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

STATISTICAL LEARNING

Tiêu đề

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

CoRT

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 $\mathcal{D}^{(k)}, k \in \hat{S} \cup \{0\}$.

Chương 2

Xác Định Vùng Z

2.1 Bài toán con 1

Trong quá trình đánh giá chéo (cross-validation), thuật toán thực hiện huấn luyện nhiều mô hình Lasso trên các tập dữ liệu con khác nhau. Để đảm bảo cấu trúc của các mô hình này không thay đổi khi biến z di chuyển, ta cần cố định tập đặc trưng được chọn (active set) và dấu của các hệ số hồi quy tương ứng cho từng fold.

Với mỗi fold $t \in \{1, \ldots, T\}$ và mỗi nguồn $k \in \{1, \ldots, K\}$, ta định nghĩa các tập đặc trưng tích cực và tập dấu tương ứng tại giá trị z như sau:

- Đối với mô hình gốc (chỉ dùng dữ liệu đích):

$$\mathcal{M}_{[t]}^{(0)}(z) = \left\{ j : \hat{\beta}_{[t],j}^{(0)}(z) \neq 0 \right\} \tag{2.1}$$

$$\mathcal{S}_{[t]}^{(0)}(z) = \operatorname{sign}\left((\hat{\beta}_{[t]}^{(0)}(z))_{\mathcal{M}_{[t]}^{(0)}(z)} \right) \tag{2.2}$$

- Đối với mô hình kết hợp (kết hợp dữ liệu đích và nguồn k):

$$\mathcal{M}_{[t]}^{(0k)}(z) = \left\{ j : \hat{\beta}_{[t],j}^{(0k)}(z) \neq 0 \right\} \tag{2.3}$$

$$\mathcal{S}_{[t]}^{(0k)}(z) = \operatorname{sign}\left((\hat{\beta}_{[t]}^{(0k)}(z))_{\mathcal{M}_{[t]}^{(0k)}(z)} \right) \tag{2.4}$$

Để đảm bảo tính nhất quán với quan sát thực tế, điều kiện ổn định cho bài toán con này là tập đặc trưng và dấu của tất cả các mô hình phải trùng khớp với giá trị quan sát $\mathcal{M}_{obs}$ và $\mathcal{S}_{obs}$. Cụ thể, ta tìm khoảng $\mathcal{Z}_{train}$ sao cho:

$$\mathcal{Z}_{train} = \bigcap_{t=1}^{T} \left(\left\{ \mathcal{M}_{[t]}^{(0)}(z) = \mathcal{M}_{[t],obs}^{(0)},\ \mathcal{S}_{[t]}^{(0)}(z) = \mathcal{S}_{[t],obs}^{(0)} \right\} \right)$$
$$\cap \bigcap_{k=1}^{K} \bigcap_{t=1}^{T} \left(\left\{ \mathcal{M}_{[t]}^{(0k)}(z) = \mathcal{M}_{[t],obs}^{(0k)},\ \mathcal{S}_{[t]}^{(0k)}(z) = \mathcal{S}_{[t],obs}^{(0k)} \right\} \right) \tag{2.5}$$

Điều kiện này được thỏa mãn bằng cách giải hệ bất phương trình tuyến tính (từ điều kiện KKT) cho từng mô hình Lasso.

2.2 Bài toán con 2

Sau khi đã cố định cấu trúc của các mô hình Lasso (trong Bài toán con 1), bước tiếp theo của thuật toán là so sánh kết quả của chúng trên tập kiểm định (validation fold) để đưa ra quyết định "bỏ phiếu"cho dữ liệu nguồn.

Với mỗi fold $t \in \{1, \ldots, T\}$ và mỗi nguồn $k \in \{1, \ldots, K\}$, hàm mất mát (loss function) được sử dụng là bình phương sai số (Squared Error).

