ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

STATISTICAL LEARNING

Selection Inference for CoRT

TTNT2025

Sinh viên:
Phạm Bách Hiệp
Nguyễn Ngọc Nhớ

Giảng viên:
Võ Nguyễn Lê Duy

Tóm Tắt

Trong báo cáo này, ta sẽ

Mục Lục

Chương 1

SI For CoRT

1.1 CoRT Algorithm

Thuật toán 1 Adaptive Knowledge Transfer

Input: $\mathcal{D}^{(k)}$ với $0 \leq k \leq K$

1: **Phân chia dữ liệu**: Ngẫu nhiên chia dữ liệu mục tiều $\mathcal{D}^{(0)}$ thành T (là số lẻ) phân bằng nhau và kí hiệu chúng là $D^{(0)}_{[1]}, \ldots, D^{(0)}_{[T]}$.
2: **for** $k = 1$ to K **do**
3: **for** $t = 1$ to T **do**
4: Tìm $\hat{\boldsymbol{\beta}}^{(0)}_{[t]}$ bằng cách chạy Lasso trên $(\mathcal{D}^{(0)} \setminus \mathcal{D}^{(0)}_{[t]})$.
5: Tìm $\hat{\boldsymbol{\beta}}^{(0k)}_{[t]}$ bằng cách chạy Lasso trên $(\mathcal{D}^{(0)} \setminus \mathcal{D}^{(0)}_{[t]}) \cup \mathcal{D}^{(k)}$.
6: Lần lượt tính toán hàm lỗi cho $\ell(\hat{\boldsymbol{\beta}}^{(0)}_{[t]}, \mathcal{D}^{(0)}_{[t]})$ và $\ell(\hat{\boldsymbol{\beta}}^{(0k)}_{[t]}, \mathcal{D}^{(0)}_{[t]})$,
7: **end for**
8: Lập hàm đếm $C(\boldsymbol{\beta}^{(0k)})$ với $\sum_{t=1}^{T} \mathbf{1}\left\{\ell(\hat{\boldsymbol{\beta}}^{(0k)}_{[t]}, \mathcal{D}^{(0)}_{[t]}) = \min_{\boldsymbol{\beta} \in \{\hat{\boldsymbol{\beta}}^{(0k)}_{[t]}, \hat{\boldsymbol{\beta}}^{(0)}_{[t]}\}} \ell(\boldsymbol{\beta}, \mathcal{D}^{(0)}_{[t]})\right\}$.
9: **end for**
10: Thu được ước lượng của S với $\hat{S} = \{k : C(\boldsymbol{\beta}^{(0k)}) \geq (T+1)/2\}$.
11: Thu được $\hat{\boldsymbol{\beta}}^{(0)}$ băng cách chạy CoRT với $\mathcal{D}^{(k)}, k \in \hat{S} \cup \{0\}$.

1.2 Bài toán con 1: Tính ổn định của các mô hình Huấn luyện

1.2.1 Selection event

Theo mô tả của Thuật toán 1 (Adaptive Knowledge Transfer), quá trình lựa chọn nguồn phụ thuộc trực tiếp vào các mô hình trung gian được huấn luyện trong vòng lặp kiểm chứng chéo (Cross-Validation). Cụ thể:

- **Dòng 4:** Tìm $\hat{\beta}_{[t]}^{(0)}$ bằng cách chạy Lasso trên tập huấn luyện đích $(D^{(0)} \setminus D_{[t]}^{(0)})$.
- **Dòng 5:** Tìm $\hat{\beta}_{[t]}^{(0k)}$ bằng cách chạy Lasso trên tập kết hợp $(D^{(0)} \setminus D_{[t]}^{(0)}) \cup D^{(k)}$.

