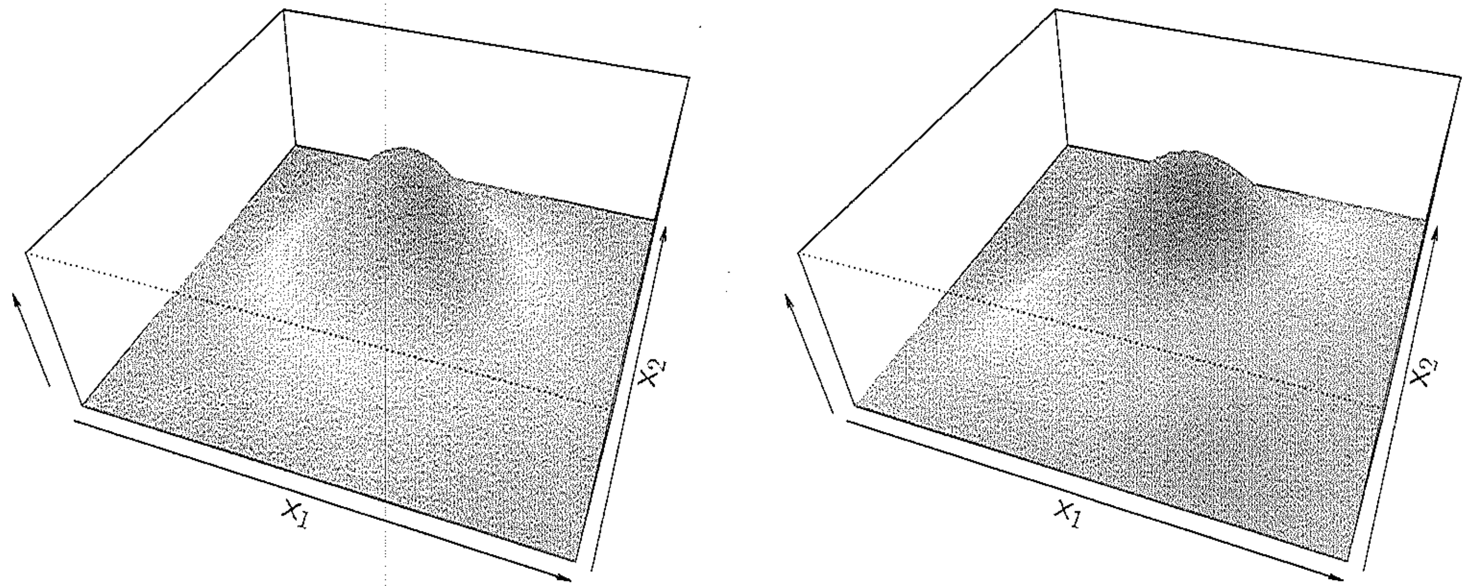
SML 2018 L09 24-Sep-2018

Linear Discriminant Analysis when $p > 1$

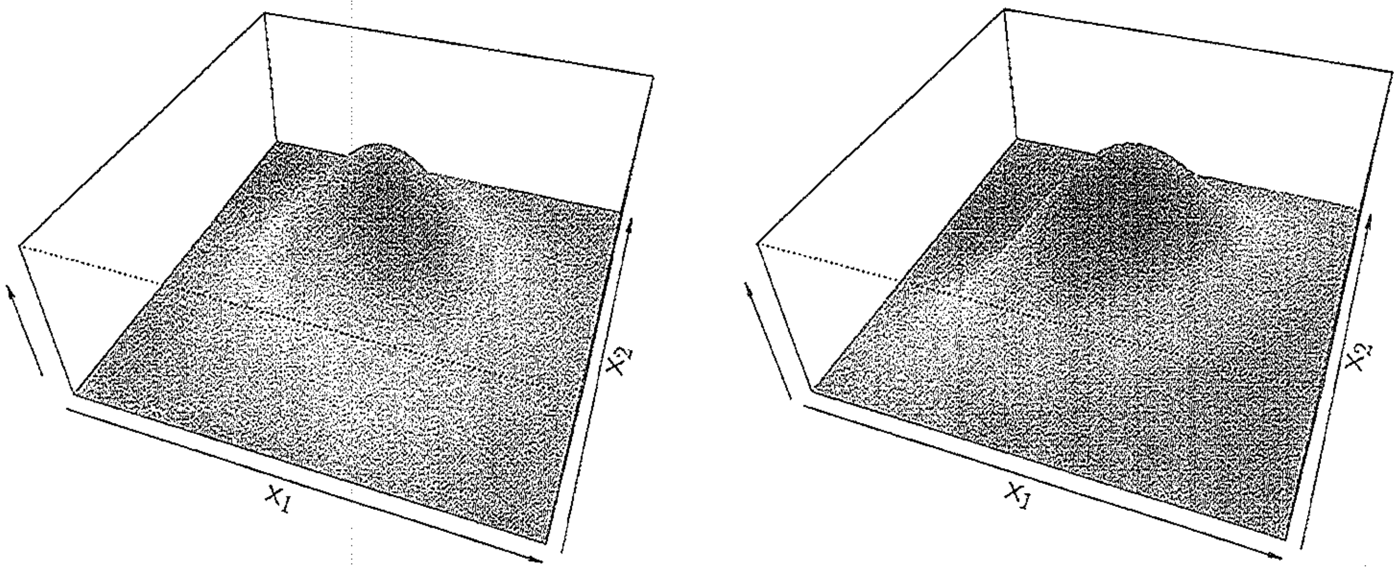


Density: $f_k(x) = \dfrac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k)}$

cancels out

$x^T \Sigma^{-1} x$ in the exponent also cancels out

