INTRODUCTION

Dimension Reduction techniques are widely used on large datasets in statistics and data mining fields. Implementing these methods can:

- Reduce dimensionality of data while still capturing the most important features in the data
- Improve accuracy of data classification
- Reduce computational runtime
- Allow visualization of data in lower dimensions.

This research focuses specifically on Kernel Linear Discriminant Analysis (KLDA). Understanding the theoretical background behind KLDA and statistical classification power help in optimal tuning parameter selection, which is still an on-going challenge. Selecting optimal tuning parameters can yield well-separated classes without over-fitting the model.

BACKGROUND THEORY

Let $\phi$ be a non-linear mapping of the input to a higher-dimensional space. Define a $p \times \ell$ dimensional variable $X = [x_1, x_2, \ldots, x_p]$, where $\ell$ is the number of the number of observations in $C$ classes and $\ell_i$ is the number of observations in the $i^{th}$ class. We try to find $w$, which is a linear combination of the $X$, such that among the projections of $X$,

$$Y = w^T \phi(X),$$

the between-class variance is maximized and the within-class variance is minimized. In the case of a two-class problem, to accomplish this, we define an objective function in which we aim to maximize:

$$J(w) = \frac{w^T S^B w}{w^T S^W w},$$

where the mean response vectors, between-class scatter matrix, and within-class scatter matrix are defined as:

$$m^B = \frac{1}{\ell} \sum_{i=1}^{C} \phi(x_i),$$

$$S^B = (m^B_1 - m^B)(m^B_2 - m^B)^T,$$

$$S^W = S_{w_1} + S_{w_2} = \sum_{x \in C_1} (\phi(x) - m^B_1)(\phi(x) - m^B_1)^T + \sum_{x \in C_2} (\phi(x) - m^B_2)(\phi(x) - m^B_2)^T.$$

The expression for $J(w)$ can be put in terms of kernel functions to make computing run much faster. The Gaussian (RBF) kernel is used in this research which is of the form:

$$\exp \left[ -\frac{|x-y|^2}{2\sigma^2} \right].$$

This is where the tuning parameter that we wish to optimize is introduced.

APPLICATION OF KLDA

KLDA was implemented on four toy datasets to assess how well the classes from each data set were being separated.

APPLICATION OF SVM

From SVM, two error rates are defined to help investigate how well KLDA is performing given different $\sigma$ values.

$$L_1 = \text{Classification error rate of SVM fit}$$

$$L_2 = \text{Relative distance of support vectors against all data points to linear SVM hyperplane}$$

We now create a loss function that we aim to minimize:

$$L = L_1 + L_2.$$

A function was created in R to take in $\sigma$ values and output the error $L$. To understand how the loss function is working, the KLDA projections were plotted and compared among their loss function values.

ACKNOWLEDGEMENTS

Thank you to the NSF Foundation for the grant that allowed the REU students to participate in the Statistical Data Mining and Machine Learning REU this summer. Thank you to my collaborators of the Dimension Reduction Group: Lynn Huang, Jackson Maris, and Ryan Wood. A special thanks Dr. Yishi Wang and Dr. Cuixian Chen at University of North Carolina at Wilmington for their time and support.

REFERENCES

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning, volume 112. Springer.