Ta định nghĩa sự chênh lệch lỗi giữa mô hình kết hợp và mô hình gốc tại giá trị z là:

$$\Delta_{[t]}^{(k)}(z) = \ell\left(\hat{\beta}_{[t]}^{(0k)}(z), \mathcal{D}_{[t]}^{(0)}\right) - \ell\left(\hat{\beta}_{[t]}^{(0)}(z), \mathcal{D}_{[t]}^{(0)}\right) \tag{2.6}$$

Để đảm bảo tính nhất quán của quá trình bỏ phiếu, dấu của sự chênh lệch lỗi này phải trùng khớp với kết quả quan sát thực tế. Cụ thể, nếu nguồn k "chiến thắng"ở fold t trong dữ liệu quan sát, nó cũng phải tiếp tục thắng trong khoảng z đang xét, và ngược lại.

- Nếu mô hình kết hợp thắng (có hàm mất mát nhỏ hơn), ta tìm khoảng z sao cho $\Delta_{[t]}^{(k)}(z) < 0$.
- Nếu mô hình kết hợp thua (có hàm mất mát lớn hơn), ta tìm khoảng z sao cho $\Delta_{[t]}^{(k)}(z) \geq 0$.

Khoảng giá trị $\mathcal{Z}_{val}$ thỏa mãn điều kiện ổn định này được xác định bởi:

$$\mathcal{Z}_{val} = \bigcap_{k=1}^{K} \bigcap_{t=1}^{T} \left\{ z \in R : \text{sign}\left(\Delta_{[t]}^{(k)}(z)\right) = \text{sign}\left(\Delta_{[t],\text{obs}}^{(k)}\right) \right\} \tag{2.7}$$

2.3 Bài toán con 3

Bước cuối cùng của thuật toán Adaptive Knowledge Transfer là thực hiện hồi quy (CoRT) trên dữ liệu kết hợp từ tập nguồn $\hat{S} \cup \{0\}$ để thu được mô hình cuối cùng. Mục tiêu của chúng ta là kiểm định thống kê cho các đặc trưng được chọn bởi mô hình này. Do đó, điều kiện tiên quyết là tập đặc trưng cuối cùng này không được thay đổi.

Gọi $\hat{\beta}^{(final)}(z)$ là ước lượng hệ số hồi quy cuối cùng tại giá trị z. Ta định nghĩa tập đặc trưng và dấu tương ứng là:

$$\mathcal{M}_{final}(z) = \left\{ j : \hat{\beta}_j^{(final)}(z) \neq 0 \right\} \tag{2.8}$$

$$\mathcal{S}_{final}(z) = \text{sign}\left((\hat{\beta}^{(final)}(z))_{\mathcal{M}_{final}(z)} \right) \tag{2.9}$$

Khoảng giá trị $\mathcal{Z}_{final}$ thỏa mãn điều kiện này được xác định bởi:

$$\mathcal{Z}_{final} = \{ z \in R : \mathcal{M}_{final}(z) = \mathcal{M}_{obs},\ \ \mathcal{S}_{final}(z) = \mathcal{S}_{obs} \} \tag{2.10}$$

2.4 Tổng hợp khoảng cắt cụt

Để xây dựng thống kê kiểm định hợp lệ, ta cần tìm khoảng $\mathcal{Z}$ sao cho toàn bộ quy trình lựa chọn mô hình—từ huấn luyện các mô hình con, so sánh lỗi, đến chọn mô hình cuối cùng—đều cho ra kết quả nhất quán với dữ liệu quan sát.

Khoảng $\mathcal{Z}$ cuối cùng là giao của nghiệm từ cả ba bài toán con:

$$\mathcal{Z} = \mathcal{Z}_{train} \cap \mathcal{Z}_{val} \cap \mathcal{Z}_{final} \tag{2.11}$$

Trong đó:

- $\mathcal{Z}_{train}$: Đảm bảo các biến hoạt động trong mô hình Lasso trong quá trình Cross-Validation không thay đổi.
- $\mathcal{Z}_{val}$: Đảm bảo kết quả bỏ phiếu của các nguồn là không thay đổi.
- $\mathcal{Z}_{final}$: Đảm bảo tập đặc trưng cuối cùng được chọn là $\mathcal{M}_{obs}$.

Bằng cách giới hạn không gian dữ liệu trong khoảng $\mathcal{Z}$ này, ta có thể tính toán p-value có chọn lọc (selective p-value) chính xác cho các đặc trưng trong $\mathcal{M}_{obs}$.

