CKW

Statistics 8330: Data Analysis III
Classification (review)

Suggested Reading: James, Witten, Hastie, and Tibshirani (JWHT), 2013, Chapter 4

Christopher K. Wikle

University of Missouri
Department of Statistics

Classification

CKW

A great many of the problems of interest to data scientists today have to do with classification. That is, our responses are qualitative and correspond to a finite number of specific groups. Our goal is then to use various predictors to be able to classify a new observation into the appropriate response category.

Thus, we still have observation pairs $\{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$, but in this case, the y_i are qualitative. We seek to estimate a function $f(X)$ that will allow us to classify a new x_0 into the appropriate response category.

There are numerous methods that can be used for classification. In this lecture, we review methods that we learned in Data Analysis II (Stat 8320). In particular, we very briefly review *logistic regression*, *linear discriminant analysis*, and *quadratic discriminant analysis*. Additional details can be found in the Stat 8320 notes.

Classification

Consider the classification problem for credit card default "data" presented in Chapter 4 of JWHT (2013):

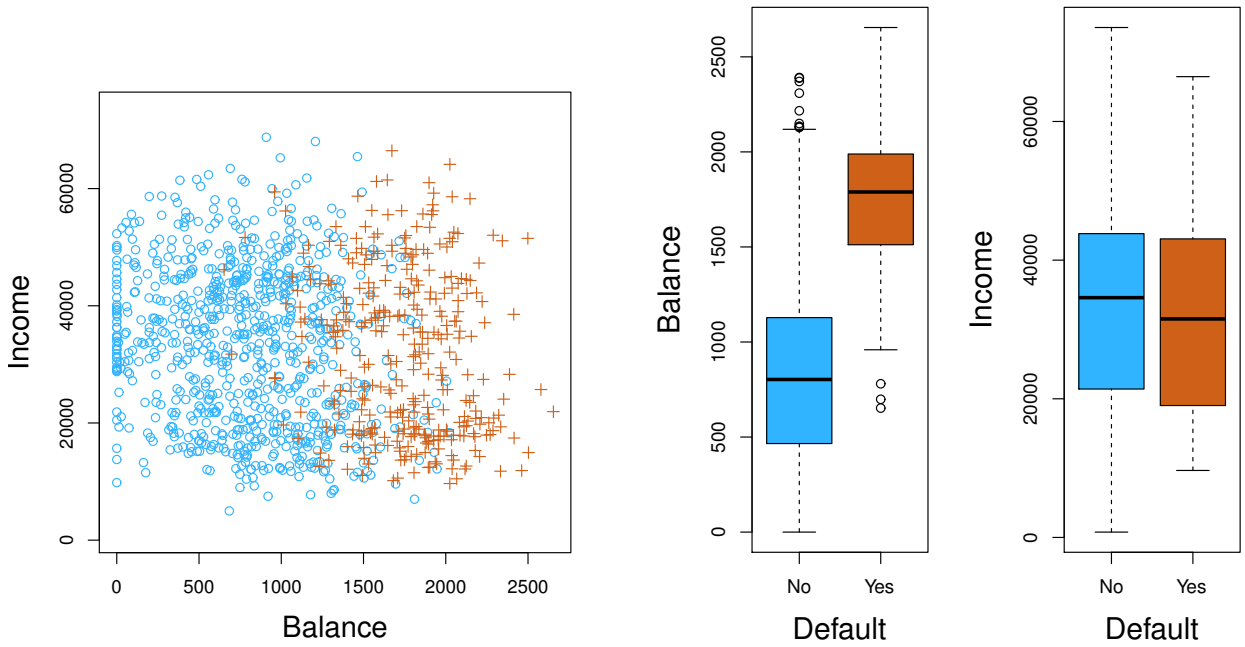

FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

Logistic Regression

CKW

As we discussed in Data Analysis II, it is not reasonable to use normal error regression models when fitting categorical data. Rather, we considered a generalized linear model (GLM) or generalized linear mixed model (GLMM) approach. The strength of such an approach is that one models the responses as conditionally independent distributions from the exponential family and then considers a transformation of the mean response (i.e., the link function) to be a *linear* function of the predictors. In the context of a binomial data distribution and a logistic link function, this is simply *logistic regression*.

