Classification

Paul Martins

Notes from ISLR book

Contents

1. Models

The models below are all classifiers. The goal is to assign the correct label to the response given a set of p predictors X.

1.1 Logistic Regression

The logistic regression estimates the conditional distribution $\hat{p}$ of the response Y given the predictors X. To keep that probability between 0 and 1, the logistic regression uses the logistic function:

$$p(x) = \frac{e^x}{1+e^x} \tag{1}$$

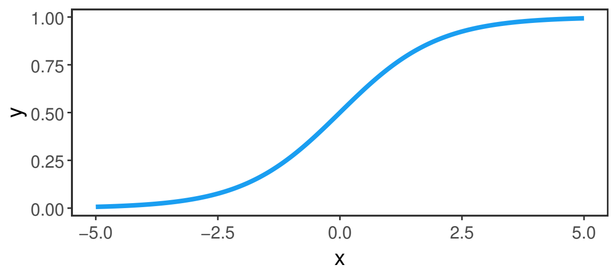

Figure 1. The logistic function

The estimation of the probability gives

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1X_1+\ldots+\hat{\beta}_pX_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1X_1+\ldots+\hat{\beta}_pX_p}} \tag{2}$$

where the coefficient estimates $\hat{\beta}$ are found through the maximum likelihood method. Note that the logit (or log-odds) is linear in X [1]

$$log\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right) = \hat{\beta}_0+\hat{\beta}_1X_1+\ldots+\hat{\beta}_pX_p \tag{3}$$

The logistic regression limitations are

- The parameter estimates are quite unstable, especially when the classes are well separated or when the number of observations n is small and the predictor distributions are gaussian.
- The model becomes harder to interpret when the response classes goes beyond 2.

[1]The quantity $\frac{p(X)}{1-p(X)}$ is called the odd.

1.2 Linear discriminant analysis

LDA uses Bayes theorem to find the estimation of the posterior probability, the probability that the response Y is of class k given that the predictors $X = x$

$$\hat{p}_k(x) = Pr(Y=k|X=x) = \frac{\hat{\pi}_k\hat{f}_k(x)}{\sum_{l=1}^K \hat{\pi}_l\hat{f}_l(x)} \tag{4}$$

where $\hat{\pi}_k$ is the estimate of the overall prior probability that a randomly chosen observation comes from class k. It is by default $\hat{\pi}_k = n_k/n$.

$\hat{f}_k(x) = Pr(X=x|Y=k)$ is the density function estimation of X for an observation from the k^{th} class. The linear discriminant assumes that $\hat{f}_k$ has a multivariate Gaussian distribution with $\hat{\mu}_k$ the estimated mean vector of X (with p components) and $\hat{\boldsymbol{\Sigma}} = Cov(X)$ the $p \times p$ estimated covariance matrix common to all K classes.[2]

$$\hat{f}_k(x) = \frac{1}{(2\pi)^{p/2}|\hat{\boldsymbol{\Sigma}}|^{1/2}} exp\Big(-\frac{1}{2}(x-\hat{\mu}_k)^T\hat{\boldsymbol{\Sigma}}^{-1}(x-\hat{\mu}_k)\Big) \tag{5}$$

Plugging everything into 4, we can show that the response is assigned the class k for which

$$\hat{\delta}_k(x) = x^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\boldsymbol{\Sigma}}^{-1} + log(\hat{\pi}_k) \tag{6}$$

is the largest. This discriminant is linear in X, hence the name of the model. The linear decision boundaries can be find by finding the x such that $\hat{\delta}_k = \hat{\delta}_l, k \neq l$.