Chương 3

Công thức toán

3.1 Bài toán con 1

Mục tiêu của phần này là xác định khoảng giá trị của z sao cho cấu trúc của các mô hình Lasso (tập đặc trưng được chọn và dấu của hệ số) giữ nguyên không đổi.

3.1.1 Điều kiện KKT và Tuyến tính hóa nghiệm

Xét một mô hình Lasso tổng quát với ma trận thiết kế $\mathbf{X}$ và vector phản hồi $\mathbf{Y}(z)$. Gọi $\mathcal{M} = \{j : \hat{\beta}_j \neq 0\}$ là tập đặc trưng hoạt động (active set) và $\mathcal{M}^c$ là tập đặc trưng không hoạt động (inactive set). Gọi $\mathbf{s}_\mathcal{M}$ là vector dấu của các hệ số tập đặc trưng hoạt động.

Điều kiện Karush-Kuhn-Tucker (KKT) cho nghiệm của bài toán Lasso là:

$$\mathbf{X}_\mathcal{M}^\top \left(\mathbf{X}_\mathcal{M} \hat{\boldsymbol{\beta}}_\mathcal{M}(z) - \mathbf{Y}(z) \right) + \lambda \mathbf{s}_\mathcal{M}(z) = \mathbf{0}, \tag{3.1}$$

$$\mathbf{X}_{\mathcal{M}^c}^\top \left(\mathbf{X}_\mathcal{M} \hat{\boldsymbol{\beta}}_\mathcal{M}(z) - \mathbf{Y}(z) \right) + \lambda \mathbf{s}_{\mathcal{M}^c}(z) = \mathbf{0}, \tag{3.2}$$

$$\mathbf{s}_j = \operatorname{sign}\left(\hat{\beta}_j(z) \right), \quad \forall j \in \mathcal{M}, \tag{3.3}$$

$$\|s_j\|_\infty < 1, \quad \forall j \in \mathcal{M}^c. \tag{3.4}$$

Từ các phương trình (3.1) và (3.2), ta suy ra nghiệm $\hat{\boldsymbol{\beta}}$ và biến đối ngẫu

s phụ thuộc tuyến tính vào $\mathbf{Y}(z)$:

$$\hat{\boldsymbol{\beta}}_{\mathcal{M}}(z) = (\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{X}_{\mathcal{M}})^{-1}\left(\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{Y}(z) - \lambda \mathbf{s}_{\mathcal{M}}\right) \tag{3.5}$$

$$\mathbf{s}_{\mathcal{M}^c}(z) = \frac{1}{\lambda}\mathbf{X}_{\mathcal{M}^c}^{\top}\left(\mathbf{I} - \mathbf{P}_{\mathcal{M}}\right)\mathbf{Y}(z) + \mathbf{X}_{\mathcal{M}^c}^{\top}\mathbf{X}_{\mathcal{M}}(\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{X}_{\mathcal{M}})^{-1}\mathbf{s}_{\mathcal{M}} \tag{3.6}$$

trong đó $\mathbf{P}_{\mathcal{M}} = \mathbf{X}_{\mathcal{M}}(\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{X}_{\mathcal{M}})^{-1}\mathbf{X}_{\mathcal{M}}^{\top}$ là ma trận chiếu lên không gian con active.

Giả sử $\mathbf{Y}(z)$ có dạng tuyến tính $\mathbf{Y}(z) = \mathbf{a} + \mathbf{b}z$. Khi đó, các đại lượng trên cũng trở thành hàm tuyến tính của z:

1. Đối với hệ số Active ($\hat{\boldsymbol{\beta}}_{\mathcal{M}}$):

$$\hat{\boldsymbol{\beta}}_{\mathcal{M}}(z) = \underbrace{(\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{X}_{\mathcal{M}})^{-1}(\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{a} - \lambda \mathbf{s}_{\mathcal{M}})}_{\text{Vector hệ số chặn: } \mathbf{u}} + \underbrace{(\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{X}_{\mathcal{M}})^{-1}\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{b}}_{\text{Vector hệ số góc: } \mathbf{v}} \cdot z \implies \hat{\boldsymbol{\beta}}_{\mathcal{M}}(z) = \mathbf{u} + \mathbf{v}z \tag{3.7}$$

