STAT562 Lecture 5: Discriminant Analysis

Beidi Qiang

SIUE

Introduction

Logistic regression involves directly modeling

$$P(class|feature) = Pr(Y = k|X = x),$$

the conditional distribution of the response Y given the predictor(s) X. We now consider an alternative approach, Discriminant Analysis.

- In Discriminant Analysis, we model $P(feature|class)$
- Discriminant analysis is more stable when the classes are well-separated.
- Discriminant model is again more stable when n is small with normally distributed predictor.
- Discriminant analysis is popular when we have more than two response classes.

Bayes' Thm

Suppose the categorical response variable Y can take on K possible classes. Bayes' theorem states that

$$p(class = k|feature = x) = Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

- $\pi_k = P(Y = k)$ is called prior probability that a randomly chosen observation comes from the kth class.
- $f_k(x) = Pr(X = x|Y = k)$ denote the density function of X for an observation that comes from the kth class.
- $Pr(Y = k|X = x)$ is called the posterior probability that an observation belongs to the kth class, given the featrue value for that observation.
- Instead of directly computing $Pr(Y = k|X = x)$ like in the logistic regression, we seek estimates of π_k and $f_k(x)$.

Linear Discriminant Analysis for One Predictor

Assume that $p = 1$, that is, we have only one predictor.

- To estimate π_k: we simply compute the fraction of the training observations that belong to the kth class.
- Estimate $f_k(x)$: we assume that $f_k(x)$ has is a normal density function with parameter μ_k and common σ^2.

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{1}{2}(x-\mu_k)^2/\sigma^2\right]$$

- we estimate μ_k and σ^2. The common estimators are just the sample mean and variance in each class.

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i$$

We further assume all classes share the same variance, then

$$\hat{\sigma}^2 = \hat{\sigma}_k^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2$$

Discriminant Function

$$\hat{p}_k(x) = \hat{P}(class = k|x) = \frac{\hat{\pi}_k \frac{1}{\sqrt{2\pi}\hat{\sigma}} exp(-(x-\hat{\mu}_k)^2/2\hat{\sigma}^2)}{\sum_{i=1}^{K} \hat{\pi}_i \frac{1}{\sqrt{2\pi}\hat{\sigma}} exp(-(x-\hat{\mu}_i)^2/2\hat{\sigma}^2)}$$

The LDA classifier will assign an observation to the class for which $\hat{P}(class = k|x)$ is largest.

- Taking the log of $\hat{p}_k(x)$ and rearranging the terms, it is not hard to show that this is equivalent assign class that gives the largest

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

- Note $\hat{\delta}_k(x)$ is called "the discriminant functions", which are linear in x.

A simple example

$K = 2$ and $n_1 = n_2 = 20$ observations were drawn from each of the two classes. The mean and variance parameters for the two density functions are $\mu_1 = -1.25$, $\mu_2 = 1.25$, and $\sigma_1 = \sigma_2 = 1$.

- $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$
- $\hat{\delta}_1(x) = (x\hat{\mu}_1 - \hat{\mu}_1^2/2)/\sigma^2$, and $\hat{\delta}_2(x) = (x\hat{\mu}_2 - \hat{\mu}_2^2/2)/\sigma^2$
- Assigns an observation to class 1 if $2x(\hat{\mu}_1^2 - \hat{\mu}_2^2) > \hat{\mu}_1 - \hat{\mu}_2$, i.e. $x > \frac{\hat{\mu}_1+\hat{\mu}_2}{2}$ is assigned to one class and $x < \frac{\hat{\mu}_1+\hat{\mu}_2}{2}$ is assigned to another.
- $x = \frac{\hat{\mu}_1+\hat{\mu}_2}{2}$ is the decision boundary.

A simple example, cont.

