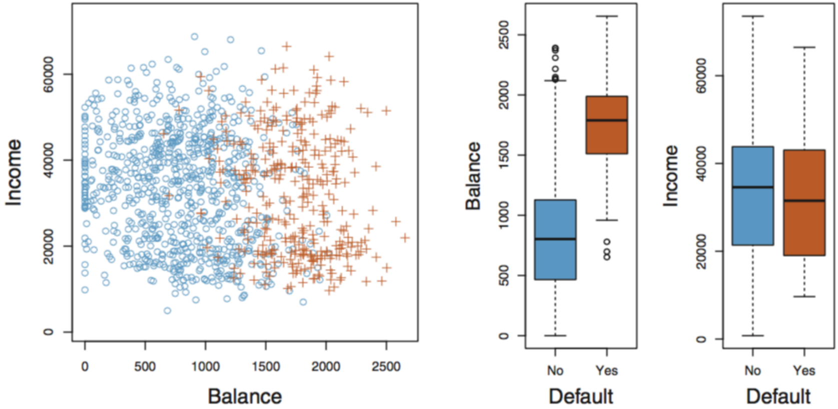
Lecture 8: Classification (Textbook 4.4)

Default Data



It shows 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.

Linear Discriminant Analysis

- Logistic regression involves directly modeling $P(Y = k|X = x)$.
- Here, **linear discriminant analysis** is to model the distribution of X in each of the classes separately, and then use Bayes' theorem to flip things around and obtain $P(Y = k|X = x)$. The Bayes' theorem is

$$P(Y = k|X = x) = P(X = x|Y = k)P(Y = k)/P(X = x).$$

- We usually assume the distribution of X in each of the classes to be normal distributions.

Why not Logistic Regression

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is more convenient when we have more than two response classes. (Multinomial logistic regression and proportional odds model)

Linear Discriminant Analysis

- Logistic regression involves directly modeling $P(Y = k|X = x)$.
- Here, **linear discriminant analysis** is to model the distribution of X in each of the classes separately, and then use Bayes' theorem to flip things around and obtain $P(Y = k|X = x)$. The Bayes' theorem is

$$P(Y = k|X = x) = P(X = x|Y = k)P(Y = k)/P(X = x).$$

- We usually assume the distribution of X in each of the classes to be normal distributions.

Using Bayes' Theorem for Classification

- Recall that the Bayes' theorem is

$$P(Y = k|X = x) = P(X = x|Y = k)P(Y = k)/P(X = x).$$

- We can slightly rewrite it as

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)},$$

  π_k is the **prior** probability that a randomly chosen observation comes from the kth class, i.e. $P(Y = k)$.

  $f_k(X) = P(X = x|Y = k)$ denotes the **density function** of X for an observation that comes from the kth class.

  $p_k(x) = P(Y = k|X = x)$ is called **posterior** probability. It is the probability that the observation belongs to the kth class, given the predictor value for that observation.

- In the ideal case, we classify a new point according to which posterior probability is highest.

Linear Discriminant Analysis for $p = 1$

- We assume that $f_k(x)$ is normal, which takes the form

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

  where μ_k and σ_k^2 are the mean and variance parameters for the kth class. We assume all $\sigma_k = \sigma$ are the same.

- Plugging this into Bayes' formula, we get $p_k(x) = P(Y = k|X = x)$ as

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

Linear Discriminant Analysis for $p = 1$

- To classify at the value $X = x$, we need to see which k has the largest $p_k(x)$.

- Taking logs, and discarding terms that do not depend on k, the Bayes classifier is to assign x to the class with the largest

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

  Note that $\delta_k(x)$ is a linear function of x. That is why it is called linear discriminant analysis (LDA).

- If $K = 2$ and $\pi_1 = \pi_2$, then the Bayes decision boundary corresponds to

$$x = \frac{\mu_1 + \mu_2}{2}.$$

Discriminant functions

- Given training data, we estimate μ_k, and σ by

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

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

  where n_k is the number of training observations in the kth class. We also estimate π_k by

$$\hat{\pi}_k = n_k/n.$$

- Plugging the estimates into $\delta_k(x)$, we get

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

  which is called **discriminant function**.