2. Đối với biến đối ngẫu Inactive ($\mathbf{s}_{\mathcal{M}^c}$):

$$\mathbf{s}_{\mathcal{M}^c}(z) = \underbrace{\left[\frac{1}{\lambda}\mathbf{X}_{\mathcal{M}^c}^{\top}(\mathbf{I} - \mathbf{P}_{\mathcal{M}})\mathbf{a} + \mathbf{X}_{\mathcal{M}^c}^{\top}\mathbf{X}_{\mathcal{M}}(\mathbf{X}_{\mathcal{M}}^{\top}\mathbf{X}_{\mathcal{M}})^{-1}\mathbf{s}_{\mathcal{M}}\right]}_{\text{Vector hệ số chặn: } \mathbf{p}} + \underbrace{\left[\frac{1}{\lambda}\mathbf{X}_{\mathcal{M}^c}^{\top}(\mathbf{I} - \mathbf{P}_{\mathcal{M}})\mathbf{b}\right]}_{\text{Vector hệ số góc: } \mathbf{q}} \cdot z$$

$$\implies \mathbf{s}_{\mathcal{M}^c}(z) = \mathbf{p} + \mathbf{q}z$$

3.1.2 Xây dựng các Bất phương trình ràng buộc

Để đảm bảo cấu trúc mô hình không đổi, z phải thỏa mãn các điều kiện sau cho mọi đặc trưng j:

Điều kiện nhất quán dấu

Với các đặc trưng active $j \in \mathcal{M}$, dấu của hệ số không được thay đổi: $\text{sign}(\hat{\beta}_j(z)) = s_j$. Điều này tương đương với $s_j \cdot \hat{\beta}_j(z) > 0$.

$$s_j(u_j + v_j z) > 0 \iff s_j v_j z > -s_j u_j \iff -s_j v_j z < s_j u_j \tag{3.8}$$

Viết dưới dạng ma trận:

$$-\text{diag}(\mathbf{s}_{\mathcal{M}})\mathbf{v}z < \text{diag}(\mathbf{s}_{\mathcal{M}})\mathbf{u} \tag{3.9}$$

Điều kiện dừng

Với các đặc trưng inactive $j \in \mathcal{M}^c$, độ lớn tương quan phải nhỏ hơn λ, tức là $|s_j(z)| < 1 \iff -1 < s_j(z) < 1$.

- Cận trên: $p_j + q_j z < 1 \implies q_j z < 1 - p_j$
- Cận dưới: $p_j + q_j z > -1 \implies -q_j z < 1 + p_j$

Viết dưới dạng ma trận khối:

$$\begin{pmatrix} \mathbf{q} \\ -\mathbf{q} \end{pmatrix} z \leq \begin{pmatrix} \mathbf{1} - \mathbf{p} \\ \mathbf{1} + \mathbf{p} \end{pmatrix} \tag{3.10}$$

Hệ Bất phương trình tổng hợp cho một mô hình

Kết hợp hai điều kiện trên, ta có hệ bất phương trình $\boldsymbol{\psi} z \leq \boldsymbol{\gamma}$ xác định khoảng ổn định cho một mô hình Lasso duy nhất:

$$\underbrace{\begin{pmatrix} -\text{diag}(\mathbf{s}_{\mathcal{M}})\mathbf{v} \\ \mathbf{q} \\ -\mathbf{q} \end{pmatrix}}_{\boldsymbol{\psi}} z \leq \underbrace{\begin{pmatrix} \text{diag}(\mathbf{s}_{\mathcal{M}})\mathbf{u} \\ \mathbf{1} - \mathbf{p} \\ \mathbf{1} + \mathbf{p} \end{pmatrix}}_{\boldsymbol{\gamma}} \tag{3.11}$$

3.1.3 Áp dụng cho Mô hình Augmented

Trong thuật toán Adaptive Knowledge Transfer, mô hình Augmented tại fold t với nguồn k sử dụng dữ liệu kết hợp giữa dữ liệu đích và dữ liệu nguồn. Ta định nghĩa các siêu ma trận (super-matrices) như sau:

