Music Genre Classification

Abhinav Khanna
Student Researcher
akhanna@princeton.edu

Amandeep Singh
Student Researcher
email@princeton.edu

Abstract

1 Introduction

2 Related Work

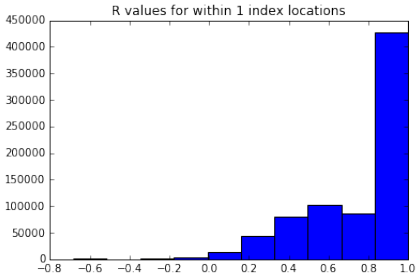
3 Methods

3.1 Description of data

The training data used in this paper comes from a set of 33 reference samples that have their entire genome bisulfite sequence. We primarily focus on the data for chromosome 1 for our analysis. The data consists of 33 reference samples with their methlyation proportions for each CpG site on the chromosome. Our test data consists of a sample tissue for which only a small subset of the methlyation values are accounted for. The first few columns of the training data specify the start location, the end location, and the positive or negative strand of the chromosome that the location is present on. We utilize this information to impute the values for the missing methlyation values in the test chromosome sample.

To better understand the relationships between different tissues, we computed the pearson r correlation coefficient and the corresponding p-value associated with that coefficient for each pair of tissues (Figure 2). As can be seen in table 1, certain tissues have a meaningful, non-pure chance, correlation with other tissues suggesting that certain tissues may be more apt at predicting given test values than other tissues. We also know from Figure 1 that locations that are closer together have higher correlations, and are more similar to each other than locations that are further apart on the chromosome. The finding that nearby locations have higher correlation is supported by other research in this area, such as the paper by Weiwei Zhang and Tim D Spector.

Figure 1: Pearson R Correlation for neighboring index locations



Examining the distribution of the mean methylation values across different tissues the distribution appears to have a right skew, and looks like it could be representative of a beta distribution. Because

Figure 2: Table of Correlation Values for Different Tissues

	0	1	2	3	4
0	(0, 0)	(0.712834867218, 0.0)	(0.757290076957, 0.0)	(0.708640674253, 0.0)	(0.377274968645, 0.0)
1	(0.712834867218, 0.0)	(0, 0)	(0.703314264614, 0.0)	(0.706602501721, 0.0)	(0.329787127457, 0.0)
2	(0.757290076957, 0.0)	(0.703314264614, 0.0)	(0, 0)	(0.738807452418, 0.0)	(0.351294390129, 0.0)
3	(0.708640674253, 0.0)	(0.706602501721, 0.0)	(0.738807452418, 0.0)	(0, 0)	(0.312640590975, 0.0)
4	(0.377274968645, 0.0)	(0.329787127457, 0.0)	(0.351294390129, 0.0)	(0.312640590975, 0.0)	(0, 0)

the mean methylation values appear like they are drawn from a beta distribution, we believe that attempting to model the similarity between our test tissue and the training tissues as a problem of judging which prior beta distribution the test sample is most likely to have been drawn from may be a worthwhile approach. The mean methylation values can be seen in Figure 2. Furthermore, plotting the methylation values across a given tissue by location yields a graph that does not look linear, suggesting that if a regressor is to be used, a non-linear one may yield better results (Figure 3).

Figure 3: Mean Methylation across 33 Tissues

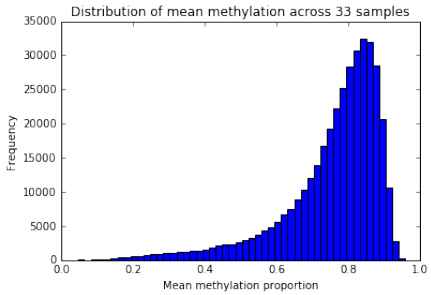
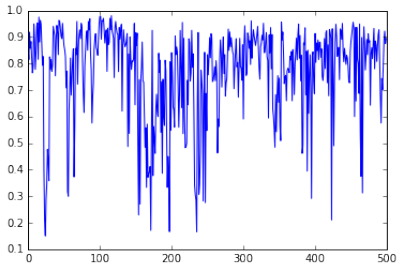


Figure 4: Methylation Values across a Single Tissue



3.2 Imputation Pipelines

The basic structure of our approach remained the same across pipelines. We began by calculating the similarity between the test tissue and the training tissues. The similarity was judged through some type of similarity metric, and different pipelines used different metrics. After measuring similarity, the pipelines combined the information from the most similar tissues to predict the values of the test tissue. The process of combining the information was also varied from pipeline to pipeline.

3.2.1 Similarity Metrics Tested

1. The probability of the test vector being drawn from the beta distribution prior of a given test tissue
2. Pearson's R correlation

3.2.2 Combination Methods Tested

1. Weighted mean - weights decided by similarity for some top x set of similar tissues
2. Support Vector Regression - Create SVRs for the top x set of similar tissues using their training data to train the regressor on each location
3. Decision Tree Regression
4. Lasso Regression

Due to time constraints, we were only able to test Pearson's R with weighted mean, and beta priors with SVRs and Decision Trees.

3.3 Evalulation

The effectiveness of each pipeline was judged by examining the Pearson R correlation of the filled in test vector to the full vector, computing the root mean squared error between the two vectors, looking at the residual errors, and graphing the predicted test vector against the full vector. Because we chose to only examine chromosome 1, there wasn't a great way to run cross validation testing on our prediction efforts, and without looking at the other chromosomes, we couldn't think of a way to guarantee that our methods generalize beyond chromosome 1.

4 Results

5 Discussion and Conclusion

Acknowledgments

We would like to acknowledge Princeton University for offering the wonderful COS 424 class that provided us with the opportunity to take on this research. We would also like to thank our Professor, Professor Englehart, and her TA team for providing us with guidance and the tools necessary to make this project happen.

References