In the context of two categories, we can model the observations as 0s and 1s. Interest is then in modeling the associated probability that the response takes a value of 1, (say, p_i) given the covariates ($p_i(X)$). More formally, we are interested in modeling $Pr(Y = k|X = x)$, which is the conditional probability that Y takes the kth discrete value ($k = \{0, 1\}$) given $X = x$.

Logistic Regression (cont.)

CKW

Recall the logit function:

$$logit(p(X)) = \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j X_j.$$

Note, this implies

$$p(X) = \frac{e^{\beta_0+\sum_j \beta_j X_j}}{1+e^{\beta_0+\sum_j \beta_j X_j}}$$

As we discussed in Stat 8320, one typically uses MLE to obtain estimates for the βs (usually through an iteratively reweighted least squares algorithm).

Given estimates of the parameters and their standard errors, one can perform inference similar to normal error regression analysis. Note that we have to be careful when performing inference on these parameters for the same reasons as in normal error regression. E.g., see the confounding example in JWHT (2013) Section 4.3.4.

Logistic Regression (cont.)

CKW

When interest is with classification, we typically use the logistic regression model for prediction rather than inference. In this case, we are interested in estimating $\hat{p}(X_0)$ and then classifying according to the category with the highest probability.

When p is large, there are algorithms that can perform stepwise model selection. We will also see other ways to perform model selection later in the course. Note that when classes are well-separated, the parameter estimates for logistic regression can be unstable.

When we have more than 2 categories, recall that we can still use the GLM framework with multinomial data models. Alternatively, we can perform more classical *discriminant analysis* as we discussed in Stat 8320. We briefly review basic linear and quadratic discriminant analysis here.

Discriminant Analysis

It is convenient to consider Bayes' theorem when thinking about the classification problem (even if one is not a "Bayesian"!).

Assume that our response Y can be one of $K \geq 2$ classes and let π_k correspond to the *prior* probability that an observation is from the kth class (ideally, before we collect data).

Then, we specify a model for the probability of an observation pair taking the particular X values, given it is from a given class:

$$f_k(X) \equiv Pr(X = x|Y = k),$$

which is just the density function of X for an observation from the kth class.

Discriminant Analysis (cont.)

CKW

Now, we are interested in classifying Y given $X = x$, so we use Bayes' theorem to obtain the *posterior probabilities*:

$$p_k(x) = Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}.$$

We classify the observation into the class for which $p_k(x)$ is largest. To calculate $p_k(x)$, we simply need the π_k (which we have before the analysis) and must specify the density $f_k(X)$. The former is either subjective, based on previous knowledge, or based empirically on the fraction of observations for each class. The later is where we put most of the work.

In general, we can specify fixed/known forms for $f_k(X)$ that depend on parameters and use the data for the kth class to estimate those parameters. Alternatively, we can try to model this density function non-parametrically (as with the KNN approach discussed previously). We consider the first approach in traditional linear and quadratic discriminant analysis.

Linear Discriminant Analysis

CKW

Recall from the introductory lecture to this class that the Bayes classifier has the lowest possible error rate out of ALL classifiers **if** all π_k and $f_k(X)$ are specified correctly (which, of course, is never possible outside of simulated examples). But, this suggests that we should strive for good estimates of these components. Traditional linear and quadratic discriminant analysis seeks good approximations by assuming these are normal densities.

Linear Discriminant Analysis: LDA

In LDA, we assume that the densities, $f_k(X)$, are normal (Gaussian) distributions and assume that the variance/covariance parameters are the same for all distributions. This leads to a linear decision boundary between the classes.

LDA ($p = 1$)

Consider the case where $p = 1$ (for now) and assumed that $f_k(X)$ is normal (Gaussian):

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right),$$

where μ_k and σ_k^2 are the mean and variances for the kth class, respectively. As mentioned, in LDA, we make the further assumption that $\sigma_1^2 = \sigma_2^2 = \ldots \sigma_K^2 = \sigma^2$.

It is easy to show that plugging this into Bayes theorem and taking the log allows one to assign the observation to the class for which the following criterion is largest (note, this is *linear* in x):

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

LDA ($p = 1$ cont.)

So, if $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier assigns to class 1 if $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$ (and, assigns to the other class otherwise). This corresponds to a Bayes decision boundary where

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}.$$

In practice, we have to estimate μ_1, μ_2 and σ^2!

