An Overview of Mixture Discriminant Analysis

John A. Ramey
http://ramhiser.com
johnramey@gmail.com

July 2, 2013

Abstract

Hastie and Tibshirani (1996) proposed a discriminant analysis model based on a mixture of Gaussians, each of which share a common covariance matrix. The mixture discriminant analysis (MDA) model provides a natural extension of the standard Gaussian assumptions underlying the well-known linear and quadratic discriminant analysis methods. However, because the estimators for the model have no closed-form, an EM algorithm was used. In this document, we provide a verbose construction of the model along with a thorough derivation of the parameter estimators as some of the details from Hastie and Tibshirani (1996) were indeed sparse. Using a simple two-dimensional simulated data set, we demonstrate that the MDA classifier identifies three classes, each of which has non-adjacent subclasses, whereas standard Gaussian assumption employed in linear and quadratic discriminant analysis is clearly inadequate and produces poor decision boundaries.

1 Introduction

In discriminant analysis or supervised learning we wish to assign correctly an unlabeled p-dimensional observation vector $\boldsymbol{x}$ to one of K unique, known classes. We assume that the true membership of $\boldsymbol{x}$ is determined by a mapping $y = f(\boldsymbol{x})$ for some unknown function f. Our goal is to estimate f from a labeled training data set $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where $\boldsymbol{x}_i \in \mathbb{R}_{p \times 1}$ is the ith training observation with true, unique membership $y_i \in \{1, \ldots, K\}$ and $\mathbb{R}_{a \times b}$ denotes the matrix space of all $a \times b$ matrices over the real field $\mathbb{R}$. Using $\mathcal{D}$ we train a classifier and select the most probable class label of $\boldsymbol{x}$

$$\begin{aligned}\hat{y} = \hat{f}(\boldsymbol{x}) &= \arg\max_k p(y = k|\boldsymbol{x}) \\ &= \arg\max_k \pi_k f_k(\boldsymbol{x}), \end{aligned} \tag{1}$$

where $f_k(\boldsymbol{x})$ is the kth class-conditional probability density function and π_k is the prior probability of class membership of the kth class. The formulation given in (1) is seen immediately from Bayes' rule and is often called a **generative classifier**.

2 Mixture Discriminant Analysis

Traditionally, in the statistics literature $f_k(\boldsymbol{x})$ is often $N_p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}_k \in \mathbb{R}_{p \times 1}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}_{p \times p}$. The resulting classifier is quadratic discriminant analysis (QDA), which is quadratic in $\boldsymbol{x}$ and has quadratic decision boundaries. It is well-known that the training sample size n must be large relative to the feature dimension p to estimate each of $\boldsymbol{\Sigma}_k$ well. The simplifying assumption $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ is often employed to reduce the number of parameters to estimate as well as to increase model parsimony. The resulting classifier is linear discriminant analysis (LDA), which is linear in $\boldsymbol{x}$ and has linear decision boundaries. Despite the reduction in the number of parameters to estimate, the LDA classifier is ill-posed if $p > n$ because the sample covariance matrix is singular, in which case either regularization methods, feature selection, or further model restrictions are employed, such as assuming $\boldsymbol{\Sigma}$ is diagonal (Ramey and Young, 2013; Dudoit et al., 2002).

Clemmensen et al. (2011) argue that although LDA is often well-suited for simple, low-dimensional settings, linear decision boundaries are often insufficient to separate classes in practice. Furthermore, a single Gaussian distribution may be insufficient in characterizing a single class. With the latter point in mind, Hastie and Tibshirani (1996) propose mixture discriminant analysis (MDA), where $f_k(\boldsymbol{x})$ is the probability density function of a finite Gaussian mixture model. That is, let

$$f_k(\boldsymbol{x}) = \sum_{r=1}^{R_k} \pi_{kr} N_p(\boldsymbol{x}_i | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \tag{2}$$

be a finite mixture density with R_k mixture components, where the rth mixture density has prior probability of π_{kr}, such that $\sum_{r=1}^{R_k} \pi_{kr} = 1$. Note that $\boldsymbol{\Sigma}$ is equal across all classes and subclasses, similar to LDA, in part for model parsimony as well as shrinkage and dimension reduction.

2.1 Model Formulation

In this document, we consider the MDA model as defined by Hastie and Tibshirani (1996) and verbosely provide the details of the likelihood and the estimation of the model parameters. Our notation largely agrees with the original formulation. We refer the interested reader to the original paper for insight regarding the additional topics of reduced-rank discrimination, optimal scoring, and shrinkage applied to the MDA model as these formulations are adequately described by the authors.

From the mixture density (2) we see that $\boldsymbol{x}_i$, $i = 1, \ldots, n$, is realized from one of the R_k subclasses, but we do not know which one. Contrarily, we emphasize that the true class label $y_i = k$ for $\boldsymbol{x}_i$ is known. Now, had we been privy to the subclass from which $\boldsymbol{x}_i$ was generated, the parameters $\boldsymbol{\theta}_{kr} = (\boldsymbol{\mu}_{kr}, \pi_{kr})$ for subclass r_k could be estimated in a straightforward manner. However, the indicator z_{ikr} that $\boldsymbol{x}_i$ was realized from subclass r_k is hidden, which suggests an EM algorithm approach to estimate the model parameters. Notice that $\sum_{r=1}^{R_k} z_{ikr} = 1$. Additionally, if $y_i = k$, we write $y_{ik} = 1$ and 0 otherwise. For clarity, note that y_{ik} is known and not random, while z_{ikr} is unknown.

We define $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{Y} = \{y_1, \ldots, y_n\}$, and $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$, where $\boldsymbol{z}_i = [z_{i1}, \ldots, z_{ir_k}]'$. The complete data likelihood is

$$\begin{aligned} L(\boldsymbol{\theta} | \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) &= \prod_{i=1}^{n} \prod_{k=1}^{K} \{\pi_k f_k(\boldsymbol{x}_i)\}^{y_{ik}} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{r=1}^{R_k} \pi_k^{y_{ik}} \{\pi_{kr} N_p(\boldsymbol{x}_i | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma})\}^{y_{ik} z_{ikr}} . \end{aligned}$$

Hence, the complete data log likelihood is

$$l(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}z_{ikr}\left\{\log\pi_{kr} + \log N_p(\boldsymbol{x}_i|\boldsymbol{\mu}_{kr},\boldsymbol{\Sigma})\right\}.$$

Notice that the log-likelihood is unconstrained so that the probabilities are unbounded. Hence, we add Lagrange multipliers to ensure that $\sum_{k=1}^{K}\pi_k = 1$ and $\sum_{i=1}^{R_k}\pi_{kr} = 1$. Hence, the constrained complete log likelihood is

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = &\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}z_{ikr}\left\{\log\pi_{kr} + \log N_p(\boldsymbol{x}_i|\boldsymbol{\mu}_{kr},\boldsymbol{\Sigma})\right\} \\ &+ \eta\left(\sum_{k=1}^{K}\pi_k - 1\right) + \sum_{k=1}^{K}\eta_k\left(\sum_{r=1}^{R_k}\pi_{kr} - 1\right). \end{aligned}$$

3 Estimation via the EM Algorithm

The parameters are estimated using the EM algorithm.

3.1 E-Step

For the set of parameter estimates $\boldsymbol{\theta}^{(t)}$ at iteration t, the expectation of the complete data log likelihood with respect to the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ and $\boldsymbol{Y}$

is

$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E[l(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{r=1}^{R_k} y_{ik} p_{ikr} \left\{ \log \pi_{kr} + \log N_p(\boldsymbol{x}_i | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \right\} \\
&+ \eta \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \sum_{k=1}^{K} \eta_k \left(\sum_{r=1}^{R_k} \pi_{kr} - 1 \right), \qquad (3)
\end{aligned}$$

where

$$p_{ikr} = E[z_{ikr}] = p(z_{ikr}|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t)}) = \frac{\pi_{kr} N_p(\boldsymbol{x}_i|\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma})}{\sum_{k=1}^{K} \pi_{kr} N_p(\boldsymbol{x}_i|\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma})}$$

is the probability that $\boldsymbol{x}_i$ is a member of subclass r_k conditional on $\boldsymbol{x}_i$ belonging to class k (i.e., $y_{ik} = 1$). The value p_{ikr} is often called the **responsibility** that cluster r_k takes for observation i. Notice that $\sum_{r=1}^{R_k} p_{ikr} = 1$. Also, we define $p_{kr} = \sum_{i=1}^{n} y_{ik} p_{ikr}$.

3.2 M-Step

Here, we wish to optimize (3) with respect to the unknown parameters given in $\boldsymbol{\theta}$. That is, we calculate

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

Below, we optimize each parameter of interest in turn by setting the gradient equal to zero and solving the resulting system of equations.

3.2.1 Estimation of π_k

First, we write

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \pi_k} &= \sum_{i=1}^{n} \frac{y_{ik}}{\pi_k} + \eta = \frac{n_k}{\pi_k} + \eta \\ \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \eta} &= \sum_{k=1}^{K} \pi_k - 1.\end{aligned}$$

From $\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \pi_k} = \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \eta} = 0$, we have $\pi_k = -\frac{n_k}{\eta}$ and $\sum_{k=1}^{K} \pi_k = 1$, which implies that $\sum_{k=1}^{K} -\frac{n_k}{\eta} = 1$. Hence, $\eta = -n$ and

$$\widehat{\pi}_k = \frac{n_k}{n}. \tag{4}$$

3.2.2 Estimation of π_{kr}

Next, we have that

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \pi_{kr}} &= \sum_{i=1}^{n} \frac{y_{ik} p_{ikr}}{\pi_{kr}} + \eta_k = \frac{p_{kr}}{\pi_{kr}} + \eta_k \\ \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \eta} &= \sum_{r=1}^{R_k} \pi_{kr} - 1.\end{aligned}$$

From $\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \pi_{kr}} = \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \eta_k} = 0$, we have $\pi_{kr} = -\frac{p_{kr}}{\eta_k}$ and $\sum_{r=1}^{R_k} \pi_{kr} = 1$, which implies that

$$-\sum_{r=1}^{R_k} \frac{p_{kr}}{\eta_k} = 1.$$

Hence, $\eta_k = -n_k$ and

$$\widehat{\pi}_{kr} = \frac{p_{kr}}{n_k}. \tag{5}$$

3.2.3 Estimation of $\boldsymbol{\mu}_{kr}$

Recall that $\frac{\partial(\boldsymbol{a}'\boldsymbol{A}\boldsymbol{a})}{\partial \boldsymbol{a}} = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{a}$ for $\boldsymbol{a} \in \mathbb{R}_{p\times 1}$ and $\boldsymbol{A} \in \mathbb{R}_{p\times p}$ (Murphy, 2012, p. 99). Letting $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu}_{kr}$ and applying the chain rule, we have that

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}_{kr}} \log N(\boldsymbol{x}|\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) &= -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}_{kr}}(\boldsymbol{x} - \boldsymbol{\mu}_{kr})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{kr}) \\ &= -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}_{kr}}\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\mu}_{kr}} \\ &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{kr}).\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\mu}_{kr}} &= \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}\frac{\partial}{\partial \boldsymbol{\mu}_{kr}} \log N(\boldsymbol{x}|\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \\ &= \sum_{i=1}^{n} y_{ik}p_{ikr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr}).\end{aligned}$$