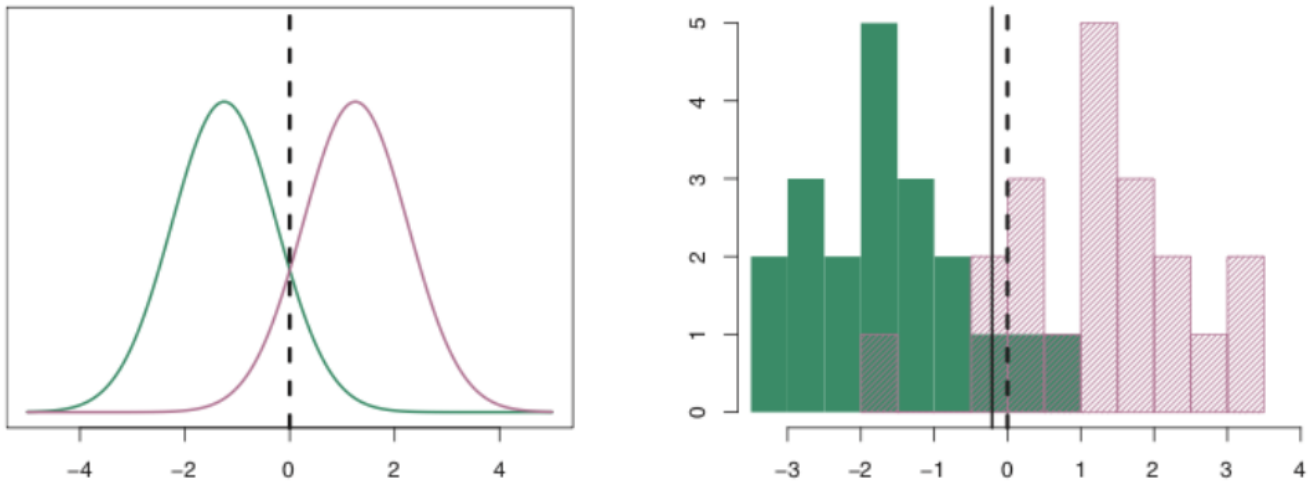


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

LDA for Multi-dimensional feature

Extend the LDA classifier to the case of multi-dimensional features.

- We assume the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \Sigma)$ with pdf:

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) \right]$$

- μ_k is a p-dimensional class-specific mean vector, and Σ is a $p \times p$ covariance matrix that is common to all K classes.
- We need to estimate the unknown parameters μ_k and Σ; the formulas are similar but more technical, we will skip the details.

Prediction

The LDA classifier will again assign an observation to the class for which $\hat{p}_k(x)$ is largest.

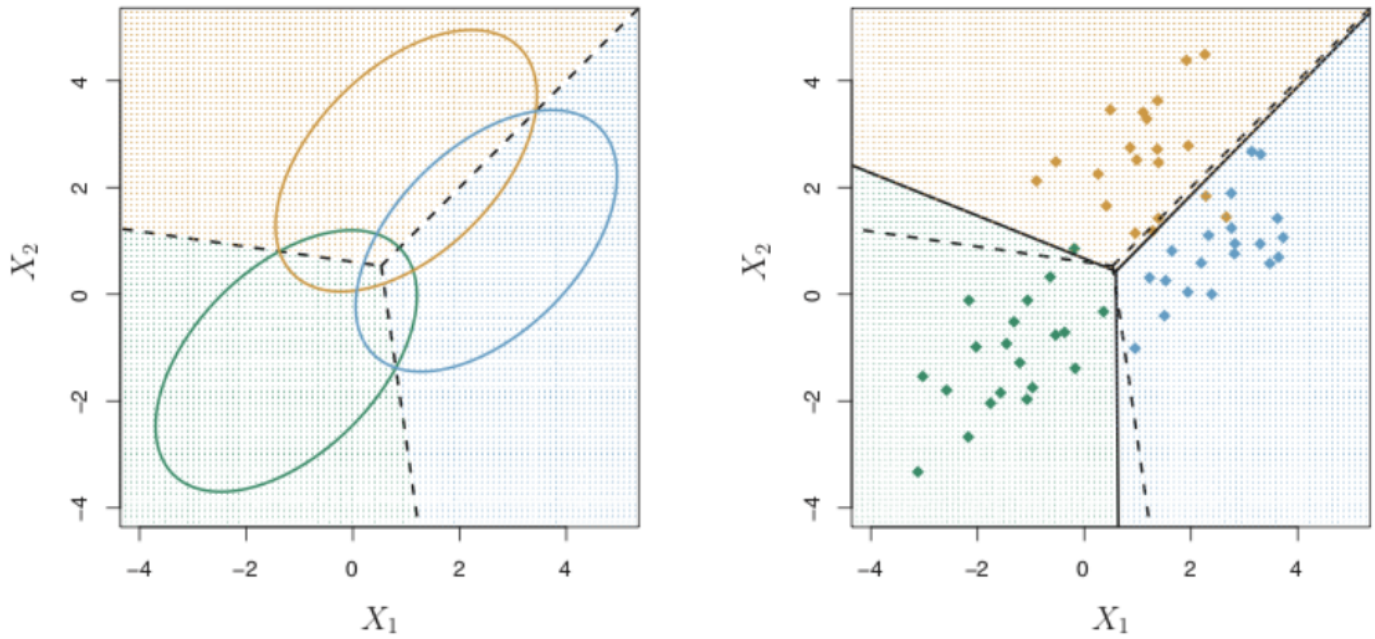
- Taking the log and rearranging the terms, we can show that this is equivalent assign class that gives the largest

$$\hat{\delta}_k(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log(\hat{\pi}_k)$$

- This is the vector/matrix version of the $\hat{\delta}_k(x)$ in single predictor case.

