Predictive Modeling of Independently Censored Failure Time Data

Kristiaan Pelckmans Johan Suykens

March 13, 2007

Abstract

This paper studies a learning machine for predicting the distribution of failure time of a new subject, based on past randomly censored observations. This is achieved through the introduction of an *health* function, which serves as a proxy between the subject's covariates, and the observed failure times. The corresponding risk is defined in terms of the concordance index (c-index) measuring the predictive performance and discriminative power of the health function with respect to failure times. The crux of this approach is now found in a formal way to translate an observed health value into a corresponding predictive distribution over the failure times. In order to formalize this approach, standard results in U-statistics are extended to the case involving independent and random censoring. From a methodological perspective, a convex approach for linear health functions and its extension to a nonlinear kernel version are studied. Artificial as well as real world benchmark datasets are used to prove the strengths and disadvantage of this method with respect to classical methods.

1 Introduction

The statistical modelling of failure time data constitutes one of the main research tracks in theoretical and applied statistics (see e.g. [8] for a thorough review). Research on machine learning and pattern recognition complements in many areas classical results in statistics, but takes a quite different perspective on the problem of inference. Instead of trying to uncover the generating probabilistic mechanism (distribution) behind the data, the objective here is to learn a predictive rule based on the observed samples, which will *generalize* well to future samples. This view is particularly pronounced in the research of empirical risk minimization [11, 4].

This paper attempts to integrate this view in a novel conceptual as well as practical approach for predictive modeling of failure time data, where the overall objective is to characterize as well as possible the occurence of failures of new subjects, based on a set of observed covariates. Since the

point-prediction of a failure time is in general unfeasable (or not appreciated), a more secure approach is proposed by introducing the notion of an health index which is one-to-one related with the covariates, and which captures the information regarding failure rates. The classical c-index is used to maximize the concordance between the healt index and the observed failure observations. Kendall's τ as reviewed in [9] and the closely related c-index [5] are found to be a proper choice for measuring the discriminative and predictive power of two random variates, closely related to the Mann-Whitney statistic, the AUC and Somers d statistic. The relation with the AUC permits the translation of recent advances devised in a context of ordinal regression [6] and information retrieval [1, 10] (trading precision and recall). From a theoretical perspective, the c-statistic relates straightforwardly to the case of U-statistic, a theory which was thoroughly investigated at least since Hoeffdings seminal work [7]. The paper [3] discusses those topics in a context of learning theory, and paves the path for application papers as the present.

This paper is organized as follows. Section 2 discusses the formal setup of this paper and makes an explicit connection with the theory on U-statistics. Section 3 discusses the resulting learning machine for extracting the optimal health index based on censored observations, optimization issues are discussed and a proof of concept is given based on artificial data. Section 4 reports on a real world dataset.

2 Predictive Modeling of Failure Time Data

This paper considers censored failure time data according to the following specifications. Let $T \geq 0$ be a random variable denoting the time to failure. Assume that (X, T) is a random vector with a joint distribution function F_{XT}. Here $X \subset \mathbb{R}^d$ denote the $d > 0$ covariates of the subject under study. For notational convenience, the case of ties is excluded by assumption, i.e. $P(T = T') = \sup_{\|w\|>0} P(w^T X = w^T X') = 0$. The case of occurence of ties can be handled properly as e.g. in [1]. Let $C > 0$ denote the time of censoring following F_C, which is assumed to be independent of both X and T. Now let $\delta \in \{0, 1\}$ denote a binary random variable indicating wether the subject is withdrawn from the study before its actual failure (right sensoring). If $\delta = 1$, let $Y \geq 0$ denote the time of failure. Formally

$$Y = \begin{cases} T & \text{if } \delta = 1 \\ C & \text{if } \delta = 0 \end{cases} \tag{1}$$

Let the set $\mathcal{D} = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$ denote the observed i.i.d. samples from a distribution $F_{XY\delta}$. Let for notational convenience the random variable Z denote the random triple (X, Y, δ) with i.i.d. samples $\{Z_i\}_{i=1}^n$.

