Nearest Neighbor and Kernel Survival Analysis: Nonasymptotic Error Bounds and Strong Consistency Rates

George H. Chen

Introduction

Survival analysis: Reason about time durations until a critical event happens, such as in:
- Health care: time until death, or time until a disease relapses
- Manufacturing: time until a device fails
- Criminology: time until a convicted criminal reoffends

Critical event need not be death but we use this as the running example

Illustrative Example Data

<table>
<thead>
<tr>
<th>Gluta</th>
<th>Immuno</th>
<th>Low rest</th>
<th>Irreg</th>
<th>High BMI</th>
<th>Time of death</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Day 2</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Observed time Y</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Day 10</td>
</tr>
</tbody>
</table>

Goal: Estimate $S(t|x) = P(\text{survive beyond time } t | \text{ feature vector } x)$

This paper analyzes $k$-NN and kernel Kaplan-Meier estimators (Beran 1981) for $S(t|x)$

Contributions
- Most general finite sample error upper bounds for $k$-NN and kernel survival estimators (nearly optimal)
- Experimental evidence that learning a kernel with random survival forests (Ishwaran et al. 2006) often works well in practice (theory for this adaptive kernel variant remains an open problem)

Problem Setup

Model: Generate each point $(X, Y, \delta)$ i.i.d.
1. Sample feature vector $X \sim P_X$
2. Sample time of death $T \sim P_T|X$
3. Sample time of censoring $C \sim P_C|X$
4. If death is before censoring ($T \leq C$):
   - Set $Y = T, \delta = 1$
   - Otherwise:
   - Set $Y = C, \delta = 0$

Goal: Estimate $\hat{S}(t|x) = P(T > t | X = x)$

$k$-NN Estimator (Beran 1981)

For given data points, estimate $P(\text{survive beyond time } t)$:
1. Sort unique times of death: $t_1 < t_2 < \cdots < t_L$
2. Construct the following table:
   
   $\begin{array}{cccccc}
   t_1 & t_2 & \cdots & t_L \\
   d_1 & d_2 & \cdots & d_L \\
   n_1 & n_2 & \cdots & n_L \\
   \end{array}$

3. Compute the following estimate (blue function):

$\hat{S}(t|x) = \prod_{i=1}^L \left(1 - \frac{d_i}{n_i} \right)$

Kaplan-Meier Estimator (Kaplan & Meier 1958)

For given data points, estimate $P(\text{survive beyond time } t)$:
1. Sort unique times of death: $t_1 < t_2 < \cdots < t_L$
2. Construct the following table:
   
   $\begin{array}{cccccc}
   t_1 & t_2 & \cdots & t_L \\
   d_1 & d_2 & \cdots & d_L \\
   n_1 & n_2 & \cdots & n_L \\
   \end{array}$

3. Compute the following estimate (blue function):

$\hat{S}(t|x) = \prod_{i=1}^L \left(1 - \frac{d_i}{n_i} \right)$

Theory

Focus on sup-norm error: $\sup_{t \in [0, \tau]} |\hat{S}(t|x) - S(t|x)|$ for time horizon $\tau$

Assumptions
- Probability of seeing $Y$ values above $\tau$ is large enough: $\exists \epsilon \in (0, 1/2)$ s.t. $P(Y > \tau | X = x) \geq \theta \forall x$
- Training set size $n$ large enough, enough $Y$ values exceed $\tau$
- Conditional distributions $P_T|X \, \& \, P_C|X$ are continuous r.v.'s
- Ensures no ties in when people die and that these conditional distributions have pdf's
- Pdf's of $P_{T|X} \, \& \, P_{C|X}$ are smooth w.r.t. feature space (Hölder index $a$)
- Feature vectors close by have similar death/censoring time distributions (close enough neighbors in training data will provide useful information for prediction)
- Feature space is separable metric space, $P_X$ is Borel probability measure
- Ensures probability of balls well-defined, probability of feature vector landing in support of $P_X$ is 1

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

$\sup_{t \in [0, \tau]} |\hat{S}(t|x) - S(t|x)| \leq O(n^{-\alpha/(2a+\delta)})$ (for large enough $n$ and intrinsic dimension $d$)

Main observation: $|\log \hat{S}(t|x) - \log S(t|x)| \leq \text{error in } k$-NN CDF estimate of $P_{(T \mid X)}$ + error in $k$-NN regression problem assuming $P_{(T \mid X)}$ known

Key proof ideas:
- Controlling bias/variance for $k$-NN estimators: Chaudhuri & Dasgupta 2014
- Also: new finite sample results for $k$-NN and kernel Nelson-Aalen estimators, choosing $k$ via validation set

Experiments

Datasets

| Dataset | Description | # subjects | # dim.
|---------|-------------|------------|--------
| pbc     | primary biliary ductal cholangiocarcinoma | 276 | 17
| gbsg1   | breast cancer | 686 | 8
| adv    | advanced cancer | 1044 | 14
| kidney  | kidney cancer | 1044 | 8

Accuracy: concordance index

Basic experiment
1. Randomly divide data into 70%/30% train/test pieces
2. Select alg. parameter(s) via 5-fold cross val on training set
3. Evaluate on test set
4. Repeat basic experiment 10 times

Compare to:
- Cox proportional hazards
- Random survival forest (RSF)
- Kernel survival estimator with kernel learned using RSF ("adaptive kernel")

Distance/kernel choice matters a lot in practice
- Adapting these tends to work best