- **Ma trận thiết kế gộp (X):**

$$X = \begin{bmatrix} \mathbf{X}^{(0)}_{[t]} \\ \mathbf{X}^{(k)} \end{bmatrix}$$

- **Vector phản hồi gộp ($Y(z)$):** Được phân rã thành thành phần hằng số $\mathbf{A}$ và thành phần phụ thuộc z là $\mathbf{B}$:

$$Y(z) = \underbrace{\begin{bmatrix} \mathbf{a}_{tgt} \\ \mathbf{Y}^{(k)} \end{bmatrix}}_{\mathbf{A}} + \underbrace{\begin{bmatrix} \mathbf{b}_{tgt} \\ \mathbf{0} \end{bmatrix}}_{\mathbf{B}} z$$

3.1.4 Hệ Bất phương trình Tổng thể

Để đảm bảo tính ổn định cho toàn bộ quá trình huấn luyện trong Cross-validation, z phải thỏa mãn điều kiện ổn định của tất cả T mô hình Baseline và $T \times K$ mô hình Augmented đồng thời.

Hệ bất phương trình tổng thể $\mathbf{\Psi}_{total} z \leq \mathbf{\Gamma}_{total}$ được cấu trúc như sau:

$$\mathbf{\Psi}_{total} = \begin{bmatrix} \vdots \\ \boldsymbol{\psi}_{[t]}^{(0)} \\ \vdots \\ \boldsymbol{\psi}_{[t]}^{(0k)} \\ \vdots \end{bmatrix}, \quad \mathbf{\Gamma}_{total} = \begin{bmatrix} \vdots \\ \boldsymbol{\gamma}_{[t]}^{(0)} \\ \vdots \\ \boldsymbol{\gamma}_{[t]}^{(0k)} \\ \vdots \end{bmatrix} \tag{3.12}$$

Trong đó:

- Khối trên tương ứng với T mô hình Baseline (chỉ dùng dữ liệu đích).
- Khối dưới tương ứng với $T \times K$ mô hình Augmented (dữ liệu đích kết hợp nguồn).

Giao của tất cả các bất phương trình này định nghĩa khoảng $\mathcal{Z}_{train}$ nơi cấu trúc của mọi mô hình con được bảo toàn.

3.2 Bài toán con 2

Mục tiêu của phần này là xác định khoảng giá trị $\mathcal{Z}_{val}$ sao cho kết quả so sánh hiệu suất giữa mô hình Augmented và mô hình Baseline trên tập kiểm định (validation fold) được giữ nguyên. Điều này đảm bảo tính ổn định của quá trình "bỏ phiếu"chọn nguồn tri thức.

3.2.1 Hàm Mất mát dưới dạng Hàm Bậc hai

Tại mỗi vòng lặp Cross-validation, ta sử dụng một fold t làm tập kiểm định. Gọi $\mathbf{X}_{[t]}^{(0)}$ là ma trận thiết kế của fold kiểm định và $\mathbf{Y}_{[t]}^{(0)}(z) = \mathbf{a}_{[t]}^{(0)} + \mathbf{b}_{[t]}^{(0)} z$ là vector phản hồi tương ứng.

Hàm mất mát được sử dụng là Bình phương Sai số (Squared Error Loss):

$$\ell(z) = \|\mathbf{Y}_{[t]}^{(0)}(z) - \mathbf{X}_{[t]}^{(0)} \hat{\boldsymbol{\beta}}(z)\|_2^2 \tag{3.13}$$

