Chapter 4. Classification methods
Part 3

March 28, 2007

1 Classification with more than 2 classes

Suppose each sample $X = (\mathbf{x}_1, ..., \mathbf{x}_p)$ belongs to one of J classes, J is 2 or more. Denote the classes by $C_j, j = 1, ..., J$. How can we do the classification?

1.1 Classification by multivariate linear regression

Suppose we have sample $X_i, i = 1, ..., n$. We form J indicator responses $Y_1, ..., Y_J$: if $X_i \in C_j$, then let $Y_{j,i} = 1$ otherwise $Y_{j,i} = 0$ (or -1). We also call the values of Y the **scores**.

Now for each Y_j, we can use linear (or other models such as logistic model, single-index model, MARS, PPR) to fit the data

$$
\begin{aligned}
Y_{j,1} &= \beta_{j0} + \beta_j^\top X_1 + \varepsilon_{j1} \\
Y_{j,2} &= \beta_{j0} + \beta_j^\top X_2 + \varepsilon_{j1} \\
&... \\
Y_{j,n} &= \beta_{j0} + \beta_j^\top X_n + \varepsilon_{j1}.
\end{aligned}
$$

$j = 1, 2, ...J$. let $b_j = (\beta_{j0}, \beta_j^\top)^\top$

$$
\mathbf{X} = \begin{pmatrix} 1 & X_1^\top \\ 1 & X_2^\top \\ ... & \\ 1 & X_n^\top \end{pmatrix}, \quad \mathbf{Y} = (Y_1, ..., Y_J) = \begin{pmatrix} Y_{1,1} & Y_{2,1} & ... & Y_{J,1} \\ Y_{1,2} & Y_{2,2} & ... & Y_{J,2} \\ ... & & & \\ Y_{1,n} & Y_{2,n} & ... & Y_{J,n} \end{pmatrix}
$$

$$
B = \begin{pmatrix} \beta_{10} & \beta_{20} & ... & \beta_{J0} \\ \beta_1 & \beta_2 & ... & \beta_J \end{pmatrix} = (b_1, ..., b_J).
$$

We have J models with the same X. The estimation for each model is based on minimizing

$$
||Y_j - \mathbf{X} b_j||^2
$$

The solution (estimator) is

$$\hat{b}_j = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X} Y_j$$

Then the fitted Y_j is

$$\hat{Y}_j = \mathcal{S} Y_j$$

where $\mathcal{S} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. The fitted error is

$$||Y_j - \mathbf{X} b_j||^2 = ||(I - \mathcal{S}) Y_j||^2$$

Then the models can be written as

$$\mathbf{Y} = \mathbf{X} B + \mathcal{E}$$

To estimate B, we need to minimize

$$\sum_{j=1}^{J} ||Y_j - \mathbf{X} b_j||^2 = tr\{(\mathbf{Y} - \mathbf{X}B)^\top (\mathbf{Y} - \mathbf{X}B)\}$$

Again, the estimator is

$$\hat{B} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

The fitted error is

$$\begin{aligned} \sum_{j=1}^{J} ||Y_j - \mathbf{X} \hat{b}_j||^2 &= tr\{(\mathbf{Y} - \mathbf{X}\hat{B})^\top (\mathbf{Y} - \mathbf{X}B)\} \\ &= tr\{\mathbf{Y}^\top (I - \mathcal{S}) \mathbf{Y}\} \end{aligned}$$

Now for a new sample X_{new}, we can predict its Y by

$$\hat{Y}_{new} = (\hat{Y}_{new,1}, ..., \hat{Y}_{new,J}) = (1, X_{new}^\top)\hat{B}$$

We class X_{new} bassed on softmax probability

$$\hat{p}_j = \frac{\exp(\hat{Y}_{new,j})}{\exp(\hat{Y}_{new,1}) + ... + \exp(\hat{Y}_{new,J})}$$

Note that

$$\hat{p}_1 + ... + \hat{p}_J = 1$$

$\hat{p}_j$ can be taken as the probability that $X_{new} \in C_j$. We can classify it easily based on the probability.

1.2 Optimal Scores

A simple criterion is that the fitted error should be small. One way to achieve this is by optimizing the scores. The original score is $\mathbf{Y}$. we consider a matrix $\Theta : J \times K$ with $K \leq J$ such that $\Theta^\top(\mathbf{Y}^\top\mathbf{Y})\Theta = I$ (identity matrix). K is called the dimension. The new score is

$$\mathbf{Y}^* = \mathbf{Y}\Theta$$

How to find the score? we need to minimize the fitted error

$$tr\{(\mathbf{Y}^*)^\top(I - \mathcal{S})\mathbf{Y}^*\} = tr\{\Theta^\top\mathbf{Y}^\top(I - \mathcal{S})\mathbf{Y}\Theta\}$$

Algorithm

Step 1

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{B}$$

Step 2 We optimize scores by matrix Θ which is the eigenvector matrix of $\hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}$ with normalization $\Theta^\top(\mathbf{Y}^\top\mathbf{Y})\Theta$.

Step 3 Go to step 1 with $\hat{B}$ replaced by $\hat{B}\Theta$.

1.3 Flexible discriminants analysis (FDA)

A refined version of the above approach is the Flexible discriminants analysis.

Besides linear regression model, we have other models for the relation between Y_j and X. Examples are PPR and MARS.

2 Examples

Example 2.1 *Speaker independent recognition of the eleven steady state vowels of British English using a specified training set of lpc derived log area ratios. (***(training set)** *,* **(validation set)***), we use SