Cameron Jester
STAT 338 – Probability
December 11th, 2023

Comparative Analysis of Dimensionality Reduction Techniques: A Study on Quadratic Discriminant Analysis, Linear Discriminant Analysis, and Naive Bayesian Classifier for Improved Classification in Machine Learning

Background

High dimension data is often defined as data sets with more variables than observations and can be associated with several downsides such as computational costs, overfitting, and often high correlations. Additionally, N-dimensional data can prove to be challenging to visualize. Minimizing dimensions in a data set can be done through dimensionality reduction. Dimensionality reduction through feature selection aims to eradicate the abovementioned issues and redundancies.

Today, several types of Dimensionality reduction techniques are employed by data scientists, data engineers, and statisticians alike. Quadratic Discriminant Analysis, Linear Discriminant Analysis, and Naive Bayesian Classifier are among the most popular methods of performing reduction. Each of the forms above has its strengths and weaknesses. This paper aims to investigate and compare the properties and effectiveness of these techniques for classification in machine learning.

A real-world application of dimensionality reduction can be found with image processing. Facial recognition is a particular example of a dataset in N-dimensional space. Facial recognition is an aspect of image processing that has been used since the early 1960's and began with a simple computer scanner that recognized a person's face by their hairline, eyes, and nose. This kind of image processing is becoming more advanced with the help of machine learning. The features on a human face are incredibly vast, and using dimensionality reduction allows certain features to be used when creating data on a human face. Decreasing the dimensions of a human face has several real-world advantages, such as making phone scanners recognize faces more quickly, allowing for more accurate social media filters, and even finding people who have committed a crime and were caught on camera. For computational advantages, dimensionality reduction allows for noise reduction, such as 'ignoring' a non-permanent skin blemish, using less space on a computer's storage, and quicker computations.

A publication in the Journal of Computational Intelligence and Neuroscience indicated that the method commonly used for feature selection of human facial recognition is linear discriminant analysis (LDA) (Computational Intelligence and Neuroscience, 2022). In the publication, 128 deep features of a human face were captured, and LDA was conducted to reduce the dimensions to three feature sets. The three feature sets allow the computer to predict someone's emotion based on the shape and features of their expression. Dimensionality reduction is essential for this and similar studies as it will enable models to make more accurate predictions, which can, in turn, improve lives as technology becomes more ingrained into society.

Probability Distribution

Decision boundaries are continuous paths, either straight or curved, that separate classes of features that models deem to be statistically different from other classes in the data. Separating data into boundaries allows humans to visualize how data will be put into separate classifiers and allows for inferential statistics to be performed through feature extraction. For example, if a model were to break up facial images into ‘smiling’ or ‘not smiling’ categories, below the decision boundary may be faces that are predicted to be ‘not smiling’ and vice versa. The names given to the Quadratic Discriminant Analysis and the Linear Discriminant Analysis provide insight into their decision boundaries, which are quadratic and linear, respectively. Visually, this means a straight line would be present for LDA, and for QDA, it would be more curved. Decision boundaries of the Naïve Bayesian Classifier can resemble the following shapes in two-dimensional space: line, circle, or parabola, and in more than two-dimensional space, it can be a plane, ellipse, or parabolic curve. Optimizing decision boundaries is the choice between machine learning classification models. Hence, picking a classifier will depend on the data set given to the model while accounting for the drawbacks and advantages of each classifier.

Linear algebra plays an important role when computing using the Multivariate Gaussian Distribution. Several key concepts, such as matrices, eigenvalues, and vectors, directly impact the models discussed in this paper. A covariance matrix summarizes the covariance between all pairs of variables in a dataset. This matrix type is a critical component to understanding the Multivariable Gaussian Distribution.

Properties

The equation to calculate a covariance between pairs of variables is as follows:

$$\mathrm{Cov}[X,Y]=E[(X-E[X])(Y-E[Y])]=E[XY]-E[X]E[Y].$$

The following properties can be applied to the above equation:

$\mathrm{Var}(aX)=a\mathrm{Var}(X)$
$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$
If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$
$\mathrm{Cov}(AX,BY)=A\mathrm{Cov}(X,Y)B^{T\,*}$
$\mathrm{Cov}(X+a,Y+b)=\mathrm{Cov}(X,Y)$

Where the expected value of E[X] and E[Y] can be calculated by:

For discrete random variables: $E(X) = \Sigma\ x(p(x))$

For continuous random variables: $E(X) = \int$ -infinity to +infimity x*f(x)dx

The values calculated by the equations above are then put into a matrix to identify variables with a high covariance, which indicates a strong relationship between variables. It is required for both QDA and LDA for the number of predictor variables to be less than the sample size of the data;

proposed guidance is for the number of observations to be five times the number of predictor variables, ex) 5 predictor variables = 25 observations. When comparing the variance tolerance of the models, the LDA is stricter in its assumptions. While this might lead to a higher prediction metric, it may lead to high model bias, especially if the training set is large, and therefore higher variance is expected. Due to QDA's more lenient variance assumptions, it would be a more suitable model for large training sets.

