



A Feature Selection Study for Medical Image deep learning

Sanghyuck Lee

2021. 04. 27.

Outline

- Introduction
- Some Mathematical Issues in last week
- Example
- Thyroid Research Project Report

Introduction

- Summary of last presentation
 - 1) False Discovery Rate (FDR)
 - Multiple comparisons in Statistics.
 - FDR in Feature Selection:
 - => Applicable by Barber & Candès (2015)
 - 2) Knockoffs filter example
 - Generating knockoffs
- In this week:
 - => Some mathematical issues of last week
 - => The continuing process with example

Some Mathematical Issues in last week

1. In initialization phase

1-1: Knockoff formula: $\tilde{X} = X\Sigma^{-1}(\Sigma - D) + \tilde{U}(2D - D\Sigma^{-1}D)^{\frac{1}{2}}$

1-2: 2 cases: $n \geq 2p$ or $n < 2p$

1-3: Standardizing original feature set

2. In calculation of D

2-1: $2\Sigma - D$ is positive definite, where $D = \text{diag}\{s\}$

2-2: $s_j = 2\lambda_{\min}(\Sigma) \wedge 1$ (Equi-correlated method), where $\tilde{X}_j^T X_j = 1 - s_j$

3. In calculation of $\tilde{U}$

3-1: matrix $\tilde{U}$ is orthonormal to the column space of X

3-2: Using $p \sim 2p$ row of Q matrix as $\tilde{U}$ matrix

Example

- $\tilde{X} = X\Sigma^{-1}(\Sigma - D) + \tilde{U}(2D - D\Sigma^{-1}D)^{\frac{1}{2}}$

$(n \times p) =$

$$(n \times p)(n \times n)\{(n \times n) - (n \times n)\} + (n \times p)\{2(n \times n) - (n \times n)(n \times n)(n \times n)\}^{\frac{1}{2}}$$

- By substituting the previously calculated Σ^{-1} , D , and $\tilde{U}$ into the formula, knockoff matrix $\tilde{X}$ can be calculated

Example

$$X = \begin{vmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & -1 & 1 \\ -1 & -1 & -1 \\ 1 & 1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \end{vmatrix}$$

$$\tilde{X} = \begin{vmatrix} 1.42 & 0.03 & 0.52 \\ -0.73 & 1.85 & -0.48 \\ -0.35 & -0.03 & -0.57 \\ 2.36 & -1.03 & -1.1 \\ 0.31 & 1.81 & -1.89 \\ -0.25 & -0.13 & 1.73 \\ 0.31 & -0.89 & 0.15 \\ 0.81 & 0.8 & 0.88 \\ -0.31 & 0.18 & -0.62 \\ -0.81 & -0.09 & -0.41 \\ 0.31 & -0.89 & 0.15 \end{vmatrix}$$

$$X^T X = \begin{vmatrix} 1 & -0.3 & -0.1 \\ -0.3 & 1 & -0.3 \\ -0.1 & -0.3 & 1 \end{vmatrix}$$

$$\tilde{X}^T \tilde{X} = \begin{vmatrix} 0.91 & -0.34 & -0.05 \\ -0.34 & 0.98 & -0.28 \\ -0.05 & -0.28 & 0.98 \end{vmatrix}$$

Example

Step 1) Feature importances $Z_i, \tilde{Z}_i$

Step 2) Feature statistics $W_i = f(Z_i, \tilde{Z}_i)$

Step 3) Feature selection procedure

Example

Step 1) Calculate feature importances Z_i , $\tilde{Z}_i$ using Ordinary Least Squares (OLS)

- Fit a OLS model of $y^T = [1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0]$ on $[X, \tilde{X}]$

$$\hat{y} = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \tilde{\beta}_1 \tilde{X}_{1i} + \tilde{\beta}_2 \tilde{X}_{2i} + \tilde{\beta}_3 \tilde{X}_{3i} + \epsilon_i$$