The paper uses the following notations for vectors and matrices: a lower case b will denote a vector of a scalar (clear from context), and $\mathbf{b}$ will denote a matrix. In particular, the data samples $\{X_i\}_{i=1}^n \subset \mathbb{R}^d$ can be organized in a matrix $\mathbf{x} \in \mathbb{R}^{n\times d}$ as $\mathbf{x}_i = X_i^T$ for all $i = 1, \ldots, n$, whereas the variables $\{y_i\}_{i=1}^n \subset \mathbb{R}$ can be oragnized in a vector y. Furthermore, let $1_n \in \mathbb{R}^n$ and $0_n \in \mathbb{R}^n$ be a constant vector with constants respectively 1 and 0, and let $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ be the identity matrix of size n such that $\mathbf{I}_{n,ij} = I(i = j)$ for all $i, j = 1, \ldots, n$.

2.1 Case Without Censoring

First the case of no censoring is considered in some detail where $Z = (X, T)$. For notational convenience, define the kernel $q_u : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ for any $Z, Z' \in \mathcal{Z}$ as follows

$$q_u(Z, Z') = I\left((u(X) - u(X_i))\ (T - T_i) > 0\right). \tag{2}$$

The key objects of study of this paper are the health function $u : \mathbb{R}^d \to \mathbb{R}^+$, the c-index between two covariates, its empirical version and the indexed random variable $M_u \subset [0, 1]$.

Definition 1 (Health function u, CI and M_u) *Let $u : \mathbb{R}^d \to \mathbb{R}$ be an health function mapping a covariate $\in \mathbb{R}^d$ to $u(x) \in \mathbb{R}$. Let (X, T) be a random variable with joint distribution F_{XT}. The health function u is said to be concordant with respect to F_{XT} to a level*

$$\begin{aligned} CI(u; F_{XT}) &= P\left(u(X) > u(X') \;\middle|\; T > T'\right) \\ &= P\left((u(X) - u(X'))\ (T - T') > 0\right) = P(q_u(Z, Z')). \end{aligned} \tag{3}$$

Informally, an health function is concordant with failure times encoded in T if the ordering of two randomly drawn subjects is with high probability similar both in failure time as in health. This quantity can be estimated using an i.i.d. sample $\{(X_i, T_i)\}_{i=1}^n$ which is ordered *as $T_i \leq T_j$ for all $i < j$, as follows*

$$\begin{aligned} CI_n(u; \mathcal{D}) &= \frac{2}{n(n-1)} \sum_{T_i<T_j} I\left(u(X_i) < u(X_j)\right) \\ &= \frac{1}{n(n-1)} \sum_{i\neq j} I\left((u(X_j) - u(X_i))\,(T_j - T_i) > 0\right) = \frac{1}{n(n-1)} \sum_{i\neq j} q_u(Z_i, Z_j). \end{aligned} \tag{4}$$

where $I(z > 0) = 1$ if $z > 0$, and zero otherwise. The second equality holds by symmetry. The transition from $i < j$ to $i \neq j$ follows from symmetry.

$$M_u(Z; Z_1, \ldots, Z_n) = \frac{1}{n} \sum_{i=1}^n I\left((u(X) - u(X_i))\ (T - T_i) > 0\right), \tag{5}$$

which denotes the degree of concordancy of a new sample Z with n i.i.d. copies $\{Z_i\}_{i=1}^n$.

From this definitions, it follows clearly that $\mathrm{CI}(u; F_Z)$ is a U-statistic of order 2 with kernel $q_u(Z, Z')$. The c-index can be related to Kendall's τ as follows

$$\tau = E\left[\mathrm{sign}\left((u(X) - u(X_i))\ (T - T_i)\right)\right] = 2\mathrm{CI}(u; F_{XT}) - 1. \tag{6}$$

The following Lemma relates the random variable M_u and the c-index.

Lemma 1 *Let $Z, Z_1, \ldots, Z_n, \ldots$ be i.i.d. samples of a joint distribution F_Z. Let Z_{-i} denote the vector $(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$. Then*

$$CI(u; F_{XT}) = E[M_u(Z; Z_1, \ldots, Z_n)] = \frac{n-1}{n} \sum_{i=1}^{n} E[M(Z_i; Z_{-i})] \tag{7}$$

Proof: Simple expansion of the terms proof the result

$$\begin{aligned} \frac{n-1}{n} \sum_{i=1}^{n} E[M(Z_i; Z_{-i})] &= \frac{n-1}{n} \sum_{i=1}^{n} \frac{1}{n-1} \sum_{j \neq i} E\left[q_u(Z_i, Z_j)\right] \\ &= \frac{1}{n} \sum_{j=1}^{n} E\left[q_u(Z_i, Z_j)\right] \qquad (8) \\ &= E\left[q_u(Z_i, Z_j)\right]. \qquad (9) \end{aligned}$$

Now (7!!!) equals the definition of $E[M_u(Z; Z_1, \ldots, Z_n)]$, and (8!!) equals $\mathrm{CI}(u; F_{XT})$ as in eq. (4).

