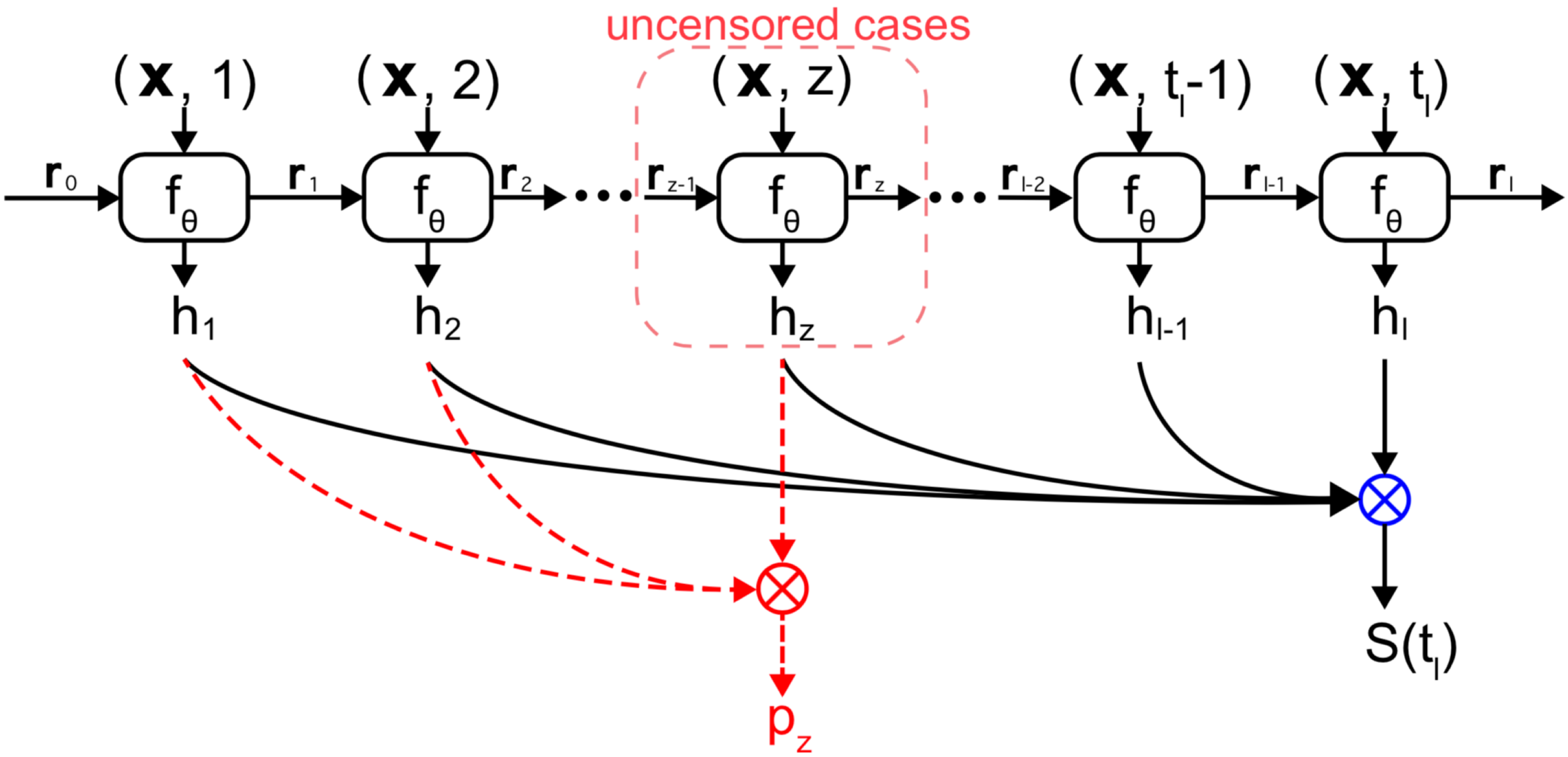
$$= \Pr(t_l < z|\boldsymbol{x};\boldsymbol{\theta})$$
$$= \boxed{\begin{aligned} &\Pr(z \notin V_1, z \notin V_2, \dots, z \notin V_l|\boldsymbol{x};\boldsymbol{\theta}) \\ = &\Pr(z \notin V_1|\boldsymbol{x};\boldsymbol{\theta}) \cdot \Pr(z \notin V_2|z \notin V_1, \boldsymbol{x};\boldsymbol{\theta}) \cdots \\ &\cdot \Pr(z \notin V_l|z \notin V_1, \dots, z \notin V_{l-1}, \boldsymbol{x};\boldsymbol{\theta}) \end{aligned}}$$
$$= \prod_{k:k\le l} [1 - \Pr(z \in V_k|z > t_{k-1}, \boldsymbol{x};\boldsymbol{\theta})]$$
$$= \prod_{k:k\le l} (1 - h_k)$$

