Project 1 - Two Category Classification Using Baysian Decision Rule

Elliot Greenlee
egreenle@vols.utk.edu

ECE 571 Pattern Recognition
February 13, 2017

Abstract

This project applied Bayesian parametric binary classification to a data set over two continuous features. In this investigation, multiple algorithms were examined for classification accuracy. The performance of a decision rule derived from the likelihood ratio was used as a baseline; experiments comparing the three versions of the discriminant rule, Gaussian and Bimodal models, and the effects of prior probability were run to determine highest accuracy. In this project, the highest accuracy achieved was by the Bimodal model, at 91.3%.

1 Introduction

1.1 Background

Pattern recognition systems are used throughout many different applications in order to solve natural problems. In each system, given data must be examined in order to find representative features. These features are used to generate a model of the natural system, a classifier, which is used to make some decision on new, unknown cases. This parametric method simplifies data to a known function. In the case of this project, the model is supervised, meaning that the training examples received are already marked with a class; the model attempts to classify the testing data into one of these two classes, rather than determining the groups itself, which would be unsupervised learning. To further simplify the classifier, Bayesian methods are used, assuming a statistical probability for the data. The most common statistical assumption, Gaussian behavior, is a well explored area of machine learning with which most readers will be familiar. [1]

1.2 Objectives

The motivation for this project was to obtain the most accurate classification model for a set of data. Along the way the performance of multiple methods was evaluated through three experiments. First, each of the three assumption cases for the discriminant function were compared. Second, the effects of prior probability on accuracy were examined. Third, Gaussian and Bimodal models were applied to the data using likelihood ratio and compared. Of course, the project also provided a greater understanding of the methods and material for ECE 571.

1.3 Achievements

The project achieved a max recognition accuracy of 91.3% on the testing data, using the Bimodal model. The accuracy of the likelihood ratio and case three discriminant function were shown to be equal as is expected at 89.8%. The case one and two discriminant functions provided an accuracy of 71.3% and 89.2%, respectively.

2 Technical Approach

2.1 Bayesian Methods

Bayes formula allows the posterior probability to be computed using the prior probability and likelihood as follows

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

From this equation can be derived the likelihood ratio. Given a probability distribution function $p(x|w_i)$ and a prior probability $P(w_i)$ for two classes, the decision rule using likelihood ratio is as follows. When it is true, sample x belongs to class 0.

$$\frac{p(x|w_0)}{p(x|w_1)} - \frac{P(w_1)}{P(w_0)} > 0$$

By setting this equation equal to 0, we obtain the general equation for the plot of the likelihood decision rule.

2.2 Gaussian and Bimodal Distributions

Different probability distributions can be applied to the rule to model the data differently. The first assumption is that the data is Gaussian because this is typical of many natural data sets. The general equation for a multivariate normal distribution function is below. In the case of two feature data, $d = 2$

$$p(x|\theta) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^t \Sigma (x-\mu)}$$

In order to determine the mean and covariance matrices for the data, maximum likelihood estimation was used. The sample mean was calculated by applying each class's feature data to the following equation.

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

The sample covariance matrix was calculated by applying each class's feature data and sample mean to the following equation.

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

The data is actually bimodal, so this model is also applied to the data. A bimodal Gaussian probability distribution is written as the sum of two individual normal Gaussian distributions as follows

$$p(x|w_i) = p_1(x|w_i) + p_2(x|w_i)$$

where

$$p_m(x|w_i) = \frac{a_{im}}{2\pi\sqrt{|\Sigma_{im}|}} e^{-\frac{1}{2}(x-\mu_{im})^t \Sigma_{im}^{-1}(x-\mu_{im})}$$

Variables are indexed first by class and then mode (Gaussian peak). Parameters for both bimodal functions were estimated based on visually applying the function to the data.

2.3 Discriminant Functions

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 decision rule for a two class system is

$$g(x) = g_0(x) - g_1(x) > 0$$

If g(x) is true, sample x is in class 0.

2.3.1 Case 1

The simplest case occurs when every feature 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.