1.2.2 Biểu diễn Selection event

Ta cần điều kiện hóa sự kiện tất cả các mô hình train này là không thay đổi. Cụ thể, tập biến hoạt động (active sets) $\hat{M}$ và dấu của các hệ số hồi quy (signs) $\hat{s}$ của mọi mô hình Baseline và Augmented trong suốt quá trình CV phải giữ nguyên giá trị quan sát được khi dữ liệu y thay đổi dọc theo đường z.

Kết luận Toán học: Áp dụng điều kiện KKT cho mỗi mô hình Lasso, ta được bất phương trình:

$$\underbrace{\begin{pmatrix} -\text{diag}(\mathbf{s}_{\mathcal{M}})\mathbf{v} \\ \mathbf{q} \\ -\mathbf{q} \end{pmatrix}}_{\boldsymbol{\psi}} z \leq \underbrace{\begin{pmatrix} \text{diag}(\mathbf{s}_{\mathcal{M}})\mathbf{u} \\ \mathbf{1} - \mathbf{p} \\ \mathbf{1} + \mathbf{p} \end{pmatrix}}_{\boldsymbol{\gamma}} \tag{1.1}$$

Từ đó ta xây dựng được ràng buộc tuyến tính cho tất cả mô hình Baseline và Augmented như sau:

$$\boldsymbol{\Psi} z \leq \boldsymbol{\Gamma} \tag{1.2}$$

trong đó:

$$\mathbf{\Psi} = \begin{bmatrix} \vdots \\ \boldsymbol{\psi}_{[t]}^{(0)} \\ \vdots \\ \boldsymbol{\psi}_{[t]}^{(0k)} \\ \vdots \end{bmatrix}, \quad \mathbf{\Gamma} = \begin{bmatrix} \vdots \\ \boldsymbol{\gamma}_{[t]}^{(0)} \\ \vdots \\ \boldsymbol{\gamma}_{[t]}^{(0k)} \\ \vdots \end{bmatrix} \tag{1.3}$$

1.3 Bài toán con 2: Tính ổn định của Bỏ phiếu Kiểm định

1.3.1 Selection event

Giai đoạn này quyết định việc một nguồn k có được đưa vào tập hợp $\hat{S}$ hay không dựa trên kết quả so sánh sai số dự báo trên tập kiểm định (Validation set). Theo Thuật toán 1:

- **Dòng 6:** Tính toán hàm lỗi $\ell(\cdot)$ cho mô hình Baseline và Augmented trên tập kiểm định $D_{[t]}^{(0)}$.
- **Dòng 8:** So sánh sai số và thực hiện bỏ phiếu thông qua hàm chỉ thị $1\{\ell_{aug} < \ell_{base}\}$.
- **Dòng 10:** Thu được ước lượng của S với $\hat{S} = \{k : C(\boldsymbol{\beta}^{(0k)}) \geq \frac{T+1}{2}\}$.

1.3.2 Biểu diễn Selection event

Ta cần điều kiện hóa sự kiện "Thắng/Thua"của mỗi cặp so sánh là không đổi. Vì nghiệm Lasso $\hat{\beta}(z)$ là tuyến tính theo z, hàm mất mát (bình phương sai số) trên tập kiểm định sẽ trở thành một hàm bậc hai theo z.

Xét hiệu sai số giữa mô hình Augmented và Baseline tại fold t:

$$\Delta_{t,k}(z) = \ell(\hat{\beta}_{[t]}^{(0k)}(z)) - \ell(\hat{\beta}_{[t]}^{(0)}(z)) \tag{1.4}$$

Kết luận Toán học: Triển khai biểu thức sai số, ta thu được đa thức bậc hai đặc trưng cho sự chênh lệch hiệu suất:

$$\Delta_{t,k}(z) = \mathcal{A}_{t,k} z^2 + \mathcal{B}_{t,k} z + \mathcal{C}_{t,k} \tag{1.5}$$

trong đó các hệ số $\mathcal{A}, \mathcal{B}, \mathcal{C}$ được tính toán dựa trên các tham số $(\mathbf{u}, \mathbf{v})$ từ Bài toán con 1 và dữ liệu kiểm định.