Probability chain rule

$P(e_1, e_2, e_3) = P(e_3|e_1, e_2)P(e_2|e_1)P(e_1)$

- $W(t_l|\boldsymbol{x};\boldsymbol{\theta}) = 1 - S(t|\boldsymbol{x};\boldsymbol{\theta})$ =1- $\prod_{k:k\le l}(1 - h_k)$
- $p_l = \Pr(z \in V_l|\boldsymbol{x};\boldsymbol{\theta}) = h_l \prod_{k:k<l}(1 - h_k)$

The Recurrent Model



Loss Functions (1/3)

- Uncensored data
  - P.D.F. loss on the true event time *z*
  - Maximize the log likelihood

$$
\begin{aligned}
L_z &= -\log \prod_{(\boldsymbol{x}^i, z^i) \in \mathbb{D}_{\text{uncensored}}} \Pr(z^i \in V_{l^i} | \boldsymbol{x}^i; \boldsymbol{\theta}) \\
&= -\log \prod_{(\boldsymbol{x}^i, z^i) \in \mathbb{D}_{\text{uncensored}}} p_l^i \\
&= -\log \prod_{(\boldsymbol{x}^i, z^i) \in \mathbb{D}_{\text{uncensored}}} h_{l^i}^i \prod_{l: l < l^i} (1 - h_l^i) \\
&= - \sum_{(\boldsymbol{x}^i, z^i) \in \mathbb{D}_{\text{uncensored}}} \left[\log h_{l^i}^i + \sum_{l: l < l^i} \log(1 - h_l^i) \right]
\end{aligned}
$$

Loss Functions (2/3)

- Uncensored data ($z < t$)
  - C.D.F. loss on the observing time t
  - Maximize the <u>log *partial* likelihood</u>

$$
\begin{aligned}
L_{\text{uncensored}} &= -\log \prod_{(\boldsymbol{x}^i, t^i) \in \mathbf{D}_{\text{uncensored}}} \Pr(t^i \geq z \mid \boldsymbol{x}^i; \boldsymbol{\theta}) \\
&= -\log \prod_{(\boldsymbol{x}^i, t^i) \in \mathbb{D}_{\text{uncensored}}} W(t^i \mid \boldsymbol{x}^i; \boldsymbol{\theta}) \\
&= - \sum_{(\boldsymbol{x}^i, t^i) \in \mathbb{D}_{\text{uncensored}}} \log \left[1 - \prod_{l : l \leq l^i} (1 - h_l^i) \right]
\end{aligned}
$$

Loss Functions (3/3)

- Censored data (z is unknown since $z > t$)
  - C.D.F. loss on the observing time t
  - Maximize the log *partial* likelihood
  - Unbiased learning