□

The following proposition states the essential tool which we will use for proving concentration properties of the $\mathrm{CI}(u; F_Z)$. It was described in [7] (or see e.g. [3], Appendix 1).

Proposition 1 (Decoupling of U-statistics) *The U-statistic with symmetrical kernel $q_u : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ can be decomposed as an average of a simpler estimate as follows*

$$\begin{aligned} \frac{1}{n(n-1)} \sum_{i \neq j} q_u(Z_i, Z_j) &= E_\pi \left[\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q_u(Z_{\pi(i)}, Z_{\pi(n-i)}) \right] \\ &= \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q_u(Z_{\pi(i)}, Z_{\pi(n-i)}) \qquad (10) \end{aligned}$$

where $\sum_\pi$ summarizes over all $n!$ equal likely permutations π of the set $\{1, \ldots, n\}$.

Now the following result follows straightforwardly as in [3]

Proposition 2 (Concentration) *With probability exceeding* $1-\delta < 0$ *and* $n > 1$*, the following inequality holds for any* $u \in \mathcal{H}$

$$CI(u; F_Z) \leq CI_n(u; \mathcal{D}) + E[R_n(\mathcal{H})] + \sqrt{\frac{4 \ln\left(\frac{2}{\delta}\right)}{n-1}} \tag{11}$$

where the Rademacher complexity $R_n(\mathcal{H})$ *is defined as*

$$R_n(\mathcal{H}) = E\left[\sup_{u \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i q_u(Z_i, Z_{n-i}) \right| \;\middle|\; X_1, \ldots, X_n \right] \tag{12}$$

where the expectation is taken over the iid random Rademacher variables $\{\sigma_i\}_{i=1}^n$ *with* $P(\sigma = 1) = P(\sigma = -1) = \frac{1}{2}$*. One can write as such that*

Proof: Fix a permutation π of the set $\{1, \ldots, n\}$. Using the classical result in [2], one gets immediately

$$E\left[q_u(Z, Z')\right] \leq \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q_u(Z_{\pi(i)}, Z_{\pi(n-i)}) + E[R_n(\mathcal{H})] + \sqrt{\frac{4 \ln\left(\frac{2}{\delta}\right)}{n-1}} \tag{13}$$

Since $\{q_u(Z_{\pi(i)}, Z_{\pi(n-i)})\}_{i=1}^{\lfloor n/2 \rfloor}$ contains at least $(n-1)/2$ i.i.d. samples, and $0 \leq q_u(\cdot, \cdot) \leq 1$. Observe that taking the expectation E_π over all $n!$ permutations π on both sides does not alter the inequality, and proofs the result.

□

The Rademacher complexity term can be related directly to the VC-dimension of a hypothesis class, but its use is not restricted to the case of indicator variables however. The following version of Markov's inequality gives us an important tool for practical use of the bounds

Proposition 3 (Markov's Inequality) *Application of Markov's inequality to the positive random variable* M_u *gives the following inequality for any* $m > 0$*, for any* $u \in \mathcal{H}$ *and with probability exceeding* $1 - \delta$

$$\begin{aligned} P\Big(M_u(Z; Z_1, \ldots, Z_n) \geq m\Big) &\leq \frac{CI(u; F_{XT})}{m} \\ &\leq \frac{CI_n(u; \mathcal{D}) + E[R_n(\mathcal{H})] + \sqrt{\frac{4 \ln\left(\frac{2}{\delta}\right)}{n-1}}}{m}. \end{aligned} \tag{14}$$

2.2 Independently Censored Observations

This subsection frames the previous discussion in a context of right-censored observations. Let Z denote now the tripple (U, Y, δ) with i.i.d. samples $\{(X_i, Y_i, \delta_i)\}_{i=1}^n$. Let the quantity of interest be defined as

$$q_u(Z, Z')c(Z, Z') = I\left((T - T')(u(X) - u(X')) > 0\right) c(Z, Z'), \tag{15}$$

where the symmetric term $c(Z, Z') = c(Z', Z) \in \{0, 1\}$ denotes wether the couple (Z, Z') is comparable. In the case of right-censoring, a formal definition of this term becomes $c(Z, Z') = I(Y < Y' \wedge \delta = 1) + I(Y > Y' \wedge \delta' = 1)$ for any $Z, Z' \in \mathcal{Z}$. Furthermore, let $c(\mathcal{D}) = \sum_{i \neq j} c(Z_i, Z_j)$ denote the number of comparable terms, and let $c(Z_i; \mathcal{D}) = \sum_{j \neq i} c(Z_i, Z_j)$ denote the number of comparable terms with Z_i.