Do nghiệm Lasso $\hat{\boldsymbol{\beta}}(z)$ đã được chứng minh là hàm tuyến tính theo z (trong miền $\mathcal{Z}_{train}$), ta có thể khai triển phần dư (residual) như sau:

$$\mathbf{r}(z) = \mathbf{Y}^{(0)}_{[t]}(z) - \mathbf{X}^{(0)}_{[t]}(\mathbf{u} + \mathbf{v}z) \tag{3.14}$$

$$= \underbrace{(\mathbf{a}^{(0)}_{[t]} - \mathbf{X}^{(0)}_{[t]}\mathbf{u})}_{\boldsymbol{\phi}} + \underbrace{(\mathbf{b}^{(0)}_{[t]} - \mathbf{X}^{(0)}_{[t]}\mathbf{v})}_{\boldsymbol{\omega}} z \tag{3.15}$$

$$= \boldsymbol{\phi} + \boldsymbol{\omega} z \tag{3.16}$$

Trong đó $\boldsymbol{\phi}$ và $\boldsymbol{\omega}$ lần lượt là vector hệ số chặn và hệ số góc của phần dư trên tập kiểm định. Lưu ý rằng $\mathbf{u}, \mathbf{v}$ là các tham số của mô hình được huấn luyện (Baseline hoặc Augmented) đã tính ở Bài toán con 1.

Khi đó, hàm mất mát trở thành một hàm bậc hai theo z:

$$\ell(z) = \|\boldsymbol{\phi} + \boldsymbol{\omega} z\|_2^2 = (\boldsymbol{\phi} + \boldsymbol{\omega} z)^\top (\boldsymbol{\phi} + \boldsymbol{\omega} z) \tag{3.17}$$

$$= \underbrace{(\boldsymbol{\omega}^\top \boldsymbol{\omega})}_{C_2} z^2 + \underbrace{2(\boldsymbol{\phi}^\top \boldsymbol{\omega})}_{C_1} z + \underbrace{(\boldsymbol{\phi}^\top \boldsymbol{\phi})}_{C_0} \tag{3.18}$$

3.2.2 Điều kiện Ổn định Bỏ phiếu (Voting Stability)

Để xác định nguồn k có được chọn hay không, thuật toán so sánh lỗi của mô hình Augmented (ℓ_{aug}) với mô hình Baseline (ℓ_{base}). Xét hàm chênh lệch lỗi:

$$\Delta(z) = \ell_{aug}(z) - \ell_{base}(z) \tag{3.19}$$

Thay các biểu thức bậc hai vào, ta có:

$$\Delta(z) = (C_2^{aug} - C_2^{base})z^2 + (C_1^{aug} - C_1^{base})z + (C_0^{aug} - C_0^{base}) = \mathcal{A}z^2 + \mathcal{B}z + \mathcal{C} \tag{3.20}$$

Gọi $\Delta(0)$ là giá trị chênh lệch lỗi quan sát được tại $z = 0$. Để kết quả bỏ phiếu không thay đổi, dấu của $\Delta(z)$ phải trùng với dấu của $\Delta(0)$. Ta có hai trường hợp:

Trường hợp 1: Augmented Model chiến thắng ($\Delta(0) < 0$)

Nếu quan sát thấy mô hình có nguồn tri thức tốt hơn mô hình cơ sở, ta cần tìm khoảng z sao cho:

$$\mathcal{A}z^2 + \mathcal{B}z + \mathcal{C} < 0 \tag{3.21}$$

Trường hợp 2: Baseline Model chiến thắng ($\Delta(0) \geq 0$)

Nếu quan sát thấy mô hình cơ sở tốt hơn hoặc bằng, ta cần tìm khoảng z sao cho:

$$\mathcal{A}z^2 + \mathcal{B}z + \mathcal{C} \geq 0 \tag{3.22}$$

3.3 Bài toán con 3

3.3.1 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 $\mathcal{Z}_{CoRT}$:

$$\mathbf{\Omega} z \leq \mathbf{\Upsilon} \tag{3.23}$$

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 $X_{combined}$):

$$\mathbf{\Omega} = \begin{pmatrix} -\text{diag}(\mathbf{s}_{CoRT})\mathbf{v}_{CoRT} \\ \mathbf{q}_{CoRT} \\ -\mathbf{q}_{CoRT} \end{pmatrix}, \quad \mathbf{\Upsilon} = \begin{pmatrix} \text{diag}(\mathbf{s}_{CoRT})\mathbf{u}_{CoRT} \\ \mathbf{1} - \mathbf{p}_{CoRT} \\ \mathbf{1} + \mathbf{p}_{CoRT} \end{pmatrix} \tag{3.24}$$