EPFL | MGT-418 : Convex Optimization | Project 2

Questions – Fall 2022

Due on Nov 23rd, 11:59PM

Survival Analysis: Support Vector Regression (graded)

Description

Survival analysis uses a feature vector $x \in \mathbb{R}^d$ to predict the time $y \in \mathbb{R}_+$ until a certain event occurs. Estimation tasks of this kind naturally arise in a wide variety of applications. For example, in medicine we aim to predict a patient's remaining lifetime, in economics we aim to predict when a company will go bankrupt, in manufacturing we aim to predict when a component or product will fail, and in marketing we aim to predict when a customer will 'churn' (unsubscribe from a service).

Intuitively, one might think that survival analysis is a simple regression task. However, a naïve regression model would yield biased predictors. To see this, assume that you work for the fictional furniture company $\iota\kappa\varepsilon\alpha$ (which targets mathematicians) and that you must predict how often one can sit on a chair before it breaks. The following experiment enables you to construct a training dataset. You develop a machine that simulates a person sitting down.[1] Using 100 machines of this type, you stress-test $n = 100$ chairs for one week, that is, you simulate 10,000 people sitting down on each chair. After the week, some chairs are broken, and thus you have observed their actual lifetime y. However, other chairs are still intact, and thus you have observed the lower bound $y = 10{,}000$ on their lifetime. Next, denote by x_i a feature vector and by y_i the observed liftime (or its lower bound) of the i-th chair. The dataset $\{(x_i, y_i)\}_{i=1}^n$ constructed in this way is biased. To eliminate the bias, we define a binary variable z_i indicating whether or not the lifetime of chair i has been observed, that is, we set

$$z_i = \begin{cases} 1 & \text{if } y_i < 10{,}000, \\ 0 & \text{if } y_i \geq 10{,}000. \end{cases}$$

Given the training dataset $\{(x_i, y_i, z_i)\}_{i=1}^n$ and a set of pairs $\mathcal{E} \subseteq \{1, \ldots, n\}^2$ with $y_i \geq y_j$ and $z_j = 1$ for every $(i, j) \in \mathcal{E}$, we use the generalized regression model $(\mathcal{P})$ to calibrate the parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ of a linear predictor $y \approx w^\top x + b$. The objective function of this regression model consists of a regularization term $\frac{1}{2} w^\top w$, a ranking loss with weight r_1 and a regression loss with weight r_2. The ranking loss ensures that if $(i, j) \in \mathcal{E}$, that is, if chair j broke before chair i, then a good predictor should predict a higher lifetime for chair i than for chair j. The regression loss ensures that if $z_i = 1$ ($z_i = 0$), then a good predictor should predict that the lifetime of chair i matches (exceeds) y_i.

$$\begin{array}{lll}
\displaystyle\min_{w,\, b,\, \xi \geq 0,\, \xi' \geq 0,\, \epsilon \geq 0} & \displaystyle\frac{1}{2} w^\top w + r_1 \sum_{(i,j) \in \mathcal{E}} \epsilon_{(i,j)} + r_2 \sum_{i=1}^n (\xi_i + \xi_i') & \\
\text{s.t.} & \epsilon_{(i,j)} \geq y_i - y_j - w^\top (x_i - x_j) & \forall (i, j) \in \mathcal{E} \\
& \xi_i \geq y_i - (w^\top x_i + b) & \forall i = 1, \ldots, n \\
& \xi_i' \geq z_i((w^\top x_i + b) - y_i) & \forall i = 1, \ldots, n
\end{array} \qquad (\mathcal{P})$$

[1] `https://www.youtube.com/watch?v=WicEZwRo-WU`

Questions

1. **(25 points)** Derive the KKT conditions for problem $(\mathcal{P})$, and show that the Lagrangian dual of $(\mathcal{P})$ is given by problem $(\mathcal{D})$ below. Explain why strong duality holds.