The crux of the argument is again a decomposition of U-statistics

Theorem 1 (Decomposition of Censored U-statistics) *Let $\{W_{ij} \geq 0\}_{i \neq j}$ be appropriate defined weights with $W_{ij} = 0$ if Z_i and Z_j are not comparable due to censoring. Consider the following estimator of Kendall's τ for randomly censored data*

$$CI_n^c(u; \mathcal{D}) = \frac{2}{n(n-1)} \sum_{i \neq j} q_u(Z_i, Z_j) W_{ij}. \tag{16}$$

Let n_c denote the number of censored observations, or $n_c = \sum_{i=1}^n (1 - \delta_i)$. Then for all $\epsilon > 0$, the following probabilistic inequality holds

$$P\left(\left| CI_{\pi,n}^c(u; \mathcal{D}) - CI(u, F_Z)\right| \geq \epsilon\right) \leq 2 \exp\left(-(n - 1 - n_c)\epsilon^2\right). \tag{17}$$

Proof: Fix at first a permutation π of the set $\{1, \ldots, n\}$, then define $\mathrm{CI}_{\pi,n}^c(u; \mathcal{D})$ as follows

$$\mathrm{CI}_{\pi,n}^c(u; \mathcal{D}) = \frac{1}{c_\pi(\mathcal{D})} \sum_{i=1}^{\lfloor n/2 \rfloor} q_u\left(Z_{\pi(i)}, Z_{\pi(n-i)}\right) c(Z_{\pi(i)}, Z_{\pi(n-i)}), \tag{18}$$

where $c_\pi(\mathcal{D}) = \sum_{i=1}^{\lfloor n/2 \rfloor} c(Z_{\pi(i)}, Z_{\pi(n-i)})$. Under the assumption that censoring is independent, $\{c(Z_{\pi(i)}, Z_{\pi(n-i)})\}_i \subset \{0, 1\}^{\lfloor n/2 \rfloor}$ is an independent sample containing at least $n - n_c - 1$ items where n_c denote the number of censored observations.

$$P\left(\left|\mathrm{CI}_{\pi,n}^c(u; \mathcal{D}) - \mathrm{CI}(u, F_Z)\right| \geq \epsilon\right) \leq 2 \exp\left(-(n - 1 - n_c)\epsilon^2\right), \tag{19}$$

following Hoeffding for the at least $n - 1 - n_c$ iid samples.

We now consider the expectation over any of the $n!$ possible permutaions π. The expected value of $CI^c_{\pi,n}(u;\mathcal{D})$ can be written explicitly as

$$\begin{aligned}
\mathrm{CI}^c_n(u;\mathcal{D}) &= E_\pi\left[\mathrm{CI}^c_{\pi,n}(u;\mathcal{D})\right] \\
&= \frac{1}{n!}\sum_\pi \frac{1}{c_\pi(\mathcal{D})} \sum_{i=1}^{\lfloor n/2\rfloor} q_u\left(Z_{\pi(i)}, Z_{\pi(n-i)}\right) c(Z_{\pi(i)}, Z_{\pi(n-i)}) \\
&= \frac{1}{n!}\sum_\pi \sum_{i=1}^{\lfloor n/2\rfloor} q_u\left(Z_{\pi(i)}, Z_{\pi(n-i)}\right) \frac{c(Z_{\pi(i)}, Z_{\pi(n-i)})}{c_\pi(\mathcal{D})} \\
&= \frac{1}{n(n-1)}\sum_{k\neq l} q_u\left(Z_k, Z_l\right)\left(\frac{c(Z_k,Z_l)}{(n-2)!}\sum_{\pi|(k,l)}\frac{1}{c_\pi}\right). \qquad (20)
\end{aligned}$$

Where $\sum_{\pi|(k,l)}$ summarizes over all permutations π with existing $1 \leq i \leq \lfloor n/2\rfloor$ such that $\pi(i) = k, \pi(n-i) = l$. Define $W_{ij} = \left(\frac{c(Z_k,Z_l)}{(n-2)!}\sum_{\pi|(k,l)}\frac{1}{c_\pi}\right)$ for all i, j. Note that if no censoring occurs, the term W_{ij} equals one for all i, j. If $c(Z_i, Z_j) = 0$, also the weight W_{ij} will be zero, and the result follows.