A general equation for the decision rule line for two features is given by the following, where y is feature two, and μ is labeled by class and then index into the vector starting at 1.

$$y = \frac{\mu_{01} - \mu_{11}}{\mu_{12} - \mu_{02}} x + \frac{\mu_{12}^2 + \mu_{11}^2 - \mu_{02}^2 - \mu_{01}^2}{2(\mu_{12} - \mu_{02})} + \frac{\sigma^2(\ln P(w_0) - \ln P(w_1))}{\mu_{12} - \mu_{02}}$$

2.3.2 Case 2

Another simple 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.

A general equation for the decision rule line for two features is given by the following, where y is feature two, μ is labeled by class and then index into the vector starting at 1, and Σ is labeled by index into the matrix, first by row and then by column starting at 1.

$$y = -\frac{Q}{R}x - \frac{L}{R} + \frac{2}{R}(\ln P(w_0) - \ln P(w_1)$$

Q, L, and R are constant terms given by

$$Q = -2\mu_{01}\Sigma_{11}^{-1} - \mu_{02}\Sigma_{21}^{-1} - \mu_{02}\Sigma_{12}^{-1} + 2\mu_{11}\Sigma_{11}^{-1} + \mu_{12}\Sigma_{21}^{-1} + \mu_{12}\Sigma_{12}^{-1}$$

$$L = \mu_{01}^2\Sigma_{11}^{-1}+\mu_{01}\mu_{02}\Sigma_{21}^{-1}+\mu_{01}\mu_{02}\Sigma_{12}^{-1}+\mu_{02}^2\Sigma_{22}^{-1}-\mu_{11}^2\Sigma_{11}^{-1}-\mu_{11}\mu_{12}\Sigma_{22}^{-1}-\mu_{11}\mu_{12}\Sigma_{12}^{-1}-\mu_{12}^2\Sigma_{22}^{-1}$$

$$R = -\mu_{01}\Sigma_{21}^{-1} - \mu_{01}\Sigma_{12}^{-1} - 2\mu_{02}\Sigma_{22}^{-1} + \mu_{11}\Sigma_{21}^{-1} + \mu_{11}\Sigma_{12}^{-1} + 2\mu_{12}\Sigma_{22}^{-1}$$

2.3.3 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

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

Because the decision rule equation is quadratic, programmatic methods were used for solving and plotting the decision rule equation. By incrementing the value of x stepwise 0.05 over the interval 0 to 1, x can be treated as a constant value. This allows the use of the NumPy roots function to solve for y values. When y is unreal or negative, the values of x and y are discarded. Otherwise the pairs (x, y) are stored and later plotted with an approximate fit line as the decision rule line.

The NumPy roots function requires the coefficients a, b, and c as in the form

$$0 = ay^2 + by + c$$

These values can be calculated as follows. Σ is indexed by class and then place in the matrix, starting at 1. V is indexed first by class and then place in the matrix, starting at 1. H is indexed by class.

$$a = \frac{1}{2}(\Sigma_{122}^{-1} - \Sigma_{022}^{-1})$$

$$b = V_{02} - V_{12} + \frac{1}{2}x(\Sigma_{121}^{-1} + \Sigma_{112}^{-1} - \Sigma_{021}^{-1} - \Sigma_{012}^{-1})$$

$$c = \frac{1}{2}x^2(\Sigma_{111}^{-1} - \Sigma_{011}^{-1}) + x(V_{01} - V_{11}) + H_0 - H_1$$

V is a constant 1 x 2 matrix which can be calculated by

$$V_i = (\Sigma_i^{-1}\mu_i)^t$$

H is a constant term which can be calculated by

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

2.4 Prior Probability

Equal (50%) prior probability was assumed for all original calculations and estimates. However, derivations and programmatic implementations took into account the values of prior probability.

In order to determine the prior probability distribution resulting in the highest accuracy, the program stepped through each pair of values $P(w_0) = j, P(w_1) = 1 - j$ from $j = 0$ (0%) to $j = 1$ (100%) by increments of 0.01 (1%).

3 Experiments and results

Three experiments were conducted on the data for this project. First, Gaussian and Bimodal models of the data were compared against each other to see which model produced a higher accuracy on the bimodal data. Second, the accuracy of three cases of the discriminant rule were evaluated to determine the effects of the assumptions made in each case. Last, the distribution of prior probabilities was varied to obtain the values for maximum accuracy and to see the effects of varying those values. Method accuracy was measured as the number of correct classifications divided by the total number of classifications, given as a percentage correct.