Setting $\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\mu}_{kr}} = 0$, we have

$$\begin{aligned}\sum_{i=1}^{n} y_{ik}p_{ikr}\boldsymbol{x}_i &= \sum_{i=1}^{n} y_{ik}p_{ikr}\boldsymbol{\mu}_{kr} \\ &= p_{kr}\boldsymbol{\mu}_{kr},\end{aligned}$$

which implies that

$$\widehat{\boldsymbol{\mu}}_{kr} = \frac{\sum_{i=1}^{n} y_{ik} p_{ikr} \boldsymbol{x}_i}{p_{kr}}. \tag{6}$$

3.2.4 Estimation of $\boldsymbol{\Sigma}$

Recall that $\frac{\partial}{\partial \boldsymbol{A}}|\boldsymbol{A}| = (\boldsymbol{A}^{-1})'$ and $\frac{\partial}{\partial \boldsymbol{A}}\mathrm{tr}(\boldsymbol{B}\boldsymbol{A}) = \boldsymbol{B}'$ for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}_{p\times p}$ (Murphy, 2012, p. 99). Hence,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}} \log N(\boldsymbol{x}|\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) &= -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} \left\{\log|\boldsymbol{\Sigma}| + (\boldsymbol{x} - \boldsymbol{\mu}_{kr})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{kr})\right\} \\
&= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} \left[\log|\boldsymbol{\Sigma}| - \mathrm{tr}\left\{(\boldsymbol{x} - \boldsymbol{\mu}_{kr})(\boldsymbol{x} - \boldsymbol{\mu}_{kr})'\boldsymbol{\Sigma}^{-1}\right\}\right] \\
&= -\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{kr})(\boldsymbol{x} - \boldsymbol{\mu}_{kr})'\boldsymbol{\Sigma}^{-1}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\Sigma}} &= \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}\frac{\partial}{\partial \boldsymbol{\Sigma}} \log N(\boldsymbol{x}|\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{kr})(\boldsymbol{x} - \boldsymbol{\mu}_{kr})'\boldsymbol{\Sigma}^{-1}.
\end{aligned}$$

Setting $\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\Sigma}} = 0$, we have

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}\boldsymbol{\Sigma}^{-1} = \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr})(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr})'\boldsymbol{\Sigma}^{-1}.$$

Premultiplying and postmultiplying by $\boldsymbol{\Sigma}$, we have

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr})(\boldsymbol{x}_i - \boldsymbol{\mu}_{kr})' = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}\boldsymbol{\Sigma} = n\boldsymbol{\Sigma},$$

which implies that

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k} y_{ik}p_{ikr}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{kr})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{kr})'. \tag{7}$$

4 Example

Using a simple simulated data set, we demonstrate that the LDA and QDA classifiers based on a Gaussian assumption can be inadequate. We contrast the decision boundaries with those obtained via the `mda` R package available on CRAN[1].

In this example, we generate $K = 3$ classes, each of which has 3 subclasses. We selected the location of the subclasses so that no subclass was adjacent either horizontally or vertically. We used all of the defaults from the `lda` and `qda` functions from the `MASS` R package. Likewise, we applied only the defaults for the MDA classifier: in particular, we allowed the number of subclasses to be determined automatically.

References

Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse Discriminant Analysis. *Technometrics*, 53(4):406–413, November 2011.

