Elvin Liu

Zhaoxing Wu

STAT 499

Autumn 2023

Classify High-Dimensional Data

During my time this quarter, I was given the opportunity to explore the classification problems posed by large feature small sample size datasets alongside the possible solutions. My mentor and I ended up focussing on a variation of Fisher's Linear Discriminant Analysis (FLDA), the Penalized Discriminant Analysis Projection Pursuit index (PDA PP-index). Before this directed reading program (DRP), I was unfamiliar with classification models as a whole, not even comprehending the concept of supervised models. I had also only just learnt Principal Component Analysis (PCA), so while the linear algebra behind FLDA was fairly straightforward, it was a little confusing to differentiate from PCA. Intuitively, the differences merely involve optimizing the between-class and within-class matrices rather than the covariance matrix. Hence, we are using pre-existing knowledge to separate classes as opposed to maximizing variance.

The earlier portions of the quarter were dedicated to gaining some degree of familiarity with classification and the curse of dimensionality. I read from the textbook *R in Action*, which showcased multiple libraries in R that could be used for classifying labels, utilizing Random Forests, Support Vector Machines, and Logistic Regression. To supplement this, I also read the paper, *High Dimensional Data Classification*, overviewing the classification errors that occur in the case of high feature count. In addition, the text went over the mathematics behind the

Discriminant Analysis functions and Support Vector Machines that mitigate the negative effects posed by high feature datasets.

In the latter half of my project, we covered the primary material, *A Projection Pursuit Index for Large P Small N*, which came attached with a GitHub page containing the paper's respective optimization functions in R. Thanks to my mentor's expertise on the subject, I understood how the penalty matrix in PDA better accommodated datasets with significantly more features than samples by correcting the data piling issue that can occur from the within-class scatter matrix. Using the provided package (from the GitHub page), I could project a well distributed simulation model using both FLDA and PDA PP-indexes to determine the linear combination of features that best separated classes. Graphical comparisons of their features revealed vastly different results, even though a penalty matrix was all that was added.

Later, I cleaned up and projected a real colon cancer dataset with multiple PDA lambda values. In comparison to support vector machines, the provided t-tests, and LDA, PDA better identified the most important features for classifying whether a tissue was normal or a cancerous tumor. In particular, its accuracy guessing true class labels on a test set (trained on training data) was leagues higher than anything else. Additionally, its 1 dimensional and 2 dimensional projections were better separated than FLDA's thanks to the penalty added to the within-class scatter matrix.

All in all, the DRP this quarter was an excellent brain teaser that I am grateful to have been part of. Nearing the beginning of my project, I was pretty lost on what exactly I was doing. Linear algebra, coding in R, rudimentary machine learning…? Little did I know the answer would be everything. I'd like to sincerely thank my mentor, Zhaoxing Wu, for her significant contributions to my learning and sage advice. My project would be nowhere near sufficient, at least quality-wise, without our weekly conferences discussing the direction of the project and the check-ups on my mathematical understanding.