Linear Discriminant Analysis when $p > 1$

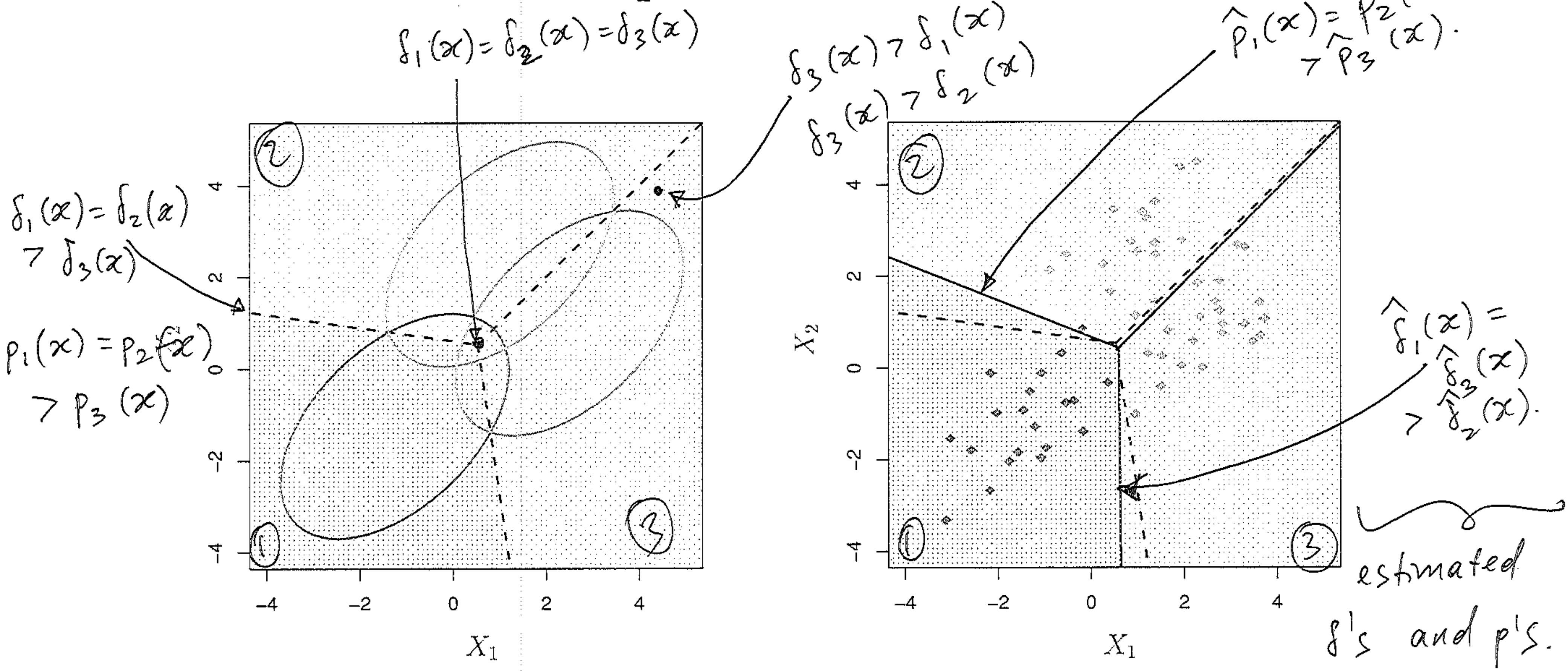


Density: $f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$

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

Despite its complex form,
$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + \ldots + c_{kp}x_p$ — a linear function.