[1] `http://cran.r-project.org/web/packages/mda`

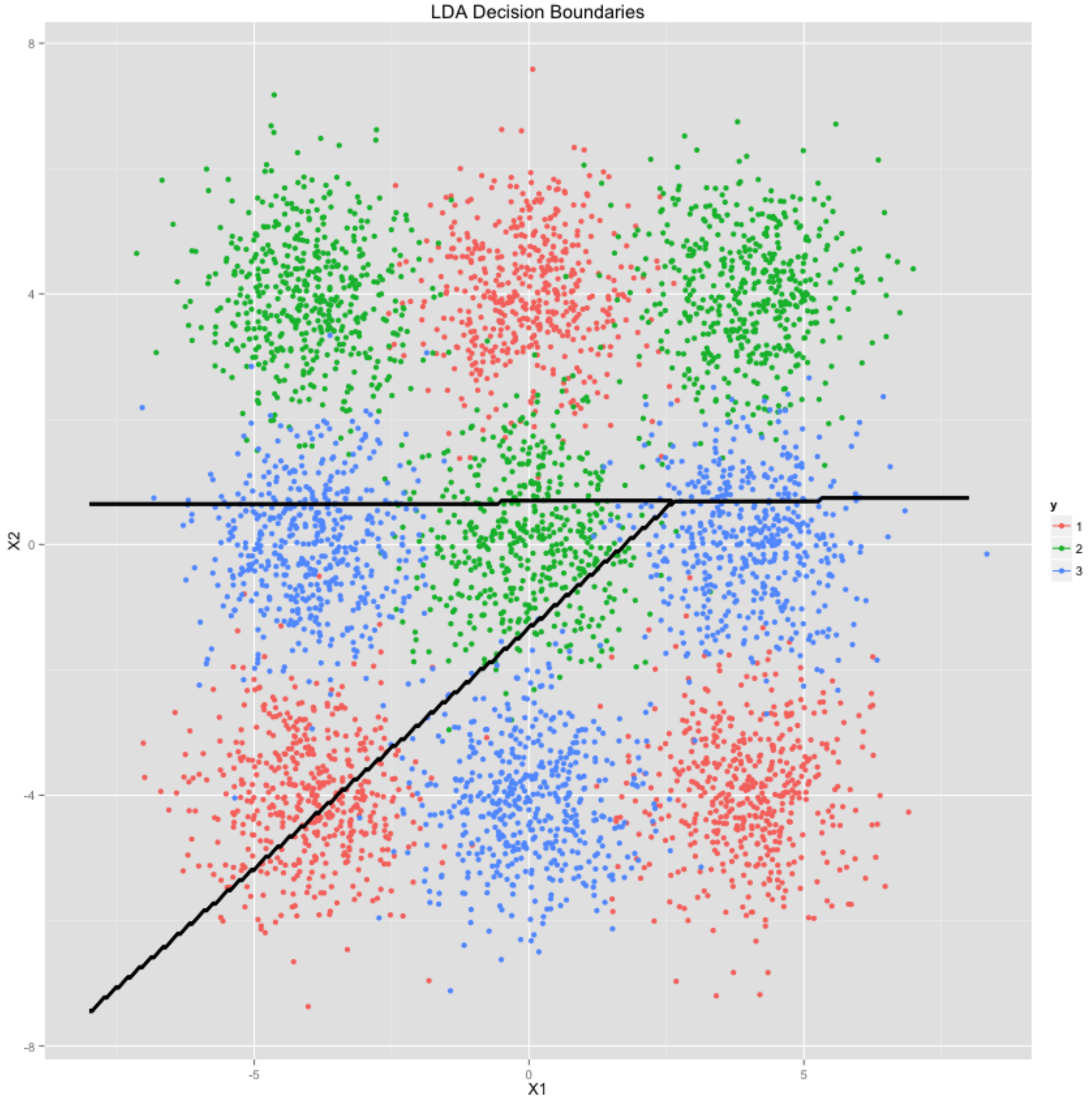

Figure 1: Decision boundaries via linear discriminant analysis.