Conversely, if the provided training set is smaller and limiting variance is crucial for classification, the LDA would be an appropriate model. Variance assumptions in the NBC are nested within the conditional independence requirement of this model, as when variables are independent, their covariance is zero. Regardless of the training set size, it is vital to recognize the limitations of assuming independence when working with real-world data.

It is essential first to discuss the similarities between the LDA and QDA. These models both assume that the predictor variable in the model comes from a Multivariable Gaussian Distribution. Comparing the Naïve Bayesian Classifier to the LDA and QDA, the assumption of multivariate normal distribution is the same at the feature level and the target class level. The Multivariable Gaussian Distribution has the following probability density function given the population mean vector = E(X) and covariance matrix = Var(X):

$$p(\vec{x}) = (2\pi)^{-p/2} \det \Sigma^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

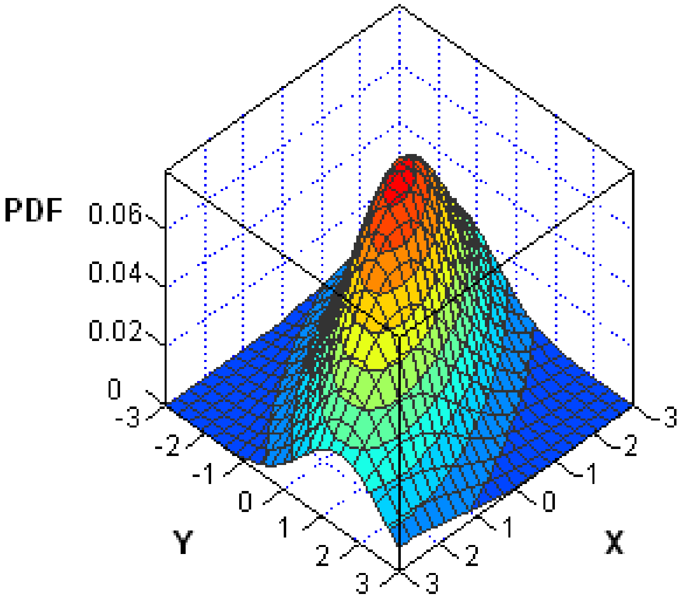
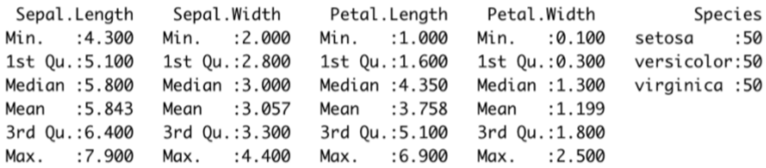


Figure 1. Visual of Multivariate Gaussian Distribution (SaS blog, 2012)

For a Multivariable Gaussian Distribution, a scatterplot matrix is typically used to model pairwise relationships between multiple variables in the dataset. Ellipsoids are used to represent the probability density function of the model. A heatmap is then applied to demonstrate areas of high density. Using Figure 1 as an example, the red at the top of the shape represents an area of high probability density. In contrast, the blue represents a lower probability between the variables on the XY plane.

Real-World Application

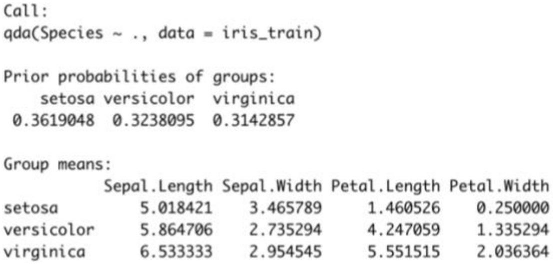
The Iris dataset found in Base R provided a conceptual understanding of the model comparison between the three methods mentioned. This dataset was created in 1936 by Ronald Fisher, a British statistician, and is now used to model machine learning classification processes. There are 150 (N=150) observations in this dataset. There are five columns, four of which are numerical (continuous), and one is categorical (discrete). The numerical variables are 'Sepal Length,' Sepal Width,' 'Petal Length,' and 'Petal Width.' The categorial variable is 'Species'; this dataset has three species: Setosa, Versicolor, and Virginica, with 50 observations per species. Below is a summary of the data to be explored before fitting a model to the data.



```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

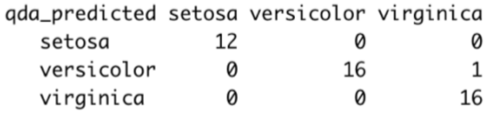
Table 1. Iris dataset statistical summary

Given the history and useability of this dataset, data cleaning is not necessary, although other data sets may require factorization and data manipulation prior to implementing this first step. Preparing the data to be classified includes creating a testing and training set using the 70-30 method, which allocates 70% of observations to training and the remaining 30% to testing. First testing the Quadratic Discriminant Analysis. Secondly, a confusion matrix is calculated, and accuracies can be calculated by hand with a small number of variables. Finally, a visualization is created. This process is then repeated for Linear Discriminant Analysis as well as Naïve Bayesian Classifier, with an excpetion with the final step.