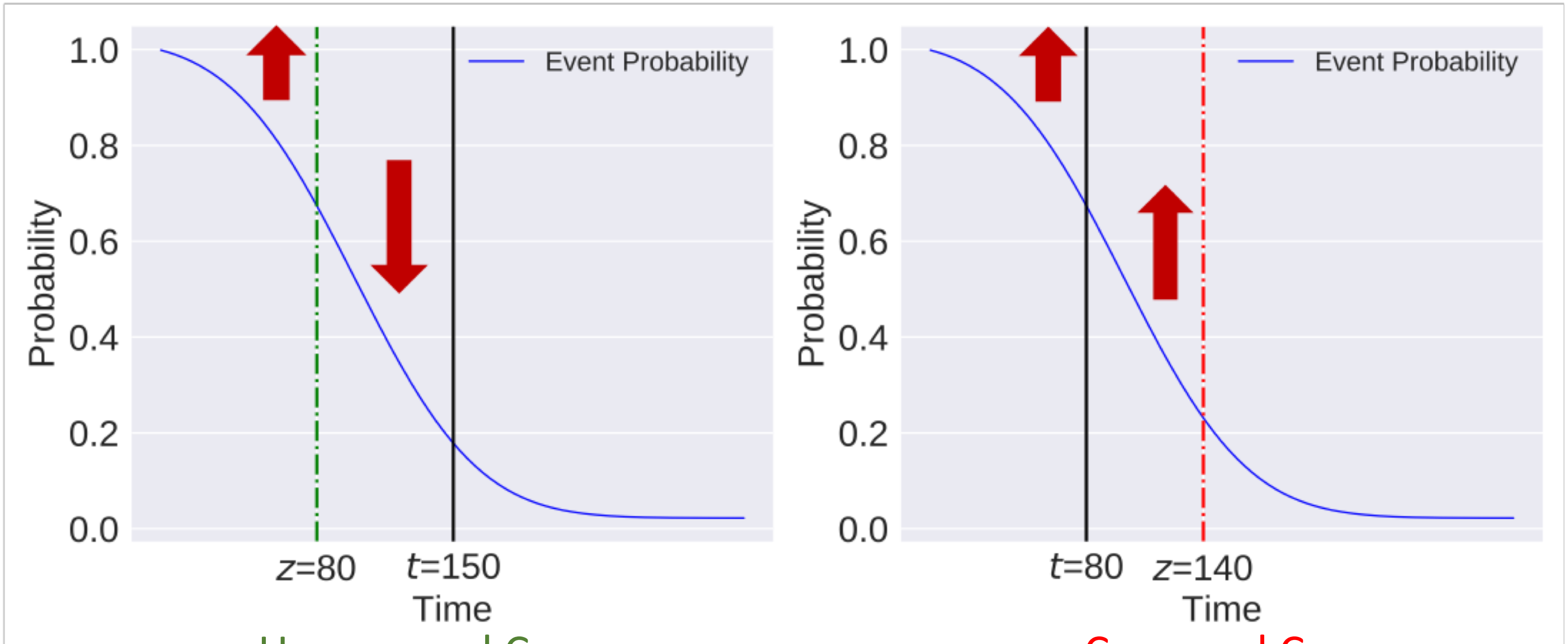
$$
\begin{aligned}
L_{\text{censored}} &= -\log \prod_{(\boldsymbol{x}^i, t^i) \in \mathbb{D}_{\text{censored}}} \Pr(z > t^i | \boldsymbol{x}^i; \boldsymbol{\theta}) \\
&= -\log \prod_{(\boldsymbol{x}^i, t^i) \in \mathbb{D}_{\text{censored}}} S(t^i | \boldsymbol{x}^i; \boldsymbol{\theta}) \\
&= -\sum_{(\boldsymbol{x}^i, t^i) \in \mathbb{D}_{\text{censored}}} \sum_{l : l \le l^i} \log(1 - h^i_l) \,.
\end{aligned}
$$

Loss Functions (cont.)

- Three losses

$$L = \underbrace{L_z}_{\text{P.D.F. Loss}} + \underbrace{\overbrace{L_{uncensored}}^{\text{Uncensored Data}} + \overbrace{L_{censored}}^{\text{Censored Data}}}_{\text{C.D.F. Loss}}$$

Intuition behind C.D.F. Losses



Uncensored Case
(z has been known)

Censored Case
(z is unknown)

- We need to
  - Push down ↓ the survival curve $S(t)$ when
    - event occurred before t, i.e., $z < t$ for uncensored data.
  - Pull up ↑ the survival curve $S(t)$ when
    - event not occurs before t, i.e., $z > t$ for censored data.

APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Experiments

- 3 real-world large-scale datasets

- 2 evaluation metrics

- 6 compared baseline models

Datasets

- 3 real-world large-scale datasets
  - Download link of the processed data:
  - https://goo.gl/nUFND4.

