



A Feature Selection Study for Medical Image deep learning

Sanghyuck Lee

2021. 04. 14.

Outline

- Introduction
- False Discovery Rate
- False Discovery Rate in Feature Selection
- Other methods for controlling False Discovery Rate

Introduction

- Barber et al. (2014):

Variable selection for linear model

→ Controlling False Discovery Rate among the selected variables

→ The knockoff filter

False Discovery Rate

- False Discovery Rate (FDR) in statistics: a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons.
- Multiple comparisons using Analysis of Variance (ANOVA):
 - 1) "The average of all groups is the same" or Not
 - 2) If not, which groups are different?

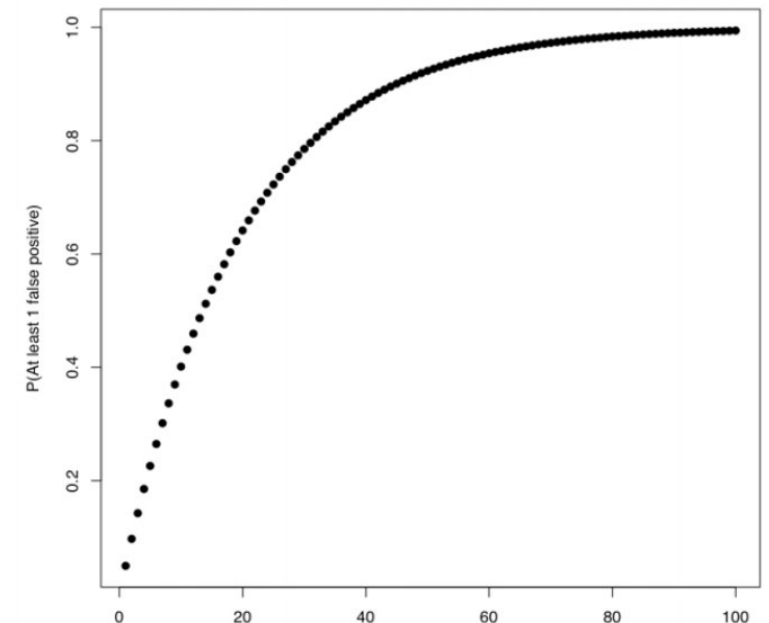
As the number of tests increases, the error increases

$$1 - (1 - \alpha)^m$$

m is the number of tests, α is significance level

$$FDR = \frac{\text{False Positive}}{\text{True Positive} + \text{False Positive}}$$

$$P(\text{Making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$



False Discovery Rate in Feature Selection

- rejected hypothesis -> selected feature -> likely belongs to the model

$$y = X\beta + z$$

y is a vector of responses, $X \in \mathbb{R}^{n \times p}$ is a known design matrix, $\beta \in \mathbb{R}^p$ is an unknown vector of coefficients and $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is Gaussian noise.

- The FDR is the expected proportion of falsely selected variables, a false discovery being a selected variable not appearing in the true model

$$\bullet \text{ } FDR = \frac{\text{False Positive}}{\text{True Positive} + \text{False Positive}} = \mathbb{E} \left[\frac{\#\{j: \beta_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j: j \in \hat{S}\} \vee 1} \right]$$

- $\#$: cardinality, $\vee$: max operator, $\hat{S}$: a set of selected features

Other methods for controlling FDR

- Permutation method
- The Benjamini–Hochberg procedure and variants
- Author's method: Using knockoffs

“The goal of the knockoff variables is to imitate the correlation structure of the original features in a very specific way that will allow for FDR control.”

Next week:

study other techniques and the history of FDR control in feature selection domain.

References

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