Consider the example from JWHT (2013, Fig. 4.4):

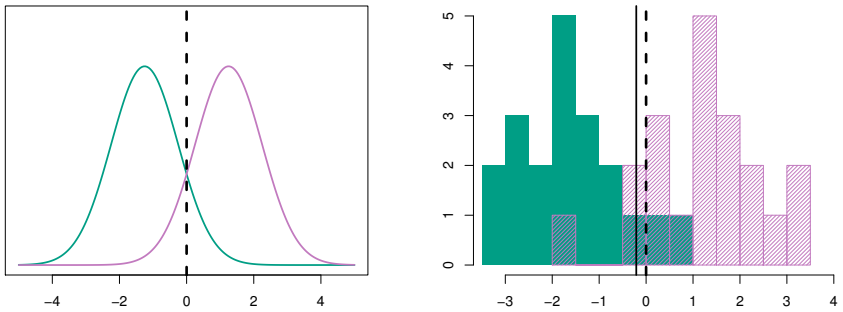

FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line.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

LDA with Multiple Predictors

Obviously, the case where $p = 1$ is not interesting. In reality, we have $p > 1$. The LDA procedure is the same – we simply have to work with multivariate normal (MVN) densities rather than univariate densities. In this case, we assume $f_k(X = x) = N_p(\mu_k, \mathbf{\Sigma})$ (i.e., p-dimensional MVN distributions with different mean vectors and common variance/covariance matrices).

It can be shown that we classify an observation $X = x$ to the class for which the linear function (in x)

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k,$$

is largest.

LDA

Consider the example in JWHT (2013) where $p = 2$ and $K = 3$:

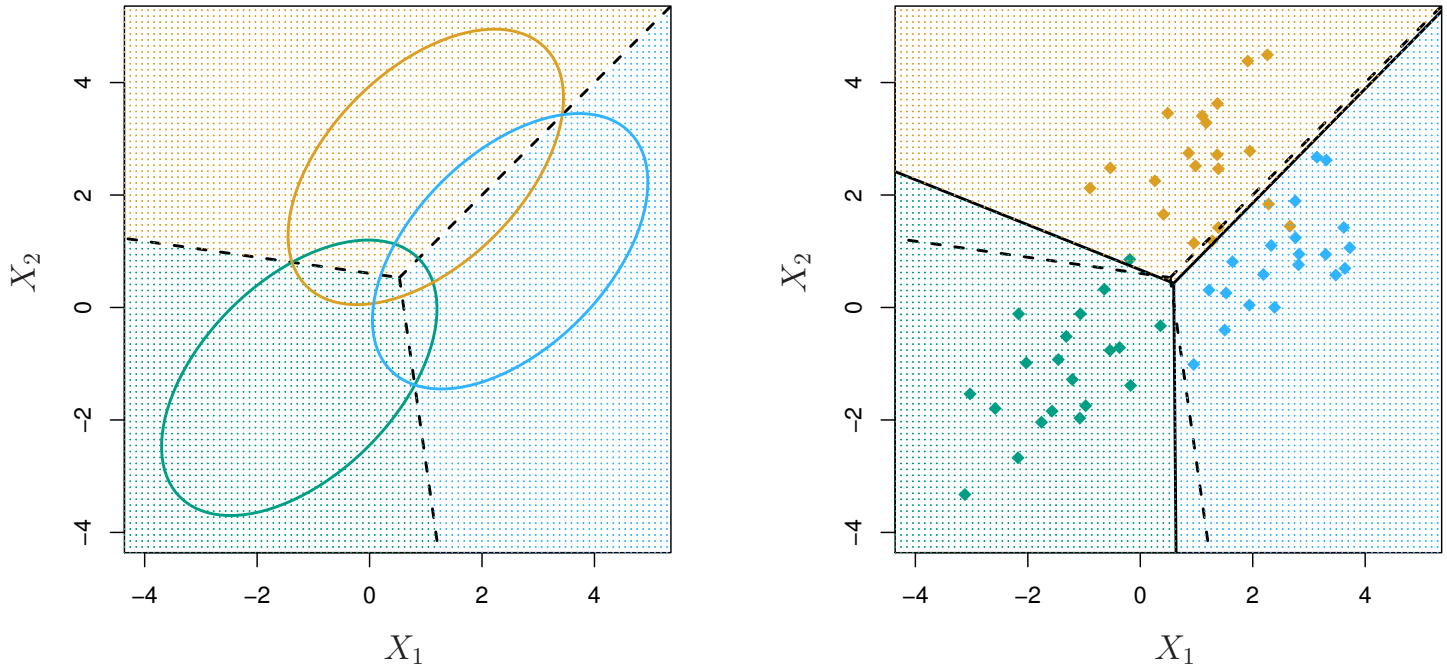

FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95 % of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.* Right: 20 *observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.*