3.1 Data

The synthetic data was provided by Ripley's Pattern Recognition and Neural Networks [2]. The data is divided into two groups: 250 training data samples and 1000 testing data samples. Each group is further divided into equal sets of class 0 and class 1 data. The data sets for class 0 and class 1 are bimodal.

Each sample has two continuous decimal features and a binary class value. Throughout this project, these two features are represented by the 2 x 1 vector x. For decision rule plotting purposes, feature one is also sometimes referred to as x, while feature two is referred to as y.

In order to extract the data for a specific class, the following code was used.

```
class0_indices = numpy.where(class == 0)[0]
x0 = x[class0_indices]
```

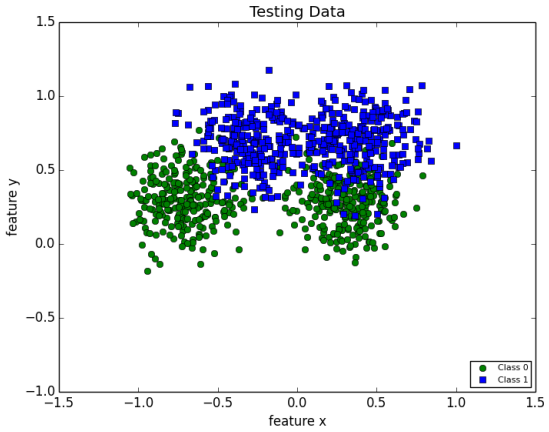

Figure 1. Testing data separated by class.

3.2 Gaussian and Bimodal Distributions

In the first experiment, models for Gaussian and Bimodal distributions were created and used to classify the data. For the Gaussian model, maximum likelihood estimation found these values for the mean.

$$\text{class } 0 = \begin{bmatrix} -0.221470244 \\ 0.32575494 \end{bmatrix}$$

$$\text{class } 1 = \begin{bmatrix} 0.07595431 \\ 0.68296891 \end{bmatrix}$$

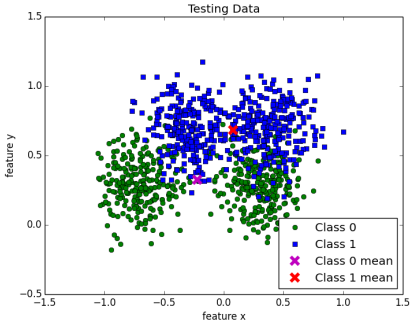

Figure 2. Class 0 and 1 sample means plotted on the testing data.

and these values for the variance.

$$\text{class } 0 = \begin{bmatrix} 0.27459508 & 0.01113883 \\ 0.01113883 & 0.03583011 \end{bmatrix}$$

$$\text{class 1} = \begin{bmatrix} 0.15846988 & -0.01545041 \\ -0.01545041 & 0.02971875 \end{bmatrix}$$

The sample mean and sample variance calculated create a two dimensional Gaussian equation for each class.

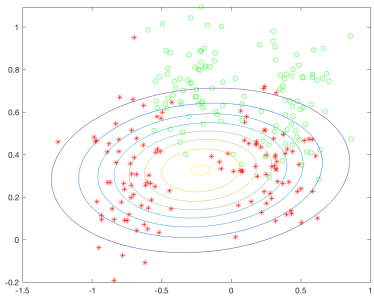

Figure 3. Class 0 Gaussian.

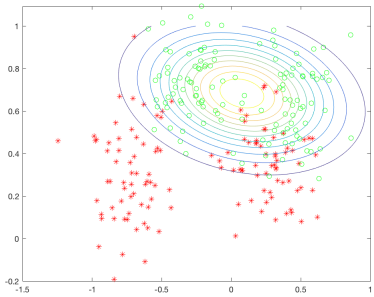

Figure 4. Class 1 Gaussian.

The parameters for the bimodal model were found through visual estimation and are as follows.