- CLINIC from medicine research
- MUSIC from information systems
- BIDDING from economics

Dataset	Total #	Censored Data #	Censored Rate	AET ($\mathbb{D}_{\text{full}}$)	AET ($\mathbb{D}_{\text{uncensored}}$)	AET ($\mathbb{D}_{\text{censored}}$)	Feature #
CLINIC	6,036	797	0.1320	9.1141	5.3319	33.9762	14
MUSIC	3,296,328	1,157,572	0.3511	122.1709	105.2404	153.4522	6
BIDDING	19,495,974	14,848,243	0.7616	82.0744	25.0484	99.9244	12

Evaluation Metrics

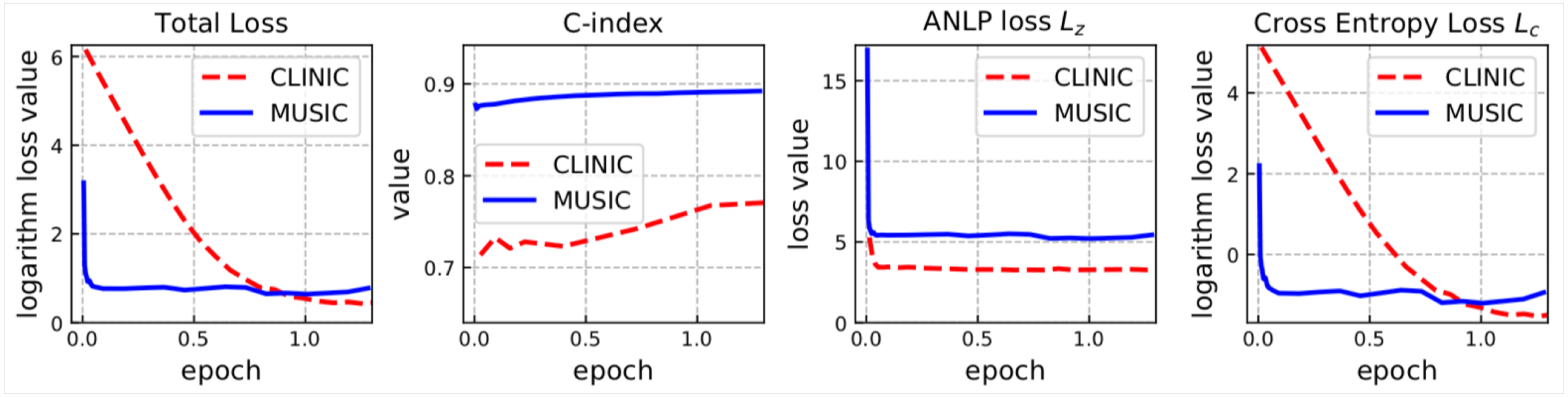
- ANLP
  - Averaged negative log probability
    - of the true event time z
- C-index
  - Time-dependent concordance index
  - measures the ranking performance of the censorship prediction at the given time.
  - The same as Area under ROC Curve in IR

