

Secure Multi-Party Computation for Additive Models

Joyee Ghosh (advised by Prof. Jerry Reiter)

Introduction

- Y : response variable
- X_1, \dots, X_p : p predictors
- **goal**: model the dependence of Y on X_1, \dots, X_p
- multiple linear regression model:

$$Y = \alpha + X_1\beta_1 + X_2\beta_2 + \dots + X_p\beta_p + \epsilon \quad (1)$$

- additive model extends the linear regression model

$$Y = \alpha + \sum_j f_j(X_j) + \epsilon \quad (2)$$

- $E(\epsilon)=0$ and $\text{Var}(\epsilon)=\sigma^2$
- regression splines for fitting the f_j s

Smoothing Splines

- Consider the regression model

$$Y_i = r(x_i) + \epsilon_i \quad (3)$$

- estimate r by choosing $\widehat{r}_n(x)$ to minimize the penalized sum of squares

$$M(\lambda) = \sum_i (Y_i - \widehat{r}_n(x_i))^2 + \lambda J(r) \quad (4)$$

$J(r)$ is some roughness penalty

λ is a tuning parameter controlling the amount of smoothness

Cubic Splines

- **Definition:** Let $\xi_1 < \xi_2 < \dots < \xi_k$ be a set of ordered points called knots, contained in some interval (a,b) . A *cubic spline* is a continuous function r such that (i) r is a cubic polynomial over $(\xi_1, \xi_2), \dots$ and (ii) r has continuous first and second derivatives at the knots.
- It turns out that the function $\widehat{r}_n(x)$ that minimizes $M(\lambda)$ with penalty $J(r) = \int (r''(x))^2 dx$ is a natural cubic spline with knots at the data points. The estimator $\widehat{r}_n(x)$ is called a natural smoothing spline.

Regression Splines

- rather than placing a knot at each data point we choose fewer knots instead
- there is no shrinkage factor
- amount of smoothing is controlled by the number and placement of the knots
- **Truncated Power basis:** Let $\xi_1 < \xi_2 < \dots < \xi_k$ be knots contained in an interval (a,b) . Define $h_1(x) = 1$, $h_2(x) = x$, $h_3(x) = x^2$, $h_4(x) = x^3$, $h_j(x) = (x - \xi_{j-4})_+^3$ for $j=5, \dots, (k+4)$. The functions $\{h_1, \dots, h_{k+4}\}$ form a basis for the set of cubic splines at these knots, called the truncated power basis. So any cubic spline $r(x)$ with these knots can be written as

$$r(x) = \sum_{j=1}^{k+4} \beta_j h_j(x) \quad (5)$$

Horizontally Partitioned Data

- Suppose there are $K > 2$ agencies, and each agency has data on its n_j subjects: p predictors X^j and response y^j .
- The agencies want to model the dependence of y on X where y and X refer to the global data.

$$X = \begin{bmatrix} X^1 \\ X^2 \\ \vdots \\ X^k \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^k \end{bmatrix}$$

Fitting the Additive Model

The additive model that we would like to fit using regression splines is:

$$E(Y_i|X) = \beta_0 + \sum_{j=1}^p \mathbf{b}'_j(\mathbf{X}_j)\beta_j \quad (6)$$

$$= \beta_0 + \sum_{j=1}^p \sum_{l=1}^{k_j} b_{lj}(X_j)\beta_{lj} \quad (7)$$

\mathbf{b}_j : vector of k_j basis functions associated with the predictor X_j

β_j : vector of k_j parameters associated with \mathbf{b}_j

Fitting the Additive Model

- Large Linear Model:

$$E(Y|X) = \mathbf{B}\beta \quad (8)$$

- fit this using usual least squares method
- least squares estimator of β :

$$\hat{\beta} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y} \quad (9)$$

- need to compute $\mathbf{B}'\mathbf{B}$ and $\mathbf{B}'\mathbf{y}$
- usual secure regression technique

Secure Summation

- B^j : the concatenated basis matrices that agency j can compute from its data X^j for $j=1, \dots, k$

$$\mathbf{B}'\mathbf{B} = \sum_{j=1}^k (\mathbf{B}^j)' \mathbf{B}^j \quad (10)$$

- each agency j computes its own $(B^j)' B^j$ which is a square matrix
- elements of the k matrices can be combined using secure summation to obtain the elements of $\mathbf{B}'\mathbf{B}$
- $\mathbf{B}'\mathbf{y}$ can be computed similarly by secure summation on the entries of $(B_j)' y_j$

Discussion

- extension to vertically partitioned data
- knot selection for the regression splines
 - number
 - placement: percentiles of x
 - obtain an estimate of global percentiles without data integration