BMI 776

HOMEWORK 2 (2 LATE DAYS)

Noah Cohen Kalafut
Computer Science Doctoral Student
University of Wisconsin-Madison
nkalafut@wisc.edu

March 16, 2021

1 Problem 1

Done!

2 Problem 2

We are given

Pop A	CC	CG	GG	
Disease	26	62	56	144
Control	101	28	13	142
	127	90	69	286

2.1 Part A

2.1.1 χ_2^2 Test

We have the hypotheses

$$p_{null} = \text{Our SNP site is not associated with the disease.}$$
$$p_{alt} = \text{Our SNP site is associated with the disease.}$$

We can use the standard measure of statistical significance, $p < .05$.

Given that our expected value can be calculated with

$$E_{i,j} = \frac{T_{:,j} T_{i,:}}{T}$$

using $T_{i,:}$ as our total for row i.

We can compute the test statistic for our given data with

$$\chi_2^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Using T from the above table for population A and E from our equation,

$$E = \begin{bmatrix} 63.944 & 45.315 & 34.741 \\ 63.056 & 44.685 & 34.259 \end{bmatrix}$$

Then, we can use our observed values, O, from population A to produce

$$\chi_2^2 = \frac{(26-63.944)^2}{63.944)} + \frac{(62-45.315)^2}{45.315} + \frac{(56-34.741)^2}{34.741} + \frac{(101-63.056)^2}{63.056} + \frac{(28-44.685)^2}{44.685} + \frac{(13-34.259)^2}{34.259}$$
$$\approx 83.923$$

This provides a p-value 5.975×10^{-19}. As such, we have successfully rejected the null hypothesis.

Our SNP site, in all likelihood, is related to the disease within the population given.

2.1.2 Armitage Test

We have the hypotheses

$$p_{\textit{null}} = \text{Our SNP site is not associated with the disease.}$$
$$p_{\textit{alt}} = \text{Our SNP site is associated with the disease.}$$

We can use the standard measure of statistical significance, $p < .05$.

The Armitage test is given by

$$T = \sum_i t_i(T_{1,:}O_{0,i} + T_{0,:}O_{1,i})$$

For chosen weights t.

We can first look at the main portion of the equation

$$(T_{1,:}O_{0,i} + T_{0,:}O_{1,i}) = [\ -10852 \quad 4772 \quad 6080\ ]$$

If we choose our weights as $0, 1, 2$, searching for a linear trend, we get the test statistic

$$T = 16,932$$

We can then divide by the square root of the variance to get our Z-statistic.

$$\sigma^2 \approx 3767293.762$$
$$Z \approx 8.724$$

This score gives a p-value of approximately 2.685×10^{-18} which means we can safely reject the null hypothesis.

Our SNP site, in all likelihood, is related to the disease within the population given.

2.2 Part B

We have

Pop B	CC	CG	GG
Disease	457	462	466
Control	435	501	484

2.2.1 χ_2^2 Test

We have the hypotheses

$$p_{\textit{null}} = \text{Our SNP site is not associated with the disease.}$$
$$p_{\textit{alt}} = \text{Our SNP site is associated with the disease.}$$

We can use the standard measure of statistical significance, $p < .05$.

Performing our test, we get the test statistic $\chi_2^2 = 2.03$ and the p-value $p = .363$. From these, we fail to reject the null hypothesis. From this data, we cannot conclude that our SNP is related to the disease within the population given.

2.2.2 Armitage Test

We have the hypotheses

$$p_{\textit{null}} = \text{Our SNP site is not associated with the disease.}$$
$$p_{\textit{alt}} = \text{Our SNP site is associated with the disease.}$$

We can use the standard measure of statistical significance, $p < .05$.

Performing our test with the same weights as before, we get the test statistic $T = -55,085$, the Z-score $-.92$, and the p-value $p = .360$. From these, we fail to reject the null hypothesis. From this data, we cannot conclude that our SNP is related to the disease within the population given.

2.3 Part C

For this portion, we will combine our population tables

2.3.1 χ^2_2 Test

We have the hypotheses

$$p_{\textit{null}} = \text{Our SNP site is not associated with the disease.}$$
$$p_{\textit{alt}} = \text{Our SNP site is associated with the disease.}$$

We can use the standard measure of statistical significance, $p < .05$.

Performing our test, we get the test statistic $\chi^2_2 = 3.04$ and the p-value $p = .219$. From these, we fail to reject the null hypothesis. From this data, we cannot conclude that our SNP is related to the disease within the population given.

2.3.2 Armitage Test

We have the hypotheses

$$p_{\textit{null}} = \text{Our SNP site is not associated with the disease.}$$
$$p_{\textit{alt}} = \text{Our SNP site is associated with the disease.}$$

We can use the standard measure of statistical significance, $p < .05$.