Experiment Results

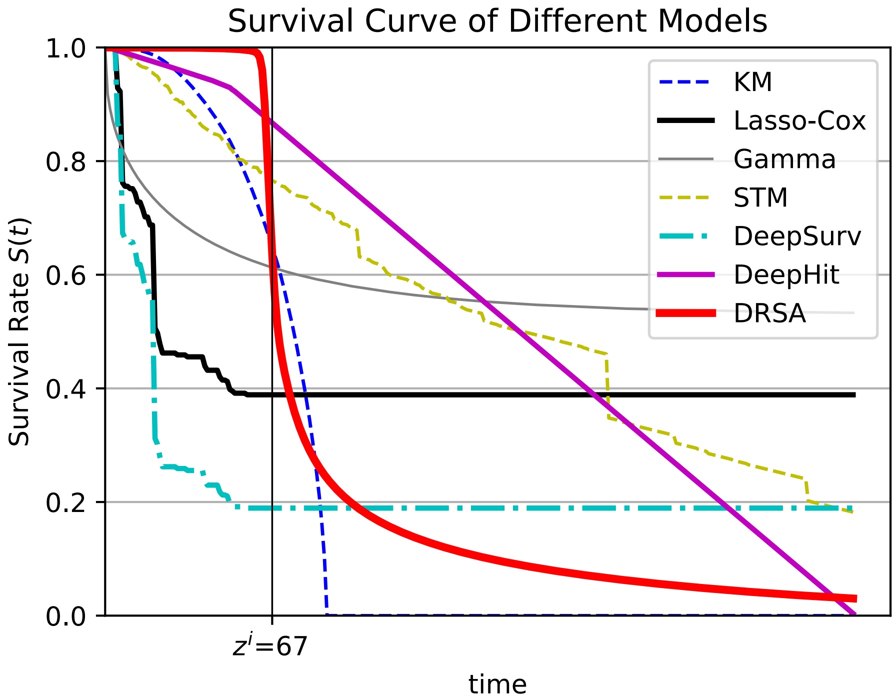
Models	C-index			ANLP		
	CLINIC	MUSIC	BIDDING	CLINIC	MUSIC	BIDDING
KM	0.710	0.877	0.700	9.012	7.270	15.366
Lasso-Cox	0.752	0.868	0.834	5.307	28.983	38.620
Gamma	0.515	0.772	0.703	4.610	6.326	6.310
STM	0.520	0.875	0.807	3.780	5.707	5.148
MTLSA	0.643	0.509	0.513	17.759	25.121	9.668
DeepSurv	0.753	0.862	0.840	5.345	29.002	39.096
DeepHit	0.733	0.878	0.858	5.027	5.523	5.544
DRN	0.765	0.881	0.823	3.441	5.412	12.255
DRSA	**0.774***	**0.892***	**0.911***	**3.337***	**5.132***	**4.774***

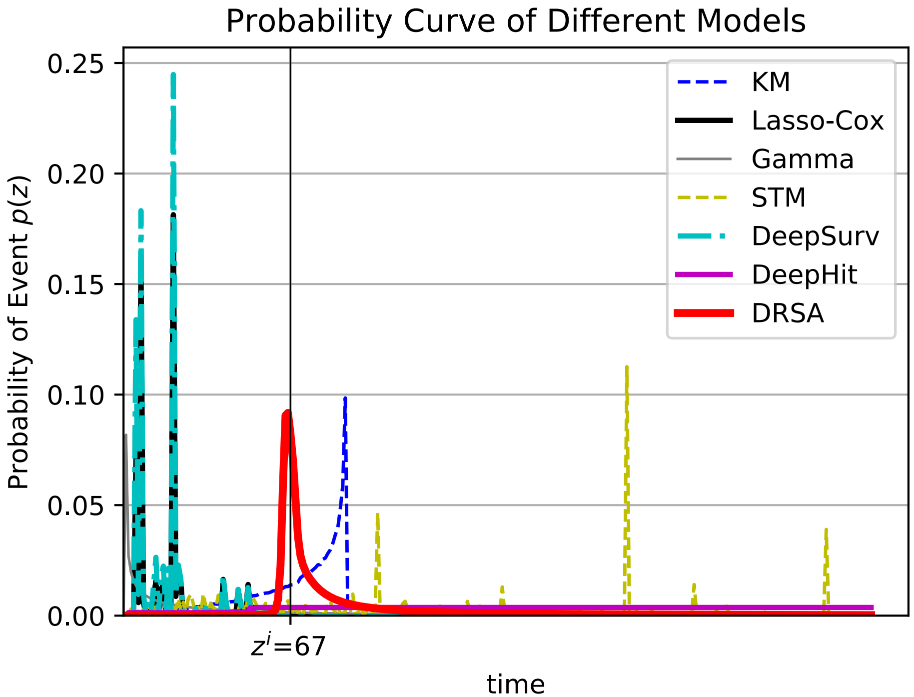
Performance comparison on C-index (the higher, the better) and ANLP (the lower, the better). (* indicates p- value < 10–6 in significance test)

Learning Curves



Survival Curves





APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Conclusion

- Thank you for attention!
- We argued that, in survival analysis,
  - Sequential patterns over time should be considered.
  - More supervision over $[z, t]$ should be made.
- We proposed
  - 1st work using auto-regressive model for survival analysis.
- DRSA (https://github.com/rk2900/drsa)
  - Utilizes recurrent neural cell predicting the conditional hazard rate;
  - Estimates the true event ratio and survival rate through probability chain rule;
  - Achieves significant improvements against strong baselines.