- LDA just assigns x to the class with the largest $\hat{\delta}_k(x)$.

Linear Discriminant Analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors $X = (X_1, ..., X_p)$.
- Recall that the posterior probability has the form

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)},$$

- Now, we assume $X|Y = k$ follows a multivariate normal distribution $N(\mu_k, \Sigma)$,

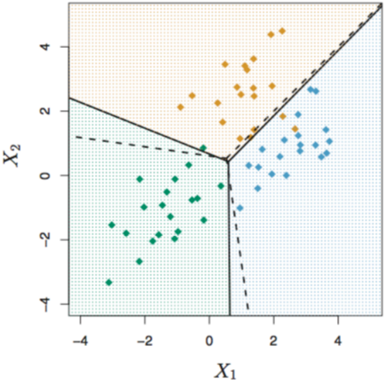
$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}.$$

- Similarly, we assign x to the class with the largest

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

- The Bayes decision boundaries are the set of x for which $\delta_k(x) = \delta_l(x)$ for $k \neq l$. Again, the boundaries are collection of straight lines, since $\delta_k(x)$ is linear in x.

Example



There are three classes (orange, green and blue) with two predictors X_1 and X_2. Dashed lines are the Bayes decision boundaries. Solid lines are their estimates based on the LDA.

LDA on the Default Data

Our goal is predict whether or not an individual will default on the basis of credit card balance.

		True default status		
		No	Yes	Total
Predicted	No	9,644	252	9,896
default status	Yes	23	81	104
	Total	9,667	333	10,000

The training error rate is $(23 + 252)/10000 = 2.75\%$. For a credit card company that is trying to identify high-risk individuals, an error rate of $252/333 = 75.7\%$ among individuals who default is unacceptable.

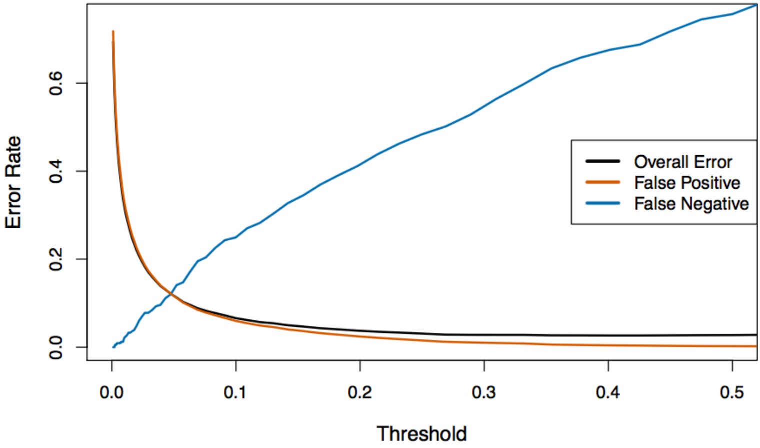
Types of Errors

- **False positive rate (FPR)**: The fraction of negative examples that are classified as positive – $23/9667 = 0.2\%$ in default data.
- **False negative rate (FNR)**: The fraction of positive examples that are classified as negative – 75.7% in default data.
- The false negative rate is too high.
- We can achieve better balance of FPR and FNR by varying the threshold

  $$\text{P(default=yes} \mid \text{X=x)} > \text{threshold},$$

  for some threshold different from 0.5.

Trade-off between FPR and FNR



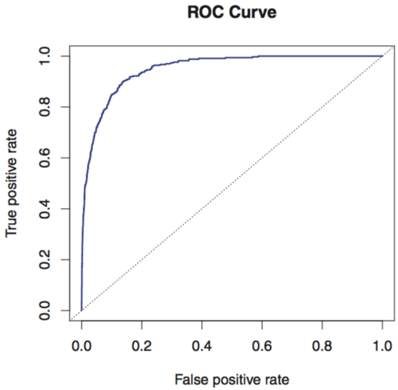
More Terminology

		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*	

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*	

This defines **sensitivity** and **specificity**.

ROC Curve



The ROC plot displays both FPR and TPR. A ROC curve for the LDA classifer on the Default data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The ideal ROC curve hugs the top left corner. The dotted line represents the "no information" classifer; this is what we would expect if student status and credit card balance are not associated with probability of default.