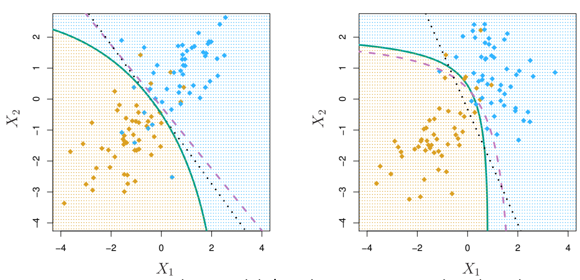

Figure 2. LDA (dotted black), Bayes (dashed purple) and QDA (green) decision boundaries. Left: the 2 classes have the same vaiance. Right: the 2 classes have different variances.

[2]In the case where $p=1$ predictor, $\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i$ and $\hat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^K\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2$, which gives $\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k)$

1.3 Quadratic discriminant analysis

The difference of QDA with LDA is that each class can have its own covariance matrice: $\hat{\mathbf{\Sigma}} \rightarrow \hat{\mathbf{\Sigma}}_k$. The discriminant is now quadratic in X

$$\hat{\delta}_k(x) = -\frac{1}{2}(x-\hat{\mu}_k)^T \hat{\mathbf{\Sigma}}_k^{-1}(x-\hat{\mu}_k) - \frac{1}{2} log|\hat{\mathbf{\Sigma}}_k| + log\hat{\pi}_k \quad (7)$$

QDA is more flexible since each class get to have a covariance matrix. This leads to quadratic decision boundaries, but also to the estimation of $Kp(p+1)/2$ parameters, against only Kp for LDA. LDA is generally recommended when there are few training observations (to reduce the variance) while QDA is a better choice with large number of training observations.

2. Diagnostics

A common diagnostic for classifier is the confusion matrix as shown in Table 1 where the "+" or Non-null class defines the specific response we want the label for.

		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 1. Confusion matrix

This allows to easily compute the measures for classification diagnostic in Table 2.

Name	Definition	Alternative
True Neg. rate	TN/N	Specificity
False Pos. rate	FP/N	1-Specificity
True Pos. rate	TP/P	Sensitivity, Recall
Pos. Pred. rate	TP/P*	Precision
Neg. Pred. rate	TN/N*	
Accuracy	(TN+TP)/(N+P)	

Table 2. Diagnostic measures

Specificity: percentage of true negative observations correctly identified by the model.

Sensitivity (recall): percentage of true positive observations correctly identified by the model.

Precision: percentage of predicted positive observations correctly identified by the model.

Accuracy: percentage of total observations correctly identified by the model.

Classifiers derived from Bayes classifiers such as LDA and QDA assign labels to the highest probability class (the class k which has the highest $\hat{\delta}_k$). If $K = 2$, it corresponds to a threshold of 0.50. This threshold can be tuned, which will affect the above values.

Two curves are used to summarised these measures: the ROC curve[3] that displays recall vs. 1-specificity and the precision-recall curve.

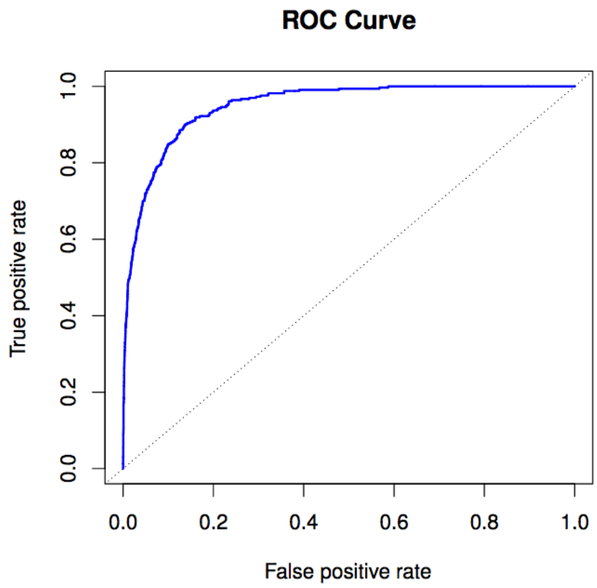

Figure 3. The ROC Curve

[3] Receiver Operating Characteristics