2023 STAT 528: Homework 4

Due Wednesday, February 22 at 11:59pm

1 Conformal Prediction

Question 1 (40 points)

Let X be a 50-dimensional random vector that is distributed according to $\mathcal{N}(0, \mathcal{I})$ and $y = 3 \tanh \beta^T X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$. Sample coordinates of β from a normal distribition $\mathcal{N}(0, 1)$ and fix this for the entire analysis. Now draw $n = \{50, 100, 200, 400\}$ iid samples of the pair (X, y) and consider the following analysis.

1. For 100 independent trials, generate the pair (X_{n+1}, y_{n+1}). Using the training data $\{(X_i, Y_i)\}_{i=1}^n$ and the test data X_{n+1}, use the split conformal prediction procedure (with a linear prediction model) to obtain a prediction interval for y_{n+1}. Compute the empirical probability of y_{n+1} belong in this prediction interval over the 100 independent trials. Report this for every n. Report also the average interval of the learned confidence interval (averaged across the 100 trials).
2. Repeat the previous step but replace the prediction model with a random forrest. How do your results change?

2 Multiple Testing

Question 2 (30 points)

In this question you will analyze data described in Efron and Hastie (2016). In a prostate cancer study, data were collected on 50 controls and 52 patients with cancer, with the genetic

activity being measured at $m = 6033$ genes. The investigators were hoping to spot *non-null genes*, ones for which patients and controls would respond differently.

At the website `https://web.stanford.edu/~hastie/CASI_files/DATA/prostz.txt` you will find data on 6033 z-scores. Under the null, these statistics follow an N(0,1) distribution.

Please provide the code used for these problems in-line.

2.1 Calculate the p-values corresponding to the z-scores, remembering that p-values should take into account extremes at both sides of the distribution. Form a histogram of these values. Be sure your histogram has an reasonable number of bins, is properly labeled, and generally follows the principles of good statistical graphics.

2.2 See which genes are flagged as significant under the procedures below. For each, clearly state/output the result (if more than 5 genes are flagged, feel free to simply output a list from R).

(a) Bonferroni controlling the FWER at 5%.

(b) Holm's procedure controlling the FWER at 5%.

(c) Benjamini and Hochberg's FDR control, apply with control at the 10% level.

(d) Control the expected number of false discoveries (EFD). Find the genes that correspond to EFD of 1, and then also with 5.

(e) Storey's q-values with a pFDR threshold of 10%. Also give the estimate of the proportion of null genes π_0. [Hint: the `R` package `qvalue` package, available at Bioconductor, implements this procedure. Installation code can be copied from `https://www.bioconductor.org/packages/release/bioc/html/qvalue.html`]

2.3 Write two paragraphs to accompany your results. In the first, briefly describe the problem and the methods you use in terms that are understandable to a non-statistician. In the second paragraph, critically evaluate the criteria, p-values you obtained in [2.1] and the multiple testing results you obtained in [2.2].

Null Hypothesis	p-value
H_{01}	0.0011
H_{02}	0.031
H_{03}	0.017
H_{04}	0.32
H_{05}	0.11
H_{06}	0.90
H_{07}	0.07
H_{08}	0.006
H_{09}	0.004
H_{10}	0.0009

Table 1: P-values for Question 3.

Question 3 (30 points)

Suppose we test $m = 10$ hypotheses, and obtain the p-values shown in Table 1.

1. Suppose that we wish to control the Type I error for each null hypothesis at level $\alpha = 0.05$. Which null hypotheses will we reject?

2. Now suppose that we wish to control the FWER at level $\alpha = 0.05$. Which null hypotheses will we reject? Justify your answer.

3. Now suppose that we wish to control the FDR at level $q = 0.05$. Which null hypotheses will we reject? Justify your answer.

4. Now suppose that we wish to control the FDR at level $q = 0.2$. Which null hypotheses will we reject? Justify your answer.

5. Of the null hypotheses rejected at FDR level $q = 0.2$, approximately how many are false positives? Justify your answer.