LDA: Estimation

To estimate the Bayes classifier, we have to estimate the relevant parameters, $\{\mu_k : k = 1, \ldots, K\}$, $\mathbf{\Sigma}$, and $\{\pi_k : k = 1, \ldots, K\}$. When p is fairly low dimensional, this isn't too difficult, but it becomes increasingly difficult as p increases due to the curse of dimensionality (recall, there are $p(p+1)/2$ unique parameters in $\mathbf{\Sigma}$).

Quadratic Discriminant Analysis (QDA)

CKW

As mentioned, the LDA Bayes decision boundary is linear. One can imagine situations where a quadratic decision boundary might lead to better classification. As it turns out, the assumption that the variance/covariance matrix is common across all classes for Gaussian densities leads to the linear rule. If we do not enforce such an assumption (i.e., allow the predictors in each class to follow MVN distributions with class-specific mean vectors **and** variance/covariance matrices) then we obtain a quadratic classifier (in x); this is called *Quadratic Discriminant Analysis* (QDA).

In particular, we assign the observation $X = x$ to the class with the largest value of:

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \mathbf{\Sigma}_k^{-1}(x - \mu_k) + \log \pi_k,$$

which is clearly quadratic in x.

QDA Estimation and Issues

To estimate the Bayes classifier for QDA, one has to estimate $\{\mu_k : k = 1, \ldots, K\}$, $\{\mathbf{\Sigma}_k : k = 1, \ldots, K\}$, and $\{\pi_k : k = 1, \ldots, K\}$. By far the biggest challenge is estimating the $Kp(p+1)/2$ parameters in the variance covariance matrices.

In general, one can see that there are many more parameters in QDA, which gives it more flexibility than LDA, at the cost of more variance. Thus, LDA vs QDA illustrates yet another example of the bias-variance tradeoff. By sharing common parameters in the variance/covariance matrix, LDA provides a less flexible classifier (i.e., it can have high bias but it has lower variance). On the other hand, since QDA has many more parameters to control the shape of the distributions, it has less bias but more variance.

In general, when one has relatively few observations, then LDA is going to provide better classifications than QDA.

QDA (cont.)

CKW

Consider the example from JWHT (2013):

