Project 2 - Classification with Dimensionality Reduction and Performance Evaluation

Elliot Greenlee
egreenle@vols.utk.edu

ECE 571 Pattern Recognition
March 12, 2017

Abstract

This report details the implementation of parametric and nonparametric decision rules. Discriminant function and k-Nearest Neighbor classification are applied to a two-category medical data set, before and after dimensionality reduction. The data is first assumed to be normally distributed, and maximum likelihood estimation is used to estimate the parameters of the Gaussian. Dimensionality reduction is performed by employing Fisher's Linear Discriminant and Principal Component Analysis. The performance of each classification method was compared, and within-method performance was evaluated across the range of prior probabilities and method parameters. By manipulating these values, an accuracy rate as high as 80.4% can be found.

1 Introduction

Humans make immediate discriminations between categories constantly throughout the day. These classifications are simple to a human, but would be difficult to implement on a computer. However, computational power also allows for more nuanced classification of features not easily understood by a human. In the case of our dataset, diabetes information for the Pima indians, features include age, number of times pregnant, and diastolic blood pressure.[2] Pattern classification methods are able to use these numeric values for discrimination between classes in a way human ability cannot.

In Bayesian theory, one calculates the posterior probability of an input belonging to a particular class by taking the product of prior probability and likelihood of that event occuring and dividing it by the evidence. Methods based on this expression are a popular solution in pattern recognition.

$$P(w_j|x) = \frac{p(x|w_j)P(w_j)}{p(x)}$$

Classes are chosen based on the highest a-posteriori probability $P(w_j|x)$, calculated using the a-prior probability of a given class $P(w_j)$, the evidence $p(x)$, and the probability distribution function for the feature with respect to the given class $p(x|w_j)$. [1]

There are two options for estimating the probability distribution of features: parametric methods and nonparametric methods. In parametric methods, the distribution is assumed to be known; for example, we assume a normal Gaussian distribution and use various simplifications of the discriminant function. In nonparametric methods, there is no assumption of the form of the distribution. In our case, k-nearest neighbor is used to estimate the distribution.

Although in theory a greater number of statistically independent features should reduce further and further the error, in practice additional features often lead to worse performance. This occurs either when the wrong model is used, or because the training sample is not infinite, so the distributions are not accurately estimated. This leads to problems with overfitting, where the model works well on the training data but poorly on the testing data, and complexity, where the model takes longer to run at $O(d^2n)$ where d is dimension and n is number of samples. The solution to these problems is dimensionality reduction. Linear methods are used because of their simplicity, projecting high dimensional data onto a lower dimensional space. Two common methods are principal component analysis, with the objective of minimizing information loss in a least-squares sense, and Fisher's Linear Discriminant, with the objective of minimizing discrimination error in a least-squares sense.

The objective of this project was to implement parametric and nonparametric decision rules, and apply them to various data reductions. Overall this gave the opportunity to explore theories of pattern recognition in practice. In this project, classification was performed on the Pima indians diabetes dataset, and a classification accuracy of 80.4% was achieved.

2 Technical Approach

2.1 Parameter Estimation

The mean and variances of the data set before and after reduction were calculated using maximum likelihood estimation. This method maximizes likelihood of of Θ with respect to a set of samples D in the formula

$$p(D|\Theta) = \prod_{k=1}^{n} p(x_k|\Theta)$$

It is generally easier to work with the logarithm of this formula. Once the parameters $\Theta = [u \;\; \Sigma]$ are found, they are substituted into the Gaussians.

Prior probability was calculated directly from the training samples, taking the number of samples of a given class over the total number of samples.

2.2 Dimensionality Reduction

Two methods for dimensionality reduction were used in this experiment, each with its own objective.

Fisher's Linear Discriminant attempts to reduce the data to one dimension while minimizing discrimination error. One might imagine the scenario of rotation a projection line near two dimensional data; Fisher's Linear Discriminant would select the rotation that best discriminates the data upon projection. Because it is reducing based on class information, it is a supervised method. The projection direction is obtained by making the data sets as far apart as possible. This occurs when the objective function

$$\frac{w^T S_B w}{w^T S_W w}$$

is maximized, where S_B is the between class scatter and S_W is the within class scatter. This maximum projection occurs at $w = S_W^{-1}(m_1 - m_2)$

Principal Component Analysis attempts to reduce the data while minimizing information loss. Typically a maximum allowable loss is chosen, in our case 10%. One might imagine the major and minor axis of a set of two dimensional elliptical data. Principal component analysis would choose the major axis as the project direction to preserve information. Because there is no consideration of class information, it is an unsupervised method. The projection vector is obtained using the Karhunen-Loeve theorem, by discarding the minimum orthogonal vectors. This is possible by obtaining the eigenvectors and values of the covariance matrix, and then discarding the smallest eigenvalues and corresponding vectors. This minimizes the squared error.