```
Call:
qda(Species ~ ., data = iris_train)

Prior probabilities of groups:
    setosa versicolor  virginica 
 0.3619048  0.3238095  0.3142857 

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         5.018421    3.465789     1.460526    0.250000
versicolor     5.864706    2.735294     4.247059    1.335294
virginica      6.533333    2.954545     5.551515    2.036364
```

Figure 2. Quadratic Discriminant Analysis model



```
 qda_predicted setosa versicolor virginica
    setosa         12          0         0
    versicolor      0         16         1
    virginica       0          0        16
```

Table 2. Quadratic Discriminant Analysis Confusion matrix with .977 accuracy

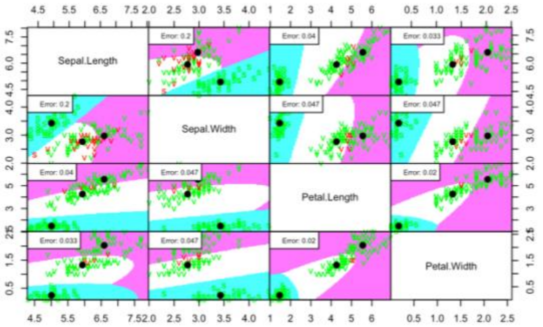
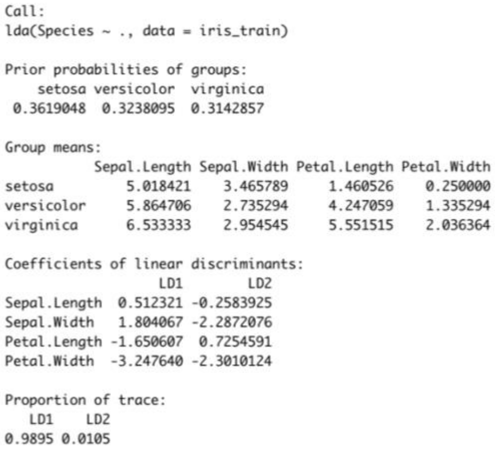
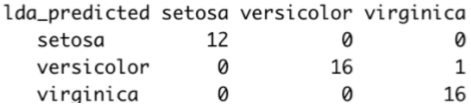


Figure 3. Quadratic Discriminant Analysis model visualization



```
Call:
lda(Species ~ ., data = iris_train)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3619048  0.3238095  0.3142857

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         5.018421    3.465789     1.460526    0.250000
versicolor     5.864706    2.735294     4.247059    1.335294
virginica      6.533333    2.954545     5.551515    2.036364

Coefficients of linear discriminants:
                   LD1        LD2
Sepal.Length  0.512321 -0.2583925
Sepal.Width   1.804067 -2.2872076
Petal.Length -1.650607  0.7254591
Petal.Width  -3.247640 -2.3010124

Proportion of trace:
   LD1    LD2
0.9895 0.0105
```

Figure 4. Linear Discriminant Analysis model