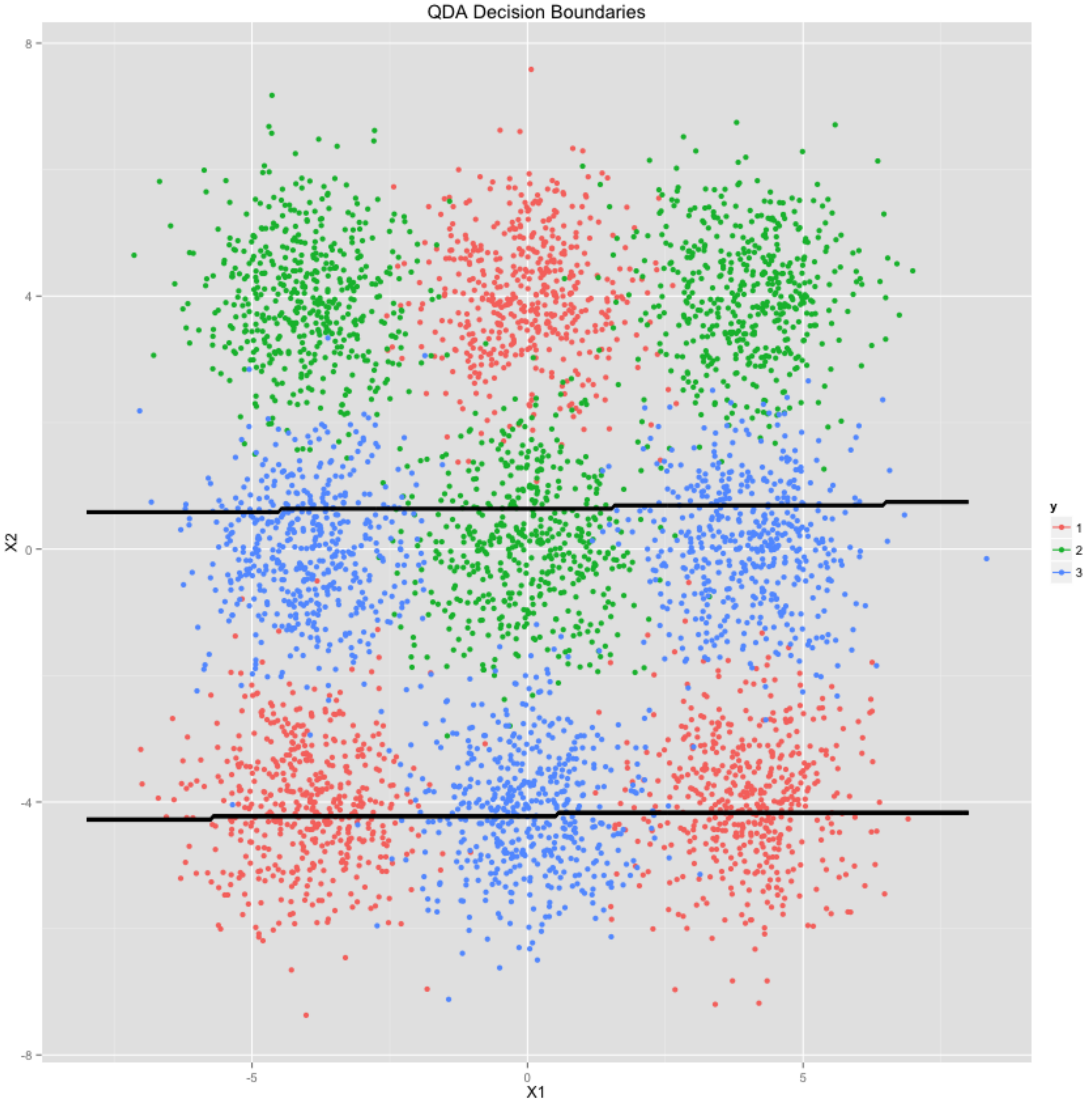

Figure 2: Decision boundaries via quadratic discriminant analysis.