2.3 Classification Methods

In our solution, we take two approaches to the probability distribution function: one parametric, and one nonparametric.

The parametric method, discriminant function, assumes a Gaussian distribution. Minimum error rate classification can be achieved through use of the discriminant function

$$g_i(x) = \ln p(x|w_i) + \ln P(w_i)$$

Given two discriminant functions $g_0(x)$ and $g_1(x)$ the larger of the two posterior probabilities chooses the correct class. Each of the cases is a variation on this formula.

Case 1: The simplest case occurs when every feature is statistically independent and has the same variance σ^2. This correlates to the assumption that the data falls into equal-size hyperspherical clusters around the means. The equation simplifies to

$$g_i(x) = \frac{(x - \mu_i)^t (x - \mu_i)}{2\sigma^2} + \ln P(w_i)$$

This is plugged into the general discriminant function equation to determine the class. The variance was calculated by averaging every element of the two covariance matrices.

Case 2: Another simplified case occurs when the covariance matrices are equal for both classes. This correlates to the assumption that the samples fall into hyperellipsoid clusters of equal size and shape around the means. The equation now only simplifies to

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \Sigma^{-1} (x - \mu_i) + \ln P(w_i)$$

This is plugged into the general discriminant function equation to determine the class. The covariance matrix was calculated by averaging the two covariance matrices.

Case 3: In the general multivariate normal case each covariance matrix is different. This correlates to any normal distribution of data. The full equation is

$$g_i(x) = -\frac{1}{2}x^t\Sigma_i^{-1}x + (\Sigma_i^{-1}\mu_i)^t x - \frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

The nonparametric method, k-nearest neighbor, determines the class of a given sample by evaluating the classes of the k nearest training samples. This distance is measured in a Euclidian sense across all features. The posterior probability is calculated just as in other Bayesian methods, however $p(x|w_m) = \frac{k_m}{n_m V}$ is the probability distribution function, $P(w_m) = \frac{n_m}{n}$ is the prior probability, and $p(x) = \frac{k}{nV}$ is the evidence.

3 Experiments and Results

3.1 Data

The data for these experiments comes from information about diabetes among the Pima indians. Both the training and testing datasets are obtained from Ripley's Pattern Recognition and Neural Networks [2]. There are 200 training samples and 332 testing samples, each with seven features and a class denomination. The training set is used to estimate Gaussian parameters and run k-nearest neighbor classification.

3.2 Comparison of Different Classification Methods

The first experiment was a comparison of the different classification methods with a prioritization of accuracy. Prior probability is determined by the training set, for fair comparison. Case 2 of the discriminant function provided the highest accuracy among the classification methods across all three reductions.
Table 1. Classification methods across various data reductions.

	Case 1	Case 2	Case 3	kNN
Normalized	74.1%	80.4%	76.5%	70.5%
Fisher's Linear Discriminant	73.5%	78.6%	73.5%	69.0%
Principle Component Analysis	80.4%	80.4%	80.4%	72.0%

3.3 Effects of Dimensionality Reduction

The second experiment was a comparison of the classification accuracy across various data reductions. Principle component analysis had the highest overall accuracy, followed by the normalized data, followed by Fisher's Linear discriminant. It appears that for this data set, some of the features reduced the accuracy of the model, but principle component analysis removes these features. See Table 1 for the full data.

3.4 Other

As the k value for k-nearest neighbor is adjusted, the accuracy varies. A typical value of k ranges to $\sqrt{n}$ where n is the number of samples. In this experiment we varied k from 1 to 20. The most accurate k values were 12, 12 and 15 for the normalized, pca reduced, and fld reduced date, respectively.

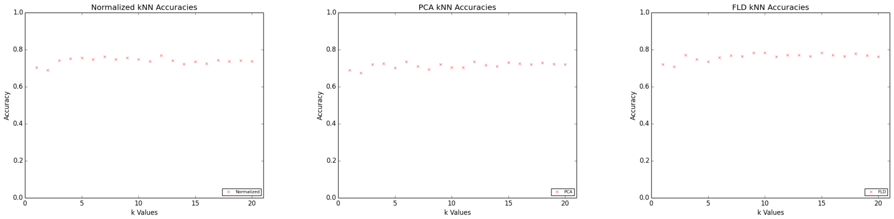

Figure 1. Accuracies as the k value is varied.