Illustration: $p = 2$ and $K = 3$ classes



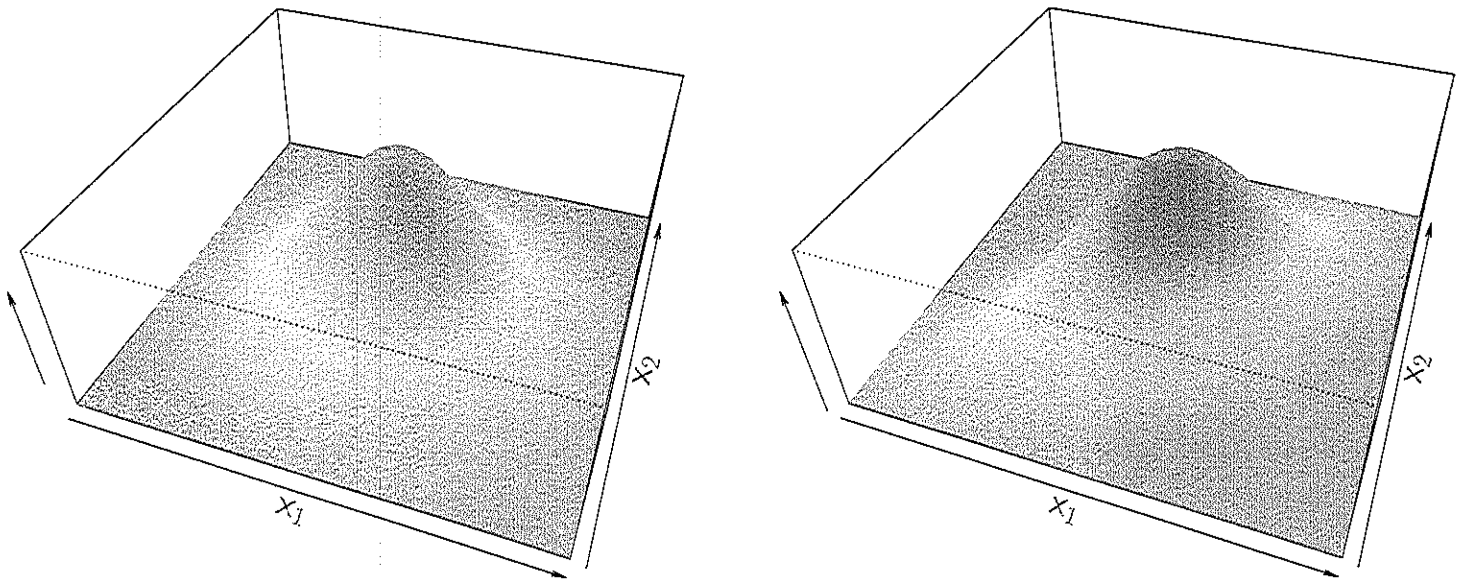
Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.

The dashed lines are known as the *Bayes decision boundaries*.

Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

Quadratic QDA

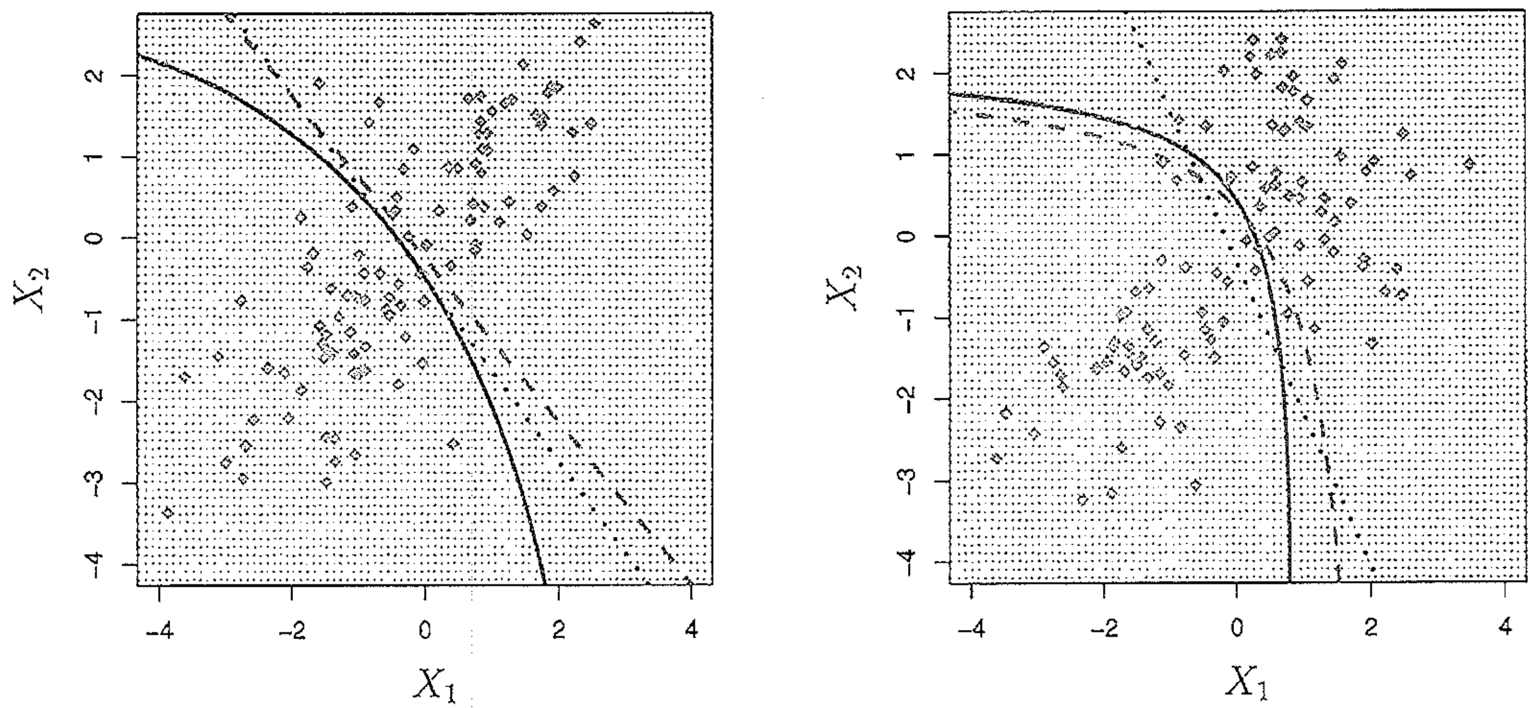
~~Linear~~ Discriminant Analysis when $p > 1$



$$\text{Density: } f_k(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma}_k^{-1}(x-\mu_k)}$$

Σ_k's and $x^T \Sigma_k^{-1} x$
no ~~not~~ longer cancel.
this leads to quadratic
decision boundaries

Quadratic Discriminant Analysis



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

Because the $\boldsymbol{\Sigma}_k$ are different, the quadratic terms matter.

no cancellations of terms involving Σ_k occur class are drawn from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common to all K classes. *Quadratic discriminant analysis* (QDA) provides an alternative approach. Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix. That is, it assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \Sigma_k)$, where Σ_k is a covariance matrix for the kth class. Under this assumption, the Bayes classifier assigns an observation $X = x$ to the class for which

quad discr analy

$$\begin{aligned} \delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \end{aligned} \quad (4.23)$$

is largest. So the QDA classifier involves plugging estimates for Σ_k, μ_k, and π_k into (4.23), and then assigning an observation $X = x$ to the class for which this quantity is largest. Unlike in (4.19), the quantity x appears as a *quadratic* function in (4.23). This is where QDA gets its name.

Why does it matter whether or not we assume that the K classes share a common covariance matrix? In other words, why would one prefer LDA to

Naive Bayes

Assumes features are independent in each class.
Useful when p is large, and so multivariate methods like QDA and even LDA break down.