Performing our test with the same weights as before, we get the test statistic $T = 120,549$, the Z-score 1.73, and the p-value $p = .084$. From these, we fail to reject the null hypothesis. From this data, we cannot conclude that our SNP is related to the disease within the population given.

2.3.3 Discussion

All tests using the results from population B failed to reject the null hypothesis, despite the results from Part C seeming more significant than those from Part B. This shows something interesting about our statistical tests. A failure to reject does not prove the opposite. Our SNP could be related to the disease with some confounding variable (which was likely isolated in population A) controlling the relationship. For example, the general population may not have more playing cards in proportion to their income, but a population consisting of magicians certainly could.

3 Problem 3

3.1 Part A

3.1.1 One-Layer Results

After the 99th epoch, the training and validation auPRCs were .846 and .780, respectively.

3.1.2 Two-Layer Results

After the 28th epoch, the training and validation auPRCs were .928 and .716, respectively.

The key to the higher accuracy (disregarding the overfitting), is in both the number of filters and the dual-layer system of the network. Firstly, the initial convolutional layer in our two-layer network has 15 filters rather than 10. Secondly, adding another convolutional layer allows us to more deeply and efficiently analyze our data. This ties into the core advantage of CNNs. By utilizing our knowledge about the arrangement of our data (i.e. It means something that entry 1,1 is next to entry 1,2), we can lessen the amount of training/luck required for our network to converge on a desirable solution without greatly increasing our computation time. This property is only accentuated with the use of more layers, until the point where our resolution (due to pooling) is too small for convolutional layers to yield proper cost/benefit. This point can vary depending on the network architecture and dataset.

3.2 Part B

3.2.1 General Questions

The input and output dimensions of our first Conv2D layer correspond to the shape of our data input and the shape after convolution. The dimensions $4,500 \rightarrow 15,486$ indicate that the convolution was done with no padding using 15 4x29x29 kernels.

The input and output dimensions of the dense layer is how our output is made comparable to the true data. After our convolutions and pooling, we end up with 195 features (input) that are all weighted and summed to make up our singular output (with an additional activation step).

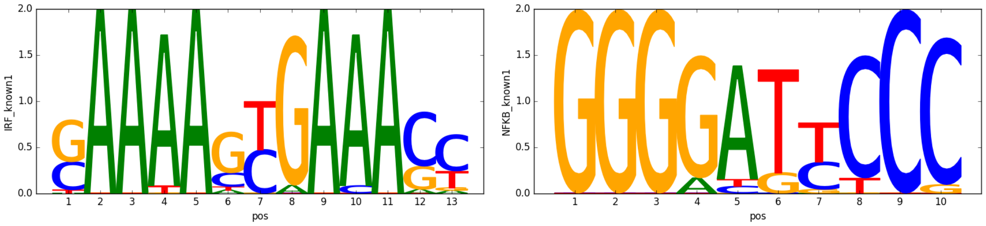
Using our chosen (5 each) positive and negative sequences, we get 4 true positives, 4 false positives, 1 true negative, and 1 false negative predicted. This corresponds to a high recall but low precision.

3.2.2 Filter Analysis

Looking at the true results in figure 1 and our results seen in figure 2, it can be seen that no one filter matches our motif. This is not unexpected, but makes the results less human-readable. A couple main concepts can be seen in the filters, however. Repeating bases, as can be seen most prominently in filter 6, are common in the filters, especially near the extrema – similar to our true motifs. Additionally, There are several instances where T will appear with high magnitude, negative/positive near the extrema and positive/negative (opposite) towards the center. It is hard to say if this is actually a learned trait, however.

It would be interesting to increase dropout and train, stopping prior to overfitting. In doing so, we could acquire a more accurate sense for what the network has actually learned.