Class 0

$$\text{mean 1} = \begin{bmatrix} -0.7 \\ 0.25 \end{bmatrix}$$

$$\text{mean 2} = \begin{bmatrix} 0.3 \\ 0.3 \end{bmatrix}$$

$$\text{covariance 1} = \begin{bmatrix} 0.08 & 0 \\ 0 & 0.08 \end{bmatrix}$$

$$\text{covariance 2} = \begin{bmatrix} 0.08 & 0 \\ 0 & 0.08 \end{bmatrix}$$

Amplitude 1 = 0.5

Amplitude 2 = 0.5

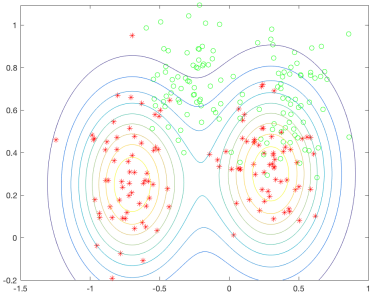

Figure 5. Class 0 bimodal Gaussian estimation.

Class 1

$$\text{mean 1} = \begin{bmatrix} -0.4 \\ 0.7 \end{bmatrix}$$

$$\text{mean 2} = \begin{bmatrix} 0.5 \\ 0.6 \end{bmatrix}$$

$$\text{covariance 1} = \begin{bmatrix} 0.08 & 0 \\ 0 & 0.08 \end{bmatrix}$$

$$\text{covariance 2} = \begin{bmatrix} 0.08 & 0 \\ 0 & 0.08 \end{bmatrix}$$

Amplitude 1 = 0.5

Amplitude 2 = 0.5

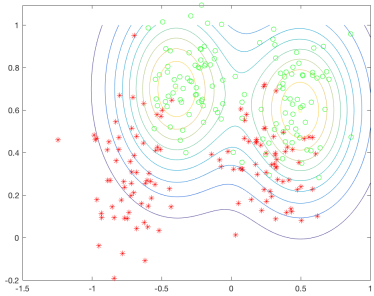

Figure 6. Class 1 bimodal Gaussian estimation.

The Gaussian model obtained an accuracy of 89.8% on the testing data, compared to an accuracy of 91.3% for the Bimodal model. Although the model looks accurate in the estimations, this performance is only a 1.4% improvement over the normal likelihood decision rule. This is a case study in the estimation power of the Gaussian assumption. For this data set, a normal curve still approximated the probability distribution of the two classes relative to each other well.

Figure 7. The decision rule line for the Gaussian likelihood ratio plotted on the testing data. It is equivalent to the discriminant rule case three decision boundary.

3.3 Discriminant Functions

All three cases of the discriminant function performed reasonably well on this data set. Case 2 outperformed case 1, and case 3 outperformed both of the others. As simplification assumptions were made, the accuracy of the model was reduced. Case 1 obtained an accuracy of 71.3% on the testing data. The decision rule line for case 1 is

$$y = -0.83262295x + 0.44378198$$

Case 2 obtained an accuracy of 89.2% on the testing data. The decision rule line for case 2 is:

$$y = -0.13486409x + 0.56584533$$

Case 3 obtained an accuracy of 89.8% on the testing data. This is equal to the accuracy of likelihood ratio, which is expected as they produce the same decision rule. Fitting the curve provides an equation of the form

$$y = 0.10165722x^2 - 0.08706527x + 0.47630219$$

Table 1. Accuracy and decision line functions for each of the three discriminant function cases.

	Accuracy	Decision Line
Case 1	71.3%	y = -0.83262295x + 0.44378198
Case 2	89.2%	y = -0.13486409x + 0.56584533
Case 3	89.8%	y = $0.10165722x^2 - 0.08706527x + 0.47630219$

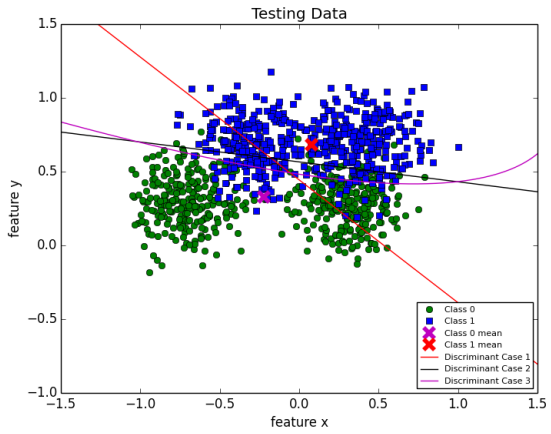

Figure 8. Decision rules plotted on the testing data.

