Nearest Neighbor and Kernel Survival Analysis

Nonasymptotic Error Bounds and Strong Consistency Rates

George H. Chen
Assistant Professor of Information Systems
Carnegie Mellon University
Survival Analysis

<table>
<thead>
<tr>
<th>Feature</th>
<th>Gluten allergy</th>
<th>Immunosuppressant</th>
<th>Low resting heart rate</th>
<th>Irregular heart beat</th>
<th>High BMI</th>
<th>Time of death</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 2</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Day 10</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Day ≥ 6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Feature vector \( x \)

Observed time \( Y \)

When we stop collecting training data, not everyone has died!

Goal: Estimate \( S(t| x) = \mathbb{P}(\text{survive beyond time } t \mid \text{feature vector } x) \)
**Problem Setup**

**Model:** Generate data point \((X, Y, \delta)\) as follows:

1. Sample feature vector \(X \sim \mathbb{P}_X\)
2. Sample time of death \(T \sim \mathbb{P}_{T|X}\)
3. Sample time of censoring \(C \sim \mathbb{P}_{C|X}\)
4. If death happens before censoring \((T \leq C)\): Set \(Y = T, \delta = 1\) 
   Otherwise: Set \(Y = C, \delta = 0\)

**Estimator (Beran 1981):**

- Find \(k\) training points closest to \(x\)
- \(k\) data points
- Kaplan-Meier estimator
- \(\hat{S}(t | x)\)

**Error:** 
\[
\sup_{t \in [0, \tau]} |\hat{S}(t | x) - S(t | x)| 
\]  
for time horizon \(\tau\)

---

**Feature space is separable metric space**
- (intrinsic dimension \(d\))
- Smooth w.r.t. feature space (Hölder index \(\alpha\))
- Borel prob. measure
- Continuous r.v. in time &
- Feature space is separable metric space

---

**Enough of the \(n\) training data have \(Y\) values > \(\tau\)**
Theory (Informal)

$k$-NN estimator with $k = \tilde{\Theta}(n^{2\alpha/(2\alpha+d)})$ has strong consistency rate:

$$\sup_{t \in [0, \tau]} |\hat{S}(t|x) - S(t|x)| \leq \tilde{O}(n^{-\alpha/(2\alpha+d)})$$

If no censoring, problem reduces to conditional CDF estimation

$\rightarrow$ Error upper bound, up to a log factor, matches conditional CDF estimation lower bound by [Chagny & Roche 2014](#).

Proof ideas also give finite sample rates for:

- Kernel Kaplan-Meier estimators
- $k$-NN & kernel Nelson-Aalen *cumulative hazard* estimators ($-\log S(t \mid x)$)
- Generalization bound for automatic $k$ using validation data

Most general finite sample theory for $k$-NN and kernel survival estimators

Existing kernel results only for Euclidean space ([Dabrowska 1989](#), [Van Keilegom & Veraverbeke 1996](#), [Van Keilegom 1998](#))
Experiments

Dataset "gbsg2" Concordance Indices

- Distance/kernel choice matters a lot in practice
- Learning the kernel typically has best performance (but no theory yet!)

<table>
<thead>
<tr>
<th>Method</th>
<th>c-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>cox</td>
<td></td>
</tr>
<tr>
<td>k-NN L2</td>
<td></td>
</tr>
<tr>
<td>k-NN L1</td>
<td></td>
</tr>
<tr>
<td>k-NN (triangle) L2</td>
<td></td>
</tr>
<tr>
<td>k-NN (triangle) L1</td>
<td></td>
</tr>
<tr>
<td>kernel (box) L2</td>
<td></td>
</tr>
<tr>
<td>kernel (box) L1</td>
<td></td>
</tr>
<tr>
<td>kernel (triangle) L2</td>
<td></td>
</tr>
<tr>
<td>kernel (triangle) L1</td>
<td></td>
</tr>
<tr>
<td>random survival forest</td>
<td></td>
</tr>
<tr>
<td>adaptive kernel</td>
<td></td>
</tr>
</tbody>
</table>