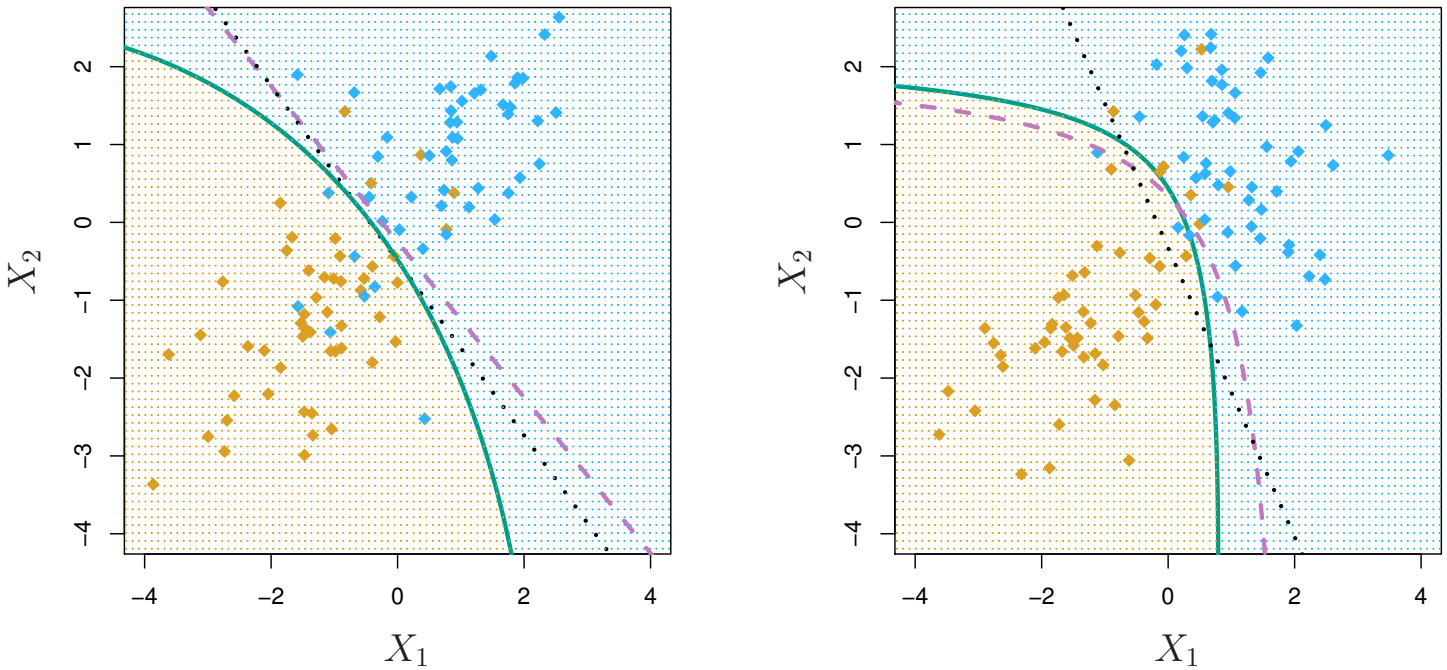

FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

Comparison of Classification Methods: logistic regression vs LDA

Recall that by definition, logistic regression provides a decision boundary that is linear in the predictors (X) (when X does not enter the model in some transformed fashion, e.g., $\sqrt{X}$, X^2, X^3, etc.) .

Thus, both LDA and logistic regression produce linear classification boundaries.

The primary difference between LDA and logistic regression is that LDA assumes that the observations come from a MVN distribution with common variance/covariance matrices across class. Logistic regression does not have this normality assumption. So, typically, LDA does better than logistic regression in situations where the normal assumptions hold. Otherwise, logistic regression is going to be more robust to deviations from the normality assumption.

Comparison of Classification Methods: KNN vs linear methods

Recall in the overview lecture we discussed KNN classification. The KNN classifier is non-parametric and simply classifies an observation into the class that the majority of its K neighbors belong to. One can control the variance/bias tradeoff by varying K; thus, it is quite flexible and can model much more complicated decision boundaries.

Typically, KNN would outperform both LDA and logistic regression when the decision boundary is highly nonlinear. However, KNN cannot provide an indication as to which predictors are most important, and one needs many observations to adequately fit these models when K is relatively small.

Comparison of Classification Methods (cont.)

Since QDA can accommodate a special type of nonlinear decision boundary (quadratic), it is essentially a compromise between the linear methods and KNN classification. In addition, although QDA is not as flexible as KNN, it can be used with fewer training observations than KNN classification in general.

Overall, simulation studies show:

- When the true decision boundary is linear, LDA and logistic regression typically perform the best
- When the true decision boundary is moderately nonlinear, QDA typically performs the best
- When the true decision boundary is highly nonlinear (and there are many training samples), KNN typically performs the best

Assessing Classification Performance

As we saw in Stat 8320, for binary classification there is a standard jargon that has arisen to report classification results (beyond the classification error rate discussed previously). In particular, we consider a *confusion matrix*:

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*

and associated terminology:

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N^*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

Assessing Classification Performance (cont.)

How do we select the probability threshold for assigning observations to classes? In the two-class case, the Bayes classifier uses a threshold value of 50% (i.e., 0.5) to assign to the default class (i.e., $Y = 1$). Yet, as we change the threshold, the true positive rate and false positive rate change.

In some applications, there may be reason to want to optimize relative to one of these rather than the total error rate. This is a subjective decision. One way to look at a range of these is through the *ROC curve* (as discussed in JWHT (2009; 4.4.3)).

ROC Curves

The *ROC curve* (note: ROC stands for Receiver Operating Characteristics) provides a graphical comparison of the two most important types of classification error, *sensitivity* and *specificity*. Recall from above and from Data Analysis II that sensitivity is the "true positive" rate and (1-specificity) is the "false positive" rate.