- $Z_i := \beta_i$ and $\tilde{Z}_i := \tilde{\beta}_i$

$$Z_1 = 0.13, Z_2 = 0.39, Z_3 = 0.07, \tilde{Z}_1 = 0.58, \tilde{Z}_2 = 0.28 \text{ and } \tilde{Z}_3 = 0.20$$

Example

Step 2) Calculate feature statistics using difference of coefficients

- $W_i = Z_i - \tilde{Z}_i$

$$Z_1 = 0.13, Z_2 = 0.39, Z_3 = 0.07, \tilde{Z}_1 = 0.58, \tilde{Z}_2 = 0.28 \text{ and } \tilde{Z}_3 = 0.20$$

$$W_1 = 0.13 - 0.58 = -0.45, W_2 = 0.39 - 0.28 = 0.11, W_3 = 0.07 - 0.20 = -0.13$$

$$\therefore W_1 = -0.45, W_2 = 0.11, W_3 = -0.13$$

Example

Step 3) Run the knockoff selection procedure at level q .

$$q = 0.2$$
$$W_1 = -0.45, \quad W_2 = 0.11, \quad W_3 = -0.13$$

$$\widehat{FDR}(t) = \frac{\#\{j: W_j \leq -t\}}{\#\{j: W_j \geq t\}}$$

$$\text{data-dependent threshold } \tau = \min \left\{ t \in W: \frac{\#\{j: W_j \leq -t\}}{\#\{j: W_j \geq t\}} \leq q \right\}$$

$t = W_1 = -0.45, \#\{j: W_j \leq -t\} = 3, \#\{j: W_j \geq t\} = 3, 1 > q \Rightarrow$ Not satisfied

$t = W_2 = 0.11, \#\{j: W_j \leq -t\} = 2, \#\{j: W_j \geq t\} = 1, 2 > q \Rightarrow$ Not satisfied

$t = W_3 = -0.13, \#\{j: W_j \leq -t\} = 3, \#\{j: W_j \geq t\} = 2, 3/2 > q \Rightarrow$ Not satisfied

$\Rightarrow$ No threshold, not all features are selected

Example

Another example with additional W s.

$$q = 0.2$$
$$W_1 = -0.45, \quad W_2 = 0.11, \quad W_3 = -0.13,$$

Additional W s: $W_4 = 0.4, \quad W_5 = 0.5, \quad W_6 = 0.6, \quad W_7 = 0.7,$
 $W_8 = 0.8, \quad W_9 = 0.9, \quad W_{10} = 1$

$$\widehat{FDR}(t) = \frac{\#\{j: W_j \leq -t\}}{\#\{j: W_j \geq t\}}, \quad \text{Threshold } \tau = \min \left\{ t \in W : \frac{\#\{j: W_j \leq -t\}}{\#\{j: W_j \geq t\}} \leq q \right\}$$

$$\widehat{FDR}(W_1) = \frac{4}{10} > q \Rightarrow \text{Not satisfied}, \quad \widehat{FDR}(W_2) = \frac{2}{8} > q \Rightarrow \text{Not satisfied}$$

$$\widehat{FDR}(W_3) = \frac{3}{9} > q \Rightarrow \text{Not satisfied}, \quad \widehat{FDR}(W_4) = \frac{1}{4} > q \Rightarrow \text{Not satisfied}$$

$$\widehat{FDR}(W_5 \sim W_{10}) = 0 < q \Rightarrow \text{satisfied}, \quad \tau = W_5 = 0.5,$$

$\therefore X_5, X_6, X_7, X_8, X_9$ and X_{10} are selected.