As the prior probability is varied for each of the four classifiers on each data reduction method, performance varies. Below are the ROC curves showing performance, arranged by data reduction method. The k that generates the best accuracy is used. Through these graphs, it can be seen that case 2 provides the highest accuracy across the range of prior probability.

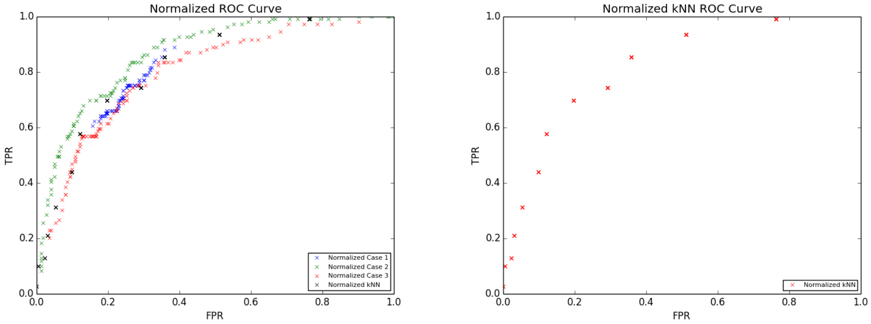

Figure 2. ROC curves as prior probability is varied for normalized data. knn data is produced twice for visibility.

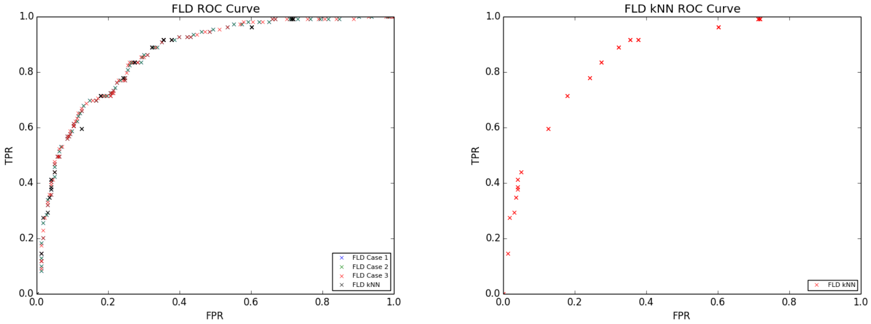

Figure 3. ROC curves as prior probability is varied for FLD reduced data. knn data is produced twice for visibility.

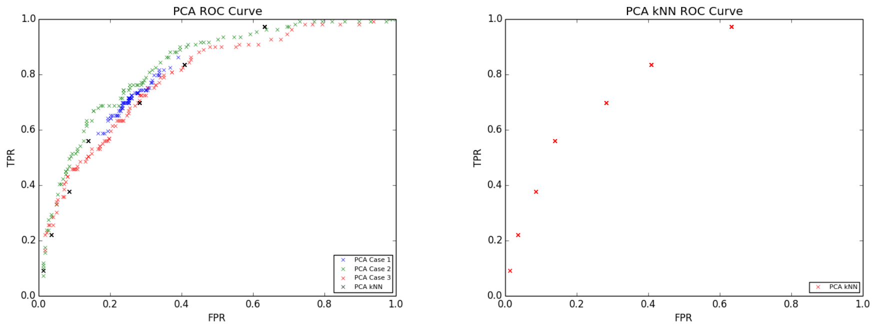

Figure 4. ROC curves as prior probability is varied for PCA reduced data. knn data is produced twice for visibility.

4 Summary

Two classification rules are implemented, the parametric discriminant function and the nonparametric k-Nearest Neighbor. Each is applied to a two-category diabetic classification problem, before and after reduction by Fisher's Linear Discriminant and Principal Component Analysis. Overall, an accuracy of 80.4% can be obtained using principal component analysis and case 2 of the discriminant function.

The objective of this project was to implement parametric and nonparametric decision rules, and apply them to various data reductions. I was able to more fully understand the practical applications of various data reduction methods, as well as the difference between parametric and nonparametric classification methods. Programming wise there were no serious challenges with this project. In the future, more experiments can be done with this data set, and more rigor can be introduced. Because runtime should be reduced by the dimensionality reduction, a comparison of timings after various reductions would be interesting. Cross validation could be used to better test the data and provide more precise results.

References

[1] Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern Classification. Second ed. New York: Wiley, 2001. Print.

[2] Ripley, B. D. Pattern Recognition and Neural Networks. N.p.: Cambridge University Press, 1996. Print.

5 Appendix

All code available at https://github.com/LambentLight/DimensionalityReduction, which includes a small README describing each file.