□

At this stage, it is still unclear how the weights $\{W_{ij}\}_{i\neq j}$ can be computed efficiently from the expression $W_{ij} = \left(\frac{c(Z_k,Z_l)}{(n-2)!}\sum_{\pi|(k,l)}\frac{1}{c_\pi}\right)$.

3 Learning Machine for Predictive Modeling

We consider the following mechanism for learning a predictive modeling of failure time data. The following definition defines the objective of learnig as well as a proper way to make predictions regarding failure time corresponding to a new subject.

Definition 2 (Learning Machine for Predictive Modeling) *Let $\mathcal{H} \subset \left\{u : \mathbb{R}^d \to \mathbb{R}\right\}$ be a class of plausible health functions. An empirical risk minimizer takes the following form*

$$\hat{u} = \arg\min_{u\in\mathcal{H}} \mathcal{R}_n(u;\mathcal{D}) = \arg\max_{u\in\mathcal{H}} CI_n(u). \qquad (21)$$

Now the previous theory can be used to deduce knowledge on the distribution of failure times corresponding with a new covariate X as follows. A corresponding failure time T has concordance

$$M(u(X), T; Z_1, \ldots, Z_n) = \sum_{i=1}^n I((u(X) - u(X_i))(T - T_i) > 0) \qquad (22)$$

wich occurs with inversely proportional probability, i.e.

$$\forall m > 0: \quad P(M(u(X), T; Z_1, \ldots, Z_n) > m) \leq \frac{CI(u)}{m}, \tag{23}$$

see Proposition 3.

This paper studies the following class of health functions

Definition 3 (Linear Health Function) *The class of linear health functions is defined as*

$$\mathcal{H}^l = \Big\{ u_w : \mathbb{R}^d \to \mathbb{R},\ w \in \mathbb{R}^d \ \Big|\ u_w(x) = w^T x, \quad x \in \mathbb{R}^d \Big\}, \tag{24}$$

Note that an intercept term is not relevant in this context since it does not affect the concerned ordering relation.

3.1 Pairwise Maximal Margin Machine

As minimizing CI_n directly amounts to a hard combinatorial problem, a convex upperbound in terms of the Hinge loss is used as classical. The Hinge loss of a term $e \in \mathbb{R}$ is defined as $\ell(e) = (1-e)_+$ where $(\cdot)_+$ returns the positive part of the argument. Minimizing this upperbound will yield the optimal estimator '$\arg\min_w \mathrm{CI}_n(u_w(X), T)$' when $n \to \infty$ (see e.g. [3], section 4). This motivates the linear estimator $\min_w \sum_{i<j} \ell\left(w^T X_j - w^T X_i\right)$. Adding a regularization term $w^T w$, and rewriting the estimator as a standard quadratic programming problem results in the following estimator which is a blueprint for the more refined estimators lateron.

Definition 4 (Linear Predictive Failure Time Estimator) *The empirical optimal health function $u_w : \mathbb{R}^d \to \mathbb{R}$ is found as follows*

$$\begin{aligned} (\hat{w}, \hat{\xi}) = \arg\min_{w,\xi} \ & \frac{1}{2} w^T w + C \sum_{i<j, \delta_i = 1} W_{ij} \xi_{ij} \\ s.t. \quad & w^T X_j - w^T X_i \geq 1 - \xi_{ij}, \quad \xi_{ij} \geq 0, \quad \forall i < j : c(Z_i, Z_j) = 1. \end{aligned} \tag{25}$$

where $\{W_{ij}\}_{i<j}$ is a set with appropriate weighting terms as in Theorem 1, dealing with the specific censoring pattern.