Figure 1: True motifs



3.2.3 DeepLIFT

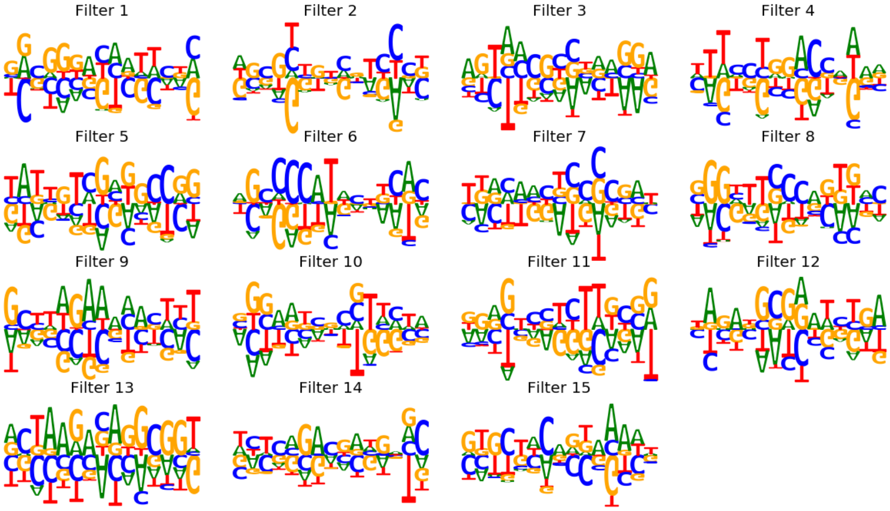
The DeepLIFT plots provide much clearer reconstructions of the given motifs. However, they mainly seem to reference the second motif (starting with GGG). This could be due to bias in the training data or this could be caused by the motif's surrounding elements generally being more distinct.

4 Problem 4

The plots produced can be seen in figures 3 and 4. The immediate observation is how conservative Bonferroni's method is in its estimations, only deeming 2 p-values significant for $\alpha = .05$. Controlling for the family-wise error rate is sound but sacrifices a lot of recall for precision.

The Benjamini-Hochberg correction procedure attempts to control for the false discovery rate (FDR). The procedure does this through the assumption that all samples are truly null, which will be nearly true in most cases. The procedure

Figure 2: Trained network first-layer filters



provides more significant values (39) than Bonferroni's method. This is because the procedure takes into account the number of samples being selected.

The Storey-Tibshirani correction procedure also attempts to control for the false discovery rate. Through a visual estimation of the proportion of truly null values in figure 4, the procedure predicts the FDR for each p-value threshold in the dataset. This results in a significant number more results than either of the previous methods (97).

For both the Benjamini-Hochberg and Storey-Tibshirani procedures, notice that the testing value ($P_k \frac{m}{k}$ and q-values) does not necessarily increase with the p-values. For this reason, the maximal p-value for which the testing value is below the chosen maximal FDR α is used as the p-value threshold.

The use of each procedure comes down to the application. In an application where extremely reliable data is needed, go with Bonferroni. If you want the most samples for the chosen FDR, Storey-Tibshirani is great. Somewhere in the middle or you have a very large sample? Benjamini-Hochberg would work well.

Figure 3: Bonferroni (Left) and Benjamini-Hochberg (Right) correction procedure results for $\alpha = .05$

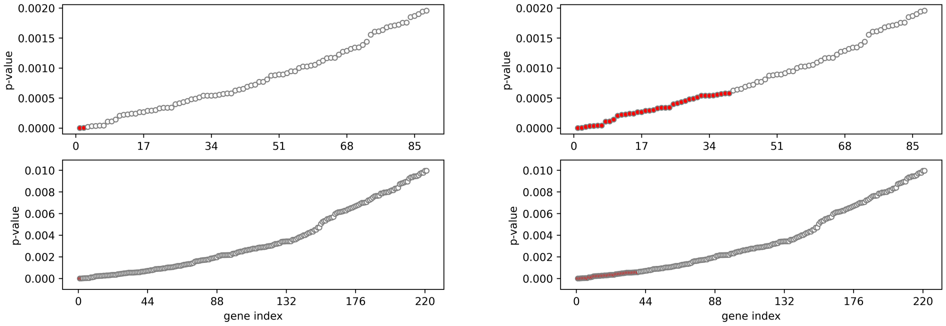
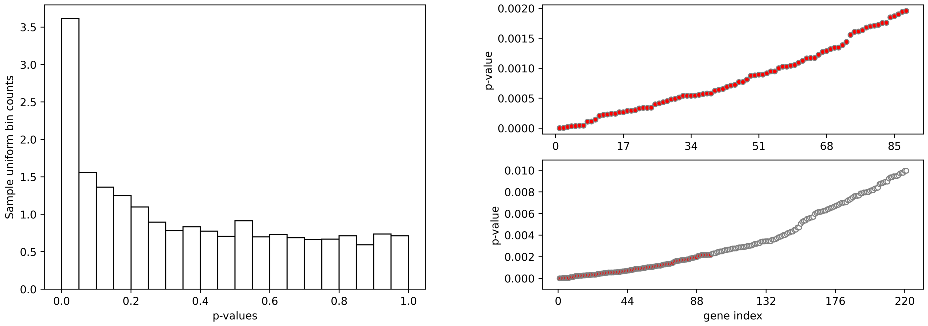


Figure 4: Storey-Tibshirani correction procedure histogram and results with $\lambda = .7$ and $\alpha = .05$


Sample uniform bin counts
p-values
p-value
p-value
gene index