Summary

Step 1) Feature importances $Z_i, \tilde{Z}_i$

- fit a model of y on $[X, \tilde{X}] \rightarrow$ Deep Pink model

Step 2) Feature statistics $W_i = f(Z_i, \tilde{Z}_i)$

Step 3) Feature selection procedure

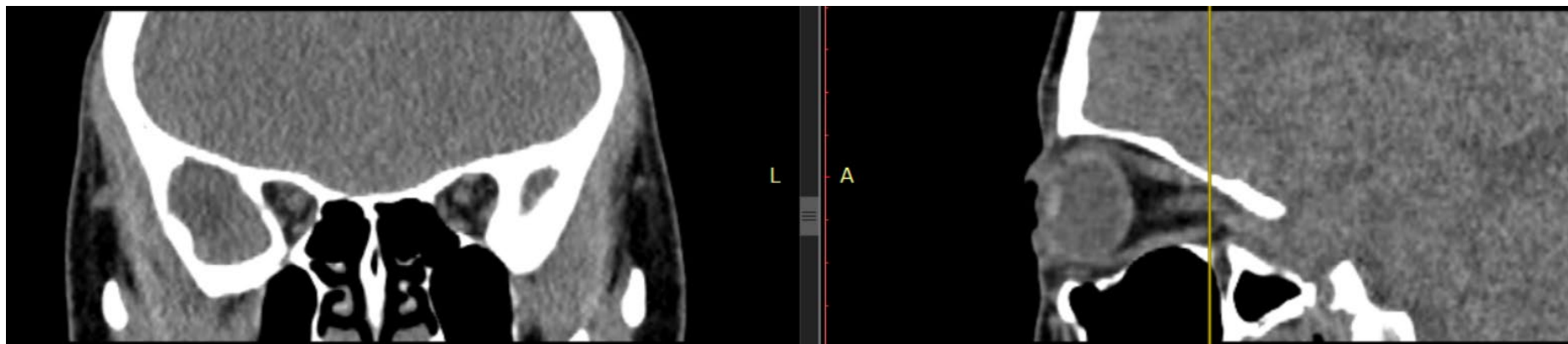
갑상선 연구 과제 보고

- 주요 업무: 1278명 환자에 대하여 84개 구조물 유형 세그멘테이션
- 세그멘테이션 완료 현황: 20 (개) / 84 (개) 구조물 유형
- 현재 진행하고 소목표: 파일럿 연구를 위한 데이터셋 확보
 - CO2와 CO3에 대해서 이미지 슬라이스 선택 현재 진행중
 - * CO2: 근육의 중간 지점
 - * CO3: 아래쪽 뼈의 V부분에 위치한 지점

갑상선 연구 과제 보고

- 파일럿 연구 진행 상황: 데이터셋 확보
- CO2: 시작과 끝 번호를 저장해서 1/2 하는 방식으로 번호 저장
- CO3: SA 뷰를 참고해서, 아래쪽 뼈의 V자 모서리에 해당하는 번호 저장

Patient Number	CT Date	시작	CO3	끝	CO2
101112	2017_01_31	110	122	131	121



갑상선 연구 과제 보고

- 파일럿 연구 진행 상황: 모델 설계 구상 (Sequential Forward Selection 기반)

초기화:

- 후보 집합 S를 84개* 구조물 타입을 포함하도록 초기화
 - Current 모델 M을 0모델로 초기화
- * 파일럿 연구에서는 40개

반복:

- S의 원소들을 Current 모델 M에 포함시켜 성능 확인
- S에서 1등 제거
- 1등을 Current 모델 M에 포함

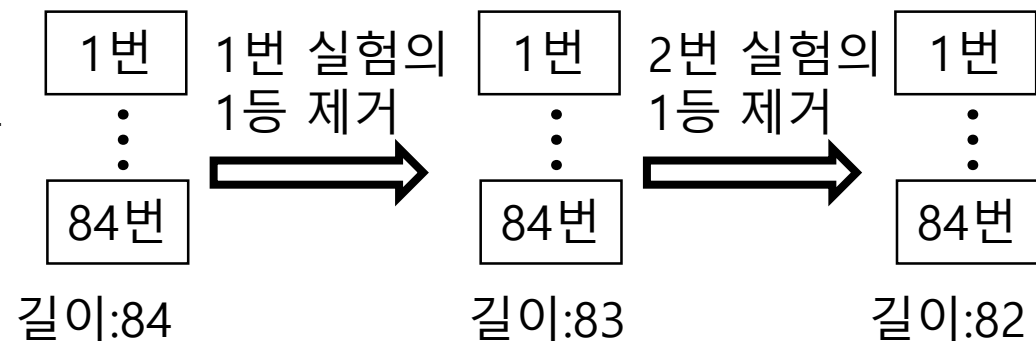
종료 조건:

- 성능 증가 정지

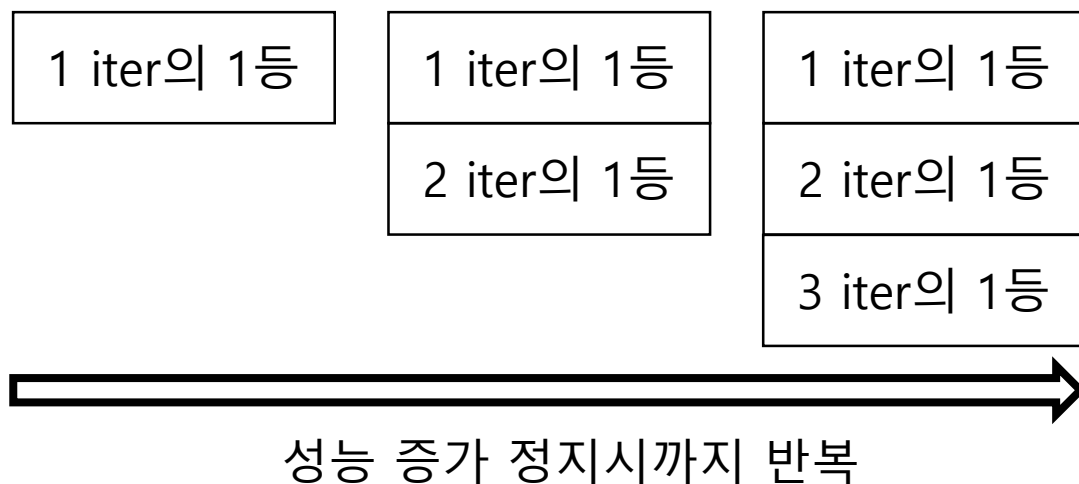
종료:

- Current 모델 M에서 마지막 요소 제거 후, M 출력

<후보집합 S>



<Current 모델 M>



갑상선 연구 과제 보고

- 파일럿 연구 진행 상황: 모델 설계 구상 (Sequential Forward Selection 기반)
- 예상 사용자 매개변수:
 - 1) 개별 성능 측정 방식:
 - 단일 뉴럴네트워크
 - 임의 개수의 뉴럴 네트워크 비교 방식
 - 뉴럴 아키텍처 서치를 이용한 최적화 방식
 - 2) 개별 실험 횟수:
 - 1번 실험
 - 임의 횟수 실험의 평균

갑상선 연구 과제 보고

- 파일럿 연구 진행 상황: 모델 설계 구상 (Sequential Forward Selection 기반)
- 예상 사용자 매개변수:
 - 3) 종료 조건:
 - 1번 반복에서 성능 증가 중지
 - 연속 2번 이상인 임의 반복 개수에서 성능 증가 중지
 - 4) 구조물 조합의 성능 평가시, 구조물 간의 평행
 - 자세한 사항은 과거 연구 검토시 진행

References

Barber, Rina Foygel, and Emmanuel J. Candès. "Controlling the false discovery rate via knockoffs." *Annals of Statistics* 43.5 (2015): 2055-2085.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995): 289-300.

Chen, Jiajie, Anthony Hou, and Thomas Y. Hou. "Some analysis of the knockoff filter and its variants." *arXiv preprint arXiv:1706.03400* (2017).

Sarkar, Sanat K., and Cheng Yong Tang. "Adjusting the Benjamini-Hochberg method for controlling the false discovery rate in knockoff assisted variable selection." *arXiv preprint arXiv:2102.09080* (2021).