- Gaussian naive Bayes assumes each Σ_k is diagonal:

$$\delta_k(x) \propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] = -\frac{1}{2}\sum_{j=1}^{p}\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log\pi_k$$

$-\frac{1}{2}\log|\Sigma_k|$

do not lose the log det piece

- can use for *mixed* feature vectors (qualitative and quantitative). If X_j is qualitative, replace $f_{kj}(x_j)$ with probability mass function (histogram) over discrete categories.

Despite strong assumptions, naive Bayes often produces good classification results.

rate and a low false positive rate. The dotted line represents the "no information classifier; this is what we would expect if student status and credit card balanc are not associated with probability of default.*

DEF'S FROM ISLR

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

minus the *specificity* of our classifier. Since there is an almost bewildering array of terms used in this context, we now give a summary. Table 4.6 shows the possible results when applying a classifier (or diagnostic test to a population. To make the connection with the epidemiology literature we think of "+" as the "disease" that we are trying to detect, and "−" as the "non-disease" state. To make the connection to the classical hypothesis testing literature, we think of "−" as the null hypothesis and "+" as the alternative (non-null) hypothesis. In the context of the `Default` data, "+'

EXAMPLES: TN, FN, FP and TP on

LDA ~~on~~ for Credit Data

		True Default Status			
		No	Yes	Total	
Predicted Default Status	No	9644	252	9896	N^*
	Yes	23	81	104	P^*
	Total	9667	333	10000	
		N	T	Total	

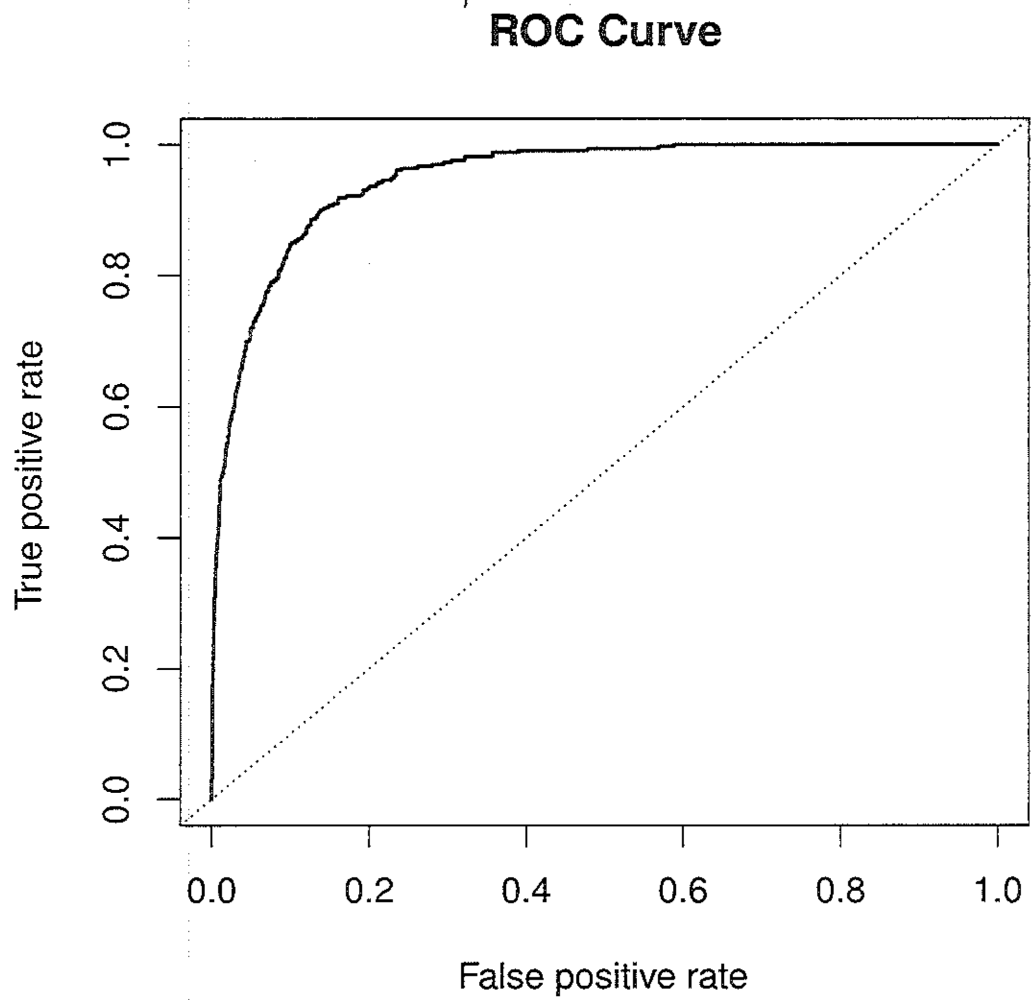
TN, FN, FP, TP

$(23 + 252)/10000$ errors — a 2.75% misclassification rate!