3.4 Prior Probability Distribution

As expected, the accuracies when varying prior probability distribution were inverted parabolas. Maximum accuracies were achieved at the following probabilities. Discriminant case 3 and the likelihood ratio found the same maximum, which is another indication that they have the same decision rule. For most methods, close to equal prior probability produced the best results

Table 2. Maximum accuracies of each method as the prior probability distributions were varied.

	$P(w_1)$	$P(w_2)$	Accuracy
Likelihood Ratio	46%	54%	90%
Discriminant Case 1	19%	81%	73.5%
Discriminant Case 2	46%	54%	89.6%
Discriminant Case 3	46%	54%	90%
Bimodal	48%	52%	91.3%

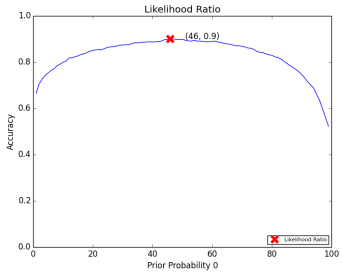

Figure 9. Variance of the accuracy for the likelihood ratio as prior probability changes.

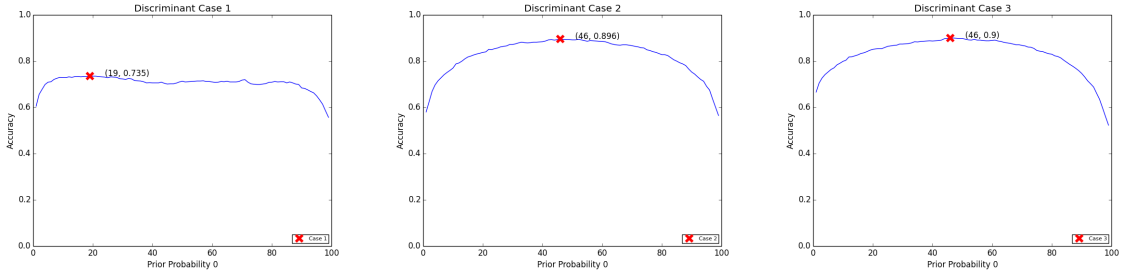

Figure 10. Variance of the accuracy for the discriminant functions as prior probability changes.

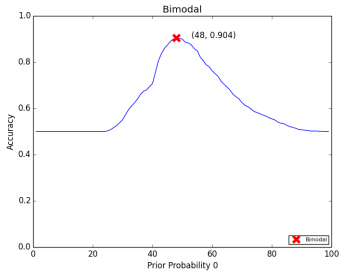

Figure 11. Variance of the accuracy for the bimodal estimate as prior probability changes.

4 Discussion

Over the course of this project's three experiments, various methods using Bayesian parametric binary classification were applied to a data set of two continuous features. The first experiment evaluated two models for the data, one Gaussian and one Bimodal, and discovered that the Bimodal model fit the bimodal data the closest, and produced the maximum accuracy of 91.3% for the entire project. The second experiment considered the three cases of the discriminant function, and found that as simplification assumptions were made, the accuracy of the model was reduced. The third experiment varied the prior probabilities for the two classes and examined the results on accuracy. For most methods, a prior probability of close to 50% produced the highest accuracy.

After completing this project, several improvements for work flow and language choices became obvious. First, although the project guidelines suggested a linear approach to meeting each goal, really the project could be broken into three separate experiments. With this in mind, project flow is a process of theoretical investigation, derivation, experimentation, and then result collection and graphing. Additionally, much time was spent deciding which parts of the project would be in Python and which would be in MATLAB. By the end it was clear that almost every part of the project except for specific two dimensional distribution graphing tasks would be possible in Python.

In the future, more work could be done to improve the performance of the Bimodal estimation model. Only visual estimation was used to determine the mean, variance, and mixing data for this model, which means that accuracy is not validated. Either an advanced library could be used to estimate the parameters, or a range of possible variables could be iterated over while monitoring performance on a training set in order to find a more accurate model. From there, the ranges could be narrowed to find the best fit to the training data. Additionally, it would be excellent to find a library that solves simultaneous matrix equations, and use it for the decision rule plotting portion of this project. Finally, it would be interested to compare this group of Bayesian methods to other learning methods like regression and gradient descent.

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/Bayesian2Class, which includes a small README describing each file.