$$
\begin{array}{lll}
\max\limits_{\alpha\in\mathbb{R}^{|\mathcal{E}|},\beta\in\mathbb{R}^n,\gamma\in\mathbb{R}^n} & -\frac{1}{2}\sum\limits_{(i,j)\in\mathcal{E}}\sum\limits_{(k,l)\in\mathcal{E}}\alpha_{(i,j)}\alpha_{(k,l)}(x_i-x_j)^\top(x_k-x_l)+\sum\limits_{i=1}^n\sum\limits_{k=1}^n\gamma_k\beta_i z_k x_i^\top x_k & \\
& -\frac{1}{2}\sum\limits_{i=1}^n\sum\limits_{k=1}^n(\beta_i\beta_k+z_iz_k\gamma_i\gamma_k)x_i^\top x_k+\sum\limits_{(i,j)\in\mathcal{E}}\alpha_{(i,j)}(y_i-y_j) & \\
& -\sum\limits_{i=1}^n\sum\limits_{(k,l)\in\mathcal{E}}(\beta_i-z_i\gamma_i)\alpha_{(k,l)}x_i^\top(x_k-x_l)+\sum\limits_{i=1}^n\beta_i y_i-\sum\limits_{i=1}^n z_i\gamma_i y_i & (\mathcal{D}) \\
\text{s.t.} & 0\le\alpha_{(i,j)}\le r_1 \qquad \forall(i,j)\in\mathcal{E} & \\
& 0\le\beta_i,\gamma_i\le r_2 \qquad \forall i=1,\dots,n & \\
& \sum\limits_{i=1}^n(\beta_i-z_i\gamma_i)=0 &
\end{array}
$$

2. **(20 points)** Use the training data to construct $y=[y_1,\dots,y_n]^\top$ as well as a column vector $\Delta y\in\mathbb{R}^{|\mathcal{E}|}$ with entries y_i-y_j for all $(i,j)\in\mathcal{E}$, and introduce the diagonal matrix $Z=\mathrm{diag}[z_1,\dots,z_n]$. Show that the dual problem $(\mathcal{D})$ can be reformulated as the explicit quadratic program

$$
\begin{array}{lll}
\max\limits_{\alpha\in\mathbb{R}^{|\mathcal{E}|},\beta\in\mathbb{R}^n,\gamma\in\mathbb{R}^n} & -\frac{1}{2}\begin{bmatrix}\alpha\\ \beta-Z\gamma\end{bmatrix}^\top\begin{bmatrix}U & W^\top\\ W & V\end{bmatrix}\begin{bmatrix}\alpha\\ \beta-Z\gamma\end{bmatrix}+\begin{bmatrix}\Delta y\\ y\end{bmatrix}^\top\begin{bmatrix}\alpha\\ \beta-Z\gamma\end{bmatrix} & \\
\text{s.t.} & 0\le\alpha_{(i,j)}\le r_1 \qquad \forall(i,j)\in\mathcal{E} & (\mathcal{DM}) \\
& 0\le\beta_i,\gamma_i\le r_2 \qquad \forall i=1,\dots,n & \\
& \sum\limits_{i=1}^n(\beta_i-z_i\gamma_i)=0 &
\end{array}
$$

for some matrices $U\in\mathbb{S}_+^{|\mathcal{E}|}$, $V\in\mathbb{S}_+^n$ and $W\in\mathbb{R}^{n\times|\mathcal{E}|}$. Verify that problem $(\mathcal{DM})$ is convex.

3. **(30 points)** Implement problem $(\mathcal{DM})$ in PYTHON or MATLAB using the code skeletons available on Moodle. Also implement the naïve regression model $\min_w \frac{1}{2}w^\top w+r_2\sum_{i=1}^n|y_i-w^\top x_i|$. In both models, set $r_1=100/|\mathcal{E}|$ and $r_2=10{,}000/n$. You may use the vectors `y_comp` and `y_comp_bar` and the matrices `X_comp` and `X_comp_bar` to construct the matrices U and V.

   Derive a formula for the output $w^\top x+b$ of the optimal predictor on a test sample x in terms of the optimal solution of the dual problem $(\mathcal{DM})$. Compare the optimal predictors obtained from $(\mathcal{DM})$ and the naïve regression model on the telecom customer churn dataset available from Moodle. Specifically, compute the *mean absolute error*, which measures the average absolute error, the *C-index*, which measures the fraction of valid comparisons that were ranked correctly, and the *average underestimated survival*, which quantifies the average duration by which the churn time was underestimated conditional on it being underestimated, on the test data.

4. **(15 points)** Apply the kernel trick to problem $(\mathcal{DM})$, and solve the kernelized version of $(\mathcal{DM})$ using the radial basis function (RBF) kernel $K(x,x')=\exp(-\frac{1}{2}\|x-x'\|_2^2/\sigma^2)$ with $\sigma^2=1.5$. Compare the mean absolute error, the C-index and the average underestimated survival of the optimal predictors with and without RBF kernel on the test data.