Some caveats:

- This is *training* error, and we may be overfitting.

Receiver Operating Characteristic (ROC)



Pick a grid of threshold values $t_1, \ldots, t_N$ for each t_i, compute $FPR(t_i)$, $TPR(t_i)$ then plot TPR vs FPR

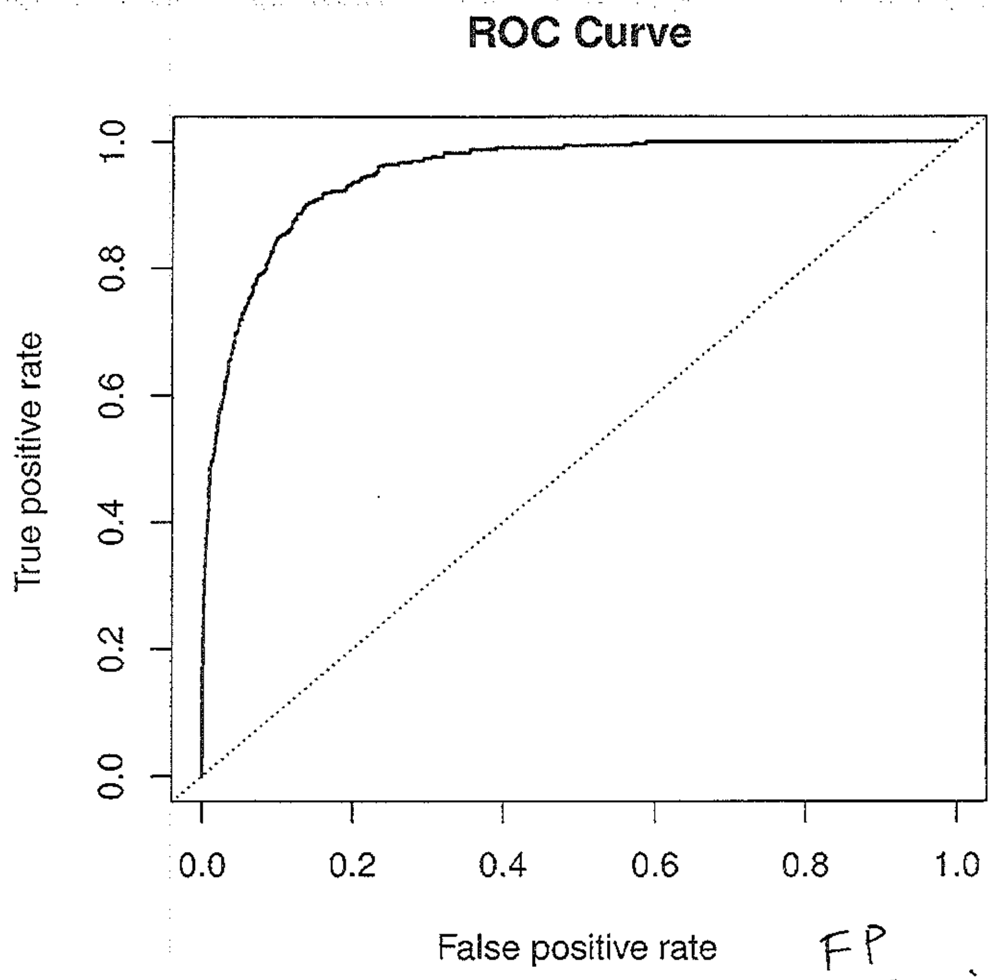
The *ROC plot* displays both simultaneously.

Idea:
for a grid of threshold values, compute the corresp. FP rate and TP rate. Then plot TP rate vs FP rate.

$\frac{TP}{P}$

P = TP + FN
(observed)

P* = TP + FP
(predicted)



$\frac{FP}{N}$; N = TN + FP
(observed)

The *ROC plot* displays both simultaneously.

QUESTION: WHAT CLASSIFIER DOES THE DIAGONAL LINE CORRESPOND to?