The ROC curve plots the true positive rate on the y-axis and the false positive rate on the x-axis as we change the threshold rate for classifying an observation as positive from 0 to 1. One then looks at the overall performance of the classifier by the area under the ROC curve (AUC). An ideal ROC curve would quickly go to one and would have an AUC equal to 1. A classifier that was no better than chance would have an AUC of 0.5.

ROC Curve Example

Consider the examples from JWHT (2013) showing ROC curves for classifiers for the heart disease example training data (note, we will learn about support vector machines (SVMs) later).

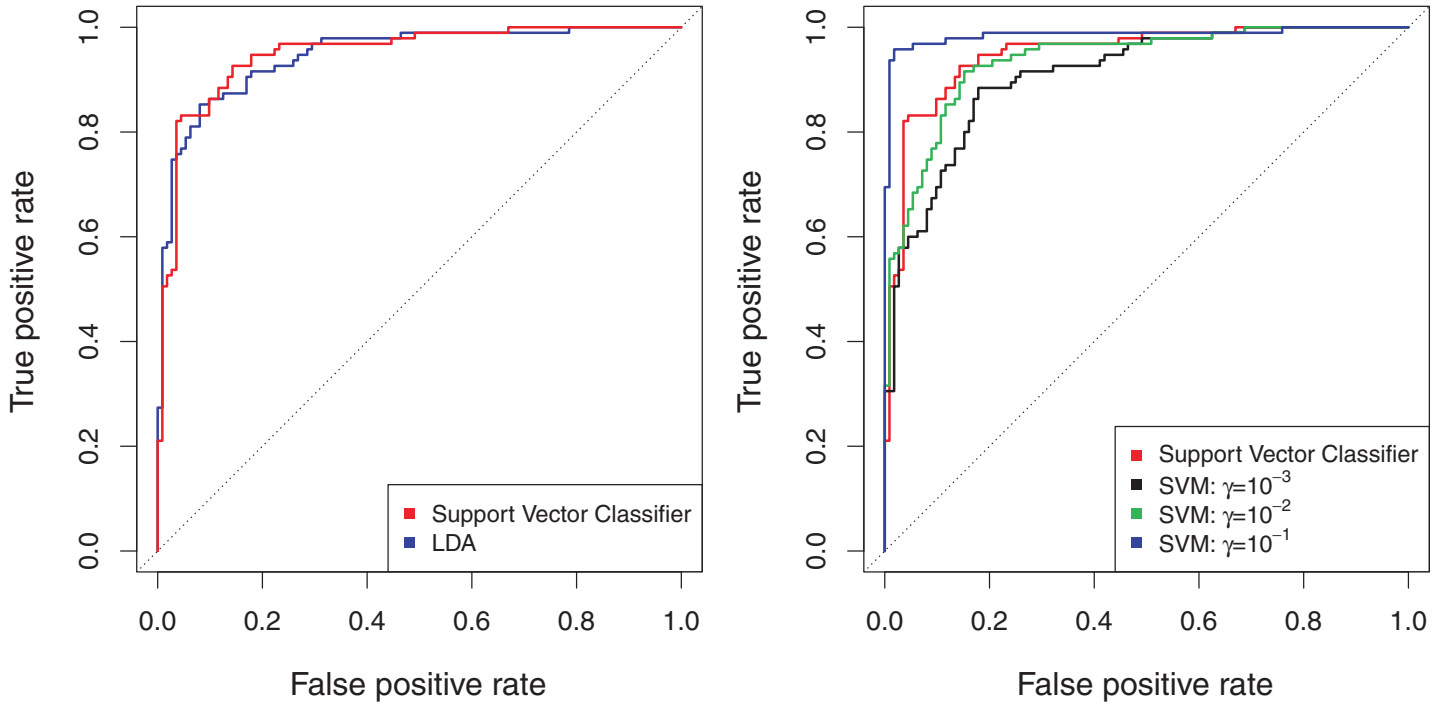

FIGURE 9.10. *ROC curves for the* `Heart` *data training set.* Left: *The support vector classifier and LDA are compared.* Right: *The support vector classifier is compared to an SVM using a radial basis kernel with* $\gamma = 10^{-3}$, 10^{-2}, *and* 10^{-1}.

Later Lectures

CKW

We will come back to the classification problem in later lectures. It is a major area of consideration in research related to statistical learning. There are many new methods that seek to solve the classification problem when p is very large.