```
lda_predicted setosa versicolor virginica
   setosa         12          0         0
   versicolor      0         16         1
   virginica       0          0        16
```

Table 3. Linear Discriminant Analysis Confusion matrix with .977 accuracy

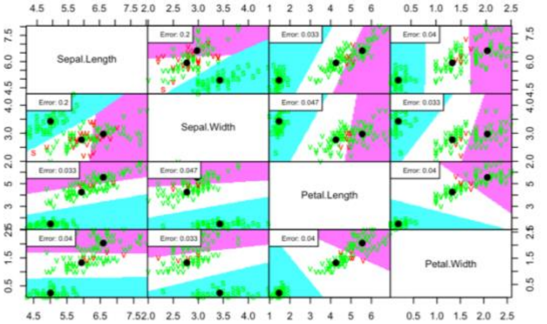
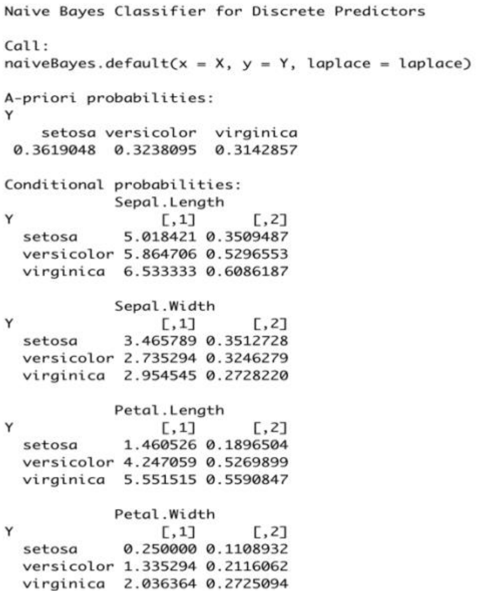
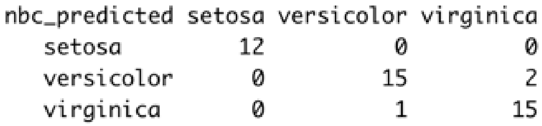


Figure 5. Linear Discriminant Analysis model visualization



```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
    setosa versicolor  virginica 
 0.3619048  0.3238095  0.3142857 

Conditional probabilities:
            Sepal.Length
Y                [,1]      [,2]
  setosa     5.018421 0.3509487
  versicolor 5.864706 0.5296553
  virginica  6.533333 0.6086187

            Sepal.Width
Y                [,1]      [,2]
  setosa     3.465789 0.3512728
  versicolor 2.735294 0.3246279
  virginica  2.954545 0.2728220

            Petal.Length
Y                [,1]      [,2]
  setosa     1.460526 0.1896504
  versicolor 4.247059 0.5269899
  virginica  5.551515 0.5590847

            Petal.Width
Y                [,1]      [,2]
  setosa     0.250000 0.1108932
  versicolor 1.335294 0.2116062
  virginica  2.036364 0.2725094
```

Figure 6. Naïve Bayesian Classifier model



```
nbc_predicted setosa versicolor virginica
   setosa         12          0         0
   versicolor      0         15         2
   virginica       0          1        15
```

Table 4. Linear Discriminant Analysis Confusion matrix with .933 accuracy

Conclusion

The experiment above uses the Iris dataset to model the differences between Quadratic Discriminant Analysis, Linear Discriminant Analysis, and Naive Bayesian Classifier yielded insightful outputs demonstrating these models' differences. The QDA and LDA models had the same outputs when applied in RStudio, with an accuracy of .977 (~98% accuracy). Both models incorrectly predicted a Virgininca flower. Comparatively, Naïve Bayesian did not perform as well compared to the other two reduction methods, but it is important to mention that a 93% accuracy is not a poor model. While it is unlikely that one would get a result that is the exact same between QDA and LDA, this dataset provides an insight into how classes are determined. When faced between two similar models, it is critical to look at the covariance matrix of the dataset to determine if a more lenient model, such as LDA that assumes the same covariance between classes, and conversely with the QDA.

Distribution Exercise

A distribution exercise showcasing understanding of Multivariate Gaussian Distribution would include covariance matrices. Students can create three separate covariance matrices to model independent, positive, and negative correlations between variables. Following this, students could plot the relationships above and begin to play around with how difference covariance will impact the visualizations of the data.

References

Bayes classifier and naive bayes. Lecture 5: Bayes classifier and naive bayes. (n.d.). https://www.cs.cornell.edu/courses/cs4780/2021fa/lectures/lecturenote05.html

Do, C. B. (2008, October 10). The Multivariate Gaussian Distribution. Stanford.

Facial Recognition History. Thales Group. (n.d.). https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/inspired/history-of-facial-recognition#:~:text=Facial%20recognition%20is%20more%20than,computer%20was%20to%20find%20matches

The Iris dates a little bit of history and biology. Medium. (n.d.). https://towardsdatascience.com/the-iris-dataset-a-little-bit-of-history-and-biology-fb4812f5a7b5

Lavrenko, V. (2014, January 15). *IAML5.10: Naive Bayes decision boundary*. YouTube. https://www.youtube.com/watch?v=0oca6pC3f0M

Linear & quadratic discriminant analysis. Linear & Quadratic Discriminant Analysis · UC Business Analytics R Programming Guide. (n.d.). https://uc-r.github.io/discriminant_analysis

Sicotte, X. B. (2018, June 22). *Xavier Bourret Sicotte*. Linear and Quadratic Discriminant Analysis - Data Blog. https://xavierbourretsicotte.github.io/LDA_QDA.html

Zheng, F. (2022, August 29). *Facial expression recognition based on LDA feature space optimization*. Computational intelligence and neuroscience. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9444369/