Khi đó, miền giá trị z thỏa mãn điều kiện ổn định bỏ phiếu được xác định bởi tập hợp các bất phương trình bậc hai:

$$\begin{cases} \mathcal{A}_{t,k} z^2 + \mathcal{B}_{t,k} z + \mathcal{C}_{t,k} < 0 & \text{nếu } \Delta_{t,k}(z_{obs}) < 0 \text{ (Augmented thắng)} \\ \mathcal{A}_{t,k} z^2 + \mathcal{B}_{t,k} z + \mathcal{C}_{t,k} \geq 0 & \text{nếu } \Delta_{t,k}(z_{obs}) \geq 0 \text{ (Baseline thắng)} \end{cases} \tag{1.6}$$

Cần lưu ý rằng dòng 10 của thuật toán ("Thu được ước lượng của S...") không yêu cầu xây dựng thêm một hệ bất phương trình mới. Đây là hệ quả trực tiếp của việc ổn định bỏ phiếu ở trên.

Cụ thể, tập hợp nguồn được chọn $\hat{S}$ được định nghĩa dựa trên tổng số phiếu bầu:

$$\hat{S} = \left\{ k : \sum_{t=1}^{T} 1\{\Delta_{t,k}(z) < 0\} \geq \frac{T+1}{2} \right\} \tag{1.7}$$

Do hệ bất phương trình ở Bài toán con 2 đã đảm bảo dấu của mọi $\Delta_{t,k}(z)$ là không đổi trong khoảng $\mathcal{Z}_{val}$, nên giá trị của hàm chỉ thị $1\{\cdot\}$ và tổng số phiếu cũng bất biến. Điều này suy ra tập hợp $\hat{S}$ là cố định trong miền $\mathcal{Z}_{val}$.

1.4 Bài toán con 3: Tính ổn định của Mô hình CoRT

1.4.1 Selection event

Sau khi xác định được tập hợp các nguồn tương đồng $\hat{S}$ từ giai đoạn trước, thuật toán tiến hành bước ước lượng tham số cuối cùng. Theo Dòng 11 của Thuật toán 1:

- **Dòng 11:** Thu được $\hat{\boldsymbol{\beta}}^{(0)}$ băng cách chạy CoRT với $\mathcal{D}^{(k)}, k \in \hat{S} \cup \{0\}$.

1.4.2 Biểu diễn Selection event

Cụ thể, tập biến hoạt động $\mathcal{M}_{CoRT}$ và vectơ dấu $\mathbf{s}_{CoRT}$ của mô hình Lasso trên dữ liệu kết hợp phải giữ nguyên giá trị quan sát được khi z thay đổi.

Kết luận Toán học: Áp dụng điều kiện KKT của bài toán Lasso trên tập dữ liệu kết hợp, ta thu được hệ bất phương trình tuyến tính xác định khoảng $\mathcal{Z}_{CoRT}$:

$$\mathbf{\Omega} z \leq \mathbf{\Upsilon} \tag{1.8}$$

Trong đó, các ma trận hệ số được xác định cụ thể từ điều kiện KKT của mô hình cuối cùng (tương tự như Bài toán con 1 nhưng áp dụng trên $X_{combined}$):

$$\mathbf{\Omega} = \begin{pmatrix} -\text{diag}(\mathbf{s}_{CoRT})\mathbf{v}_{CoRT} \\ \mathbf{q}_{CoRT} \\ -\mathbf{q}_{CoRT} \end{pmatrix}, \quad \mathbf{\Upsilon} = \begin{pmatrix} \text{diag}(\mathbf{s}_{CoRT})\mathbf{u}_{CoRT} \\ \mathbf{1} - \mathbf{p}_{CoRT} \\ \mathbf{1} + \mathbf{p}_{CoRT} \end{pmatrix} \tag{1.9}$$