As for the case of support vector machines, this optimization problem can be motivated alternatively using the concept of maximal margin. The margin of a point $x \in \mathbb{R}^d$ to a hyperplane $\{x \in \mathbb{R}^d : \ w^T x = 1\}$ is defined as $\frac{|w^T x|)}{\|w\|_2}$. The *signed* margin of a sample $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ with this hyperplane is defined as $\frac{y(w^T x)}{\|w\|_2}$. Now, for any couple (X_i, X_j) with corresponding times $T_i < T_j$ (or $i < j$), one can define a hyperplane

$\{w^T x + b_{ij} = 0 \ |(X_i, X_j), i < j\}$. Maximizing the margin between the two points can be done as '$\max_{w,b_{ij}} \frac{-(w^T X_i + b_{ij})}{\|w\|_2} + \frac{(w^T X_j + b_{ij})}{\|w\|_2} = \frac{w^T(X_j - X_i)}{\|w\|_2}$'. Or equivalently as '$\min_w \|w\|$ s.t. $w^T(X_j - X_i) \geq 1$'. Maximizing all margins between any couples (X_i, X_j) for all $i < j$ is be done by the following optimization problem

$$\min_w \frac{1}{2} w^T w \quad \text{s.t.} \quad w^T(X_j - X_i) \geq 1, \quad \forall i < j. \tag{26}$$

It is clear that if such a health u_w can be found, the corresponding concordance index equals one exactly. In general, it is advantageous to allow for violations in the constraints using slack variables. Doing so, and minimizing the one-norm of the slacks $\{\xi_{ij}\}_{i<j}$ results again in the estimator (25).

Motivated by the theoretical analysis, we modify the computation to a weaker optimality criterion, but with the same convergence properties while being usefull for larger datasets.

Definition 5 (Linear Predictive Failure Time Estimator (bis)) *Let $S \subset \{1, \ldots, \lfloor n/2 \rfloor\}$ containing no term i where $\delta_i = 1$. The empirical optimal health function $u_w : \mathbb{R}^d \to \mathbb{R}$ is found as follows*

$$\begin{aligned} (\hat{w}, \hat{\xi}) = \arg\min_{w,\xi} & \frac{1}{2} w^T w + C \sum_{i \in S} W'_{i(n-i)} \xi_{i(n-i)} \\ s.t. \quad & w^T X_{n-i} - w^T X_i \geq 1 - \xi_{i(n-i)}, \quad \xi_{i(n-i)} \geq 0, \quad \forall i \in S. \end{aligned} \tag{27}$$

where $\{W'_{i(n-i)}\}_{i<j}$ is a set with appropriate weighting terms dealing with the specific censoring pattern.

3.2 A Primal-Dual Maximal Margin Kernel Machine

It is shown how this formulation can be rewritten as a nonlinear kernel-based version based on an argument from convex optimization theory. It is shown that the derrivation yields extra insight into the structure of the solution, compared to a bunt application of the representer theorem.

Let $N \in \mathbb{N}$ be defined as $N = \frac{n(n-1)}{2}$ (or $N = |\{(ij) : i < j = 1, \ldots, n\}|$). The Lagrangian to (25) becomes $\mathcal{L}(w, \xi; \alpha, \beta) = \frac{1}{2} w^T w + C \sum_{i<j} W_{ij} \xi_{ij} - \sum_{i<j} \alpha_{ij} \left(w^T X_j - w^T X_i - 1 + \xi_{ij} \right) - \sum_{i<j} \beta_{ij} \xi_{ij}$ with multipliers $\alpha, \beta \in \mathbb{R}^N_+$. Switching the order $\max_{\alpha,\beta} \min_{w,\xi} \mathcal{L}(w, \xi; \alpha, \beta)$ to $\min_{w,\xi} \max_{\alpha,\beta} \mathcal{L}(w, \xi; \alpha, \beta)$ (Slater's condition is trivially satisfied) and taking the first order conditions for optimality gives

$$\begin{cases} \dfrac{\partial \mathcal{L}(w, \xi; \alpha, \beta)}{\partial w} = 0 \to w = \sum_{i<j} \alpha_{ij} (X_j - X_i) \\ \dfrac{\partial \mathcal{L}(w, \xi; \alpha, \beta)}{\partial \xi_{ij}} = 0 \to C W_{ij} = \alpha_{ij} + \beta_{ij} \qquad \forall i < j \end{cases} \tag{28}$$

and the dual problem becomes as such

$$\min_{\alpha} \frac{1}{2} \sum_{i<j} \sum_{k<l} \alpha_{ij}\alpha_{kl}(X_j - X_i)^T(X_l - X_k) - \sum_{i<j} \alpha_{ij}$$
$$\text{s.t.} \quad 0 \leq \alpha_{ij} \leq CW_{ij}, \quad \forall i < j. \quad (29)$$