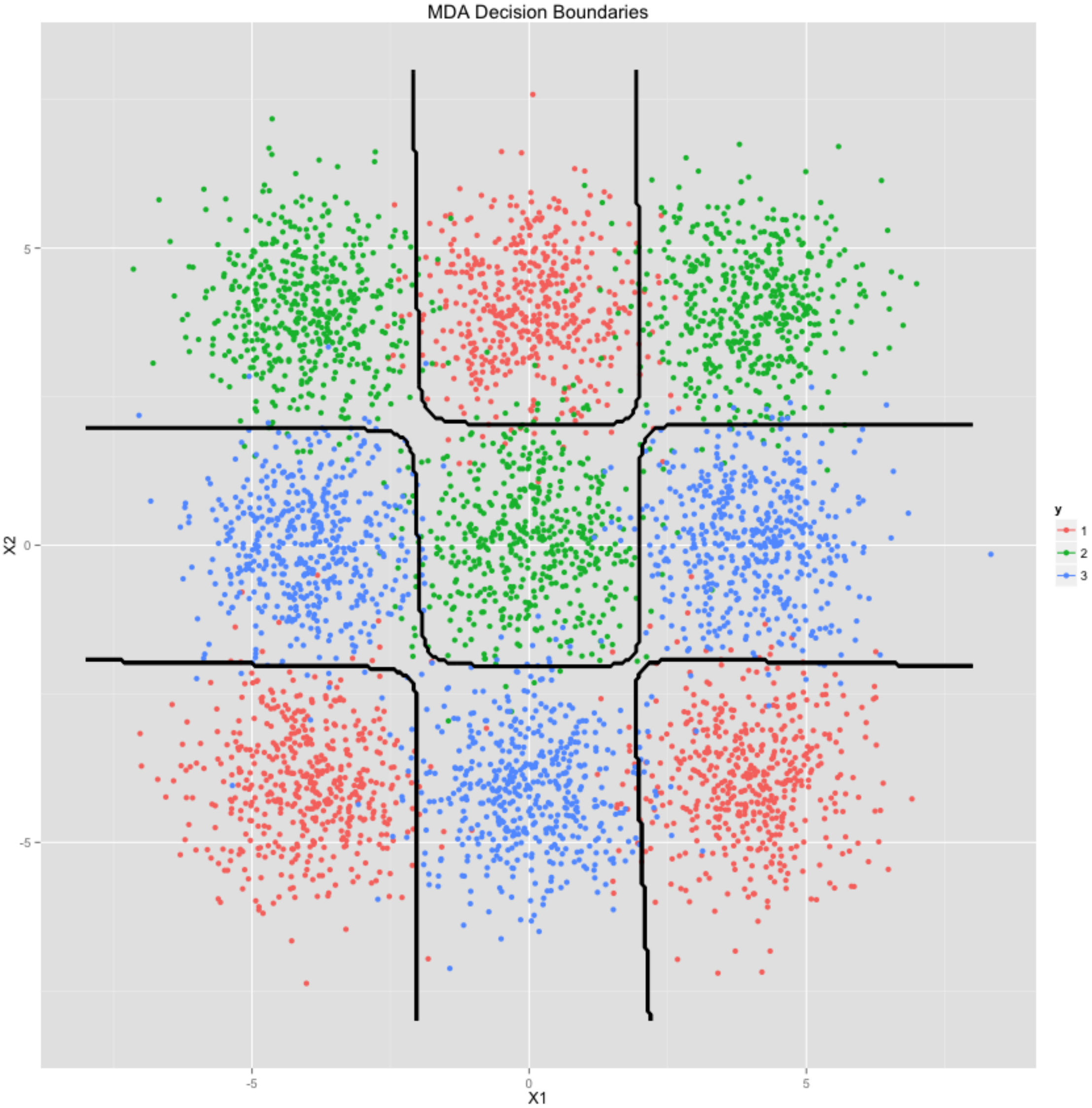

Figure 3: Decision boundaries via mixture discriminant analysis.

Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87, March 2002.

Trevor Hastie and Robert Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176, 1996.

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective.* The MIT Press, Cambridge, Massachusetts, August 2012.

John Ramey and Phil D Young. A comparison of regularization methods applied to the linear discriminant function with high-dimensional microarray data. *Journal of Statistical Computation and Simulation*, 83(3):581–596, 2013.