An example with three classes

Simulated data. Left: The true underling distribution. Right: LDA classification. LDA decision boundaries are indicated using solid black lines.



Example: Default data

The data deals with whether an individual will default on his or her credit card payment, on the basis of annual income, monthly credit card balance and student status. LDA model out put from R is given below:

```
> lda.out=lda(default~income+student+balance)
> lda.out
Call:
lda(default ~ income + student + balance)

Prior probabilities of groups:
    No    Yes
0.9667 0.0333

Group means:
      income studentYes   balance
No  33566.17  0.2914037  803.9438
Yes 32089.15  0.3813814 1747.8217

Coefficients of linear discriminants:
                    LD1
income      3.367310e-06
studentYes -1.746631e-01
balance     2.243541e-03
> lda.class=predict(lda.out,Default)$class
> table(lda.class,default)
         default
lda.class   No  Yes
      No  9645  254
      Yes   22   79
```

Quadratic Discriminant Analysis

Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a normal distribution, but unlike LDA, QDA assumes that each class has its own covariance matrix.

- QDA assumes that an observation from the kth class is of the form $N(\mu_k, \Sigma_k)$
- The QDA classifier will again assign an observation to the class for which $\hat{p}_k(x)$ is largest, which now is equivalent assign class that gives the largest

$$\hat{\delta}_k(x) = -\frac{1}{2}x^T\hat{\Sigma}_k^{-1}x + x^T\hat{\Sigma}_k^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}_k^{-1}\hat{\mu}_k + \log(\hat{\pi}_k)$$

- Note the discriminant functions $\hat{\delta}_k(x)$, are now quadratic in x.

Naive Bayes

Unlike LDA and QDA classifier that assumes the probability distribution $Pr(x|class)$ follows a p-dimensional multivariate normal distribution, Naive Bayes classifier assumes the p features are independent, i.e.

$$f_k(x) = f_{k1}(x_1) f_{k2}(x_2) \cdots f_{kp}(x_p)$$

To estimate the one-dimensional density function f_{kj}:

- If X_j is quantitative, then we can assume a univariate normal distribution for each feature. This is similar to QDA, but with an additional assumption that the covariance matrix Σ_k is diagonal.
- If X_j is quantitative, another option is to use a non-parametric estimate, such as kernel density estimator.
- If X_j is qualitative, then we can simply count the proportion of training observations for the jth predictor corresponding to each class.

Comparison of LDA and Logistic Regression

We have considered different classification approaches: logistic regression, LDA, and QDA, and K-nearest neighbors (KNN) method. How do they compare?

- In the case of two-class setting, logistic regression and LDA methods are closely connected, since the log-odds in logistic framework can also be expressed as a linear function of x, i.e. $\log[p_1(x)/p_2(x)] = c_0 + c_1 x$. The difference is only in their fitting procedures.
- LDA assumes a Gaussian distribution with a common covariance matrix. It performs better when this assumption approximately holds.
- Logistic regression can outperform LDA if these Gaussian assumptions are not met.

Compared with KNN

KNN takes a completely different approach. It is a non-parametric approach.

- KNN makes no assumptions about the shape of the decision boundary, or distribution of predictors.
- KNN will perform the best when the decision boundary is highly non-linear.
- KNN does require a larger training set so form an accurate decision boundary
- LDA and logistic regression performs better when the decision boundary is approximately linear and/or when training set is small.
- We lose some interpretability with such a non-parametric approach. We can't discuss the effect of predictors on the response with the KNN approach.

Compared with QDA

QDA serves as a compromise between the completely non-parametric KNN method and the linear LDA.

- QDA allows curvature by assuming a quadratic decision boundary.
- The QDA boundary is not as flexible as KNN.
- But QDA performs better than KNN when training set is small. Because it makes assumption of the form of the decision boundary.

Types of errors

A confusion matrix is a convenient way to display error information.

- Two types of errors: it can incorrectly assign an individual who defaults to the no default category, or it can incorrectly assign an individual who does not default to the default category.
- Note the error rate is low (22/9667) for individuals who did not default, but (unacceptably) high (254/333) among individuals who defaulted.
- Why does LDA do such a poor job of classifying the customers who default?
  LDA is trying to get the lowest total error rate, i.e. smallest possible total number of misclassified observations, irrespective of which class the errors come from!