Note the relationship with the dual expression corresponding with a maximal margin classifier (as the SVM). This can be rewritten in matrix-vector terms by introducing the matrix $D \in \{-1, 0, 1\}^{N \times n}$. Here, every kth row corresponds with a unique couple (i, j) where $i < j = 1, \ldots, n$. Now, $D_k\mathbf{x} = D_{(ij)}$ equals $X_j - X_i$, or $D_{(ij)} = (0, \ldots, -1, 0 \ldots, 1, 0 \ldots) \in \{-1, 0, 1\}^n$ for any $i < j = 1, \ldots, n$. The following Lemma summarizes results.

Lemma 2 (Dual of (25)) *The dual problem of (25) is given as*

$$\hat{\alpha} = \arg\min_{\alpha} \frac{1}{2}\alpha^T D(\mathbf{x}\mathbf{x}^T)D^T\alpha - 1_N^T\alpha \quad s.t. \quad 0_N \leq \alpha \leq CW, \quad (30)$$

where $W \in \mathbb{R}^N$ is defined as $h = \left(W_{21}, \ldots, W_{n(n-1)}\right)^T$. The corresponding optimal solution $\hat{w}$ can be evaluated in a new point $X_* \in \mathbb{R}^d$ as

$$u_w(X_*) = \hat{w}^T X_* = \sum_{i<j} \hat{\alpha}_{ij}(X_i - X_j)^T X_* = (\mathbf{x}X_*)^T \ D^T\hat{\alpha} \quad (31)$$

Application of the kernel trick gives the following result. Given a positive definite kernel function $K : \mathbb{R}^D \times \mathbb{R}^d \to \mathbb{R}$ inducing a feature map $\varphi : \mathbb{R}^d \to \mathbb{R}^{d_\varphi}$ such that $K(X, X') = \varphi(X)^T\varphi(X')$ for any $X, X' \in \mathbb{R}^d$. Let the kernel matrix $\Omega \in \mathbb{R}^{n \times n}$ be defined as $\Omega_{ij} = K(X_i, X_j)$ for all $i, j = 1, \ldots, n$, and let $\Omega(\cdot) : \mathbb{R}^d \to \mathbb{R}^n$ be defined as $\Omega(X) = (K(X_1, X), \ldots, K(X_n, X))^T \in \mathbb{R}^N$.

Lemma 3 (Kernel Health Function) *Given an appropriate kernel function and its corresponding kernel matrix, the optimal empirical faliure time predictor is found by*

$$\hat{\alpha} = \arg\min_{\alpha} \frac{1}{2}\alpha^T D\Omega D^T\alpha - 1_N^T\alpha \quad s.t. \quad 0_N \leq \alpha \leq CW, \quad (32)$$

and the corresponding optimal solution can be evaluated in a new point $X_* \in \mathbb{R}^d$ as

$$u_w(X_*) = \Omega(X_*)^T \ D^T\hat{\alpha}. \quad (33)$$

4 Conclusion

This technical report discusses an approach for predictive modeling of observational failure time data. Three contributions are made: (a) a weakly converging estimator of the concordancy in a context of independent censoring

is proposed and analysed, (b) it is illustrated how the expected concordancy and its sample estimate can be used to deduce a predictive distribution over the failure times for a new subject without imposing a metric on the failure times, and (c) the subsequent learning machine is implemented and a nonlinear version using Mercer kernels is proposed. There remain many open problems, inclusing an efficient algorithm for computing the appropriate weights in the proposed estimator, and a practical assessment of the estimator.

References

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

[2] P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[3] S. Clemencon, G. Lugosi, N. Vayatis, P. Aurer, and R. Meir. Ranking and scoring usingf empirical risk minimization, 2005.

[4] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[5] Jr.F.E. Harrell, K.L. Klee, R.M. Califf, D.B. Pryor, and R.A. Rosati. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*, 95(9):634–635, 1984.

[6] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000. MIT Press, Cambridge, MA.

[7] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[8] J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley series in probability and statistics. Wiley, 2002.

[9] W.H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

[10] A. Rakotomamonji. SVMs and area under the ROC curves. *Technical Report PSI-INSA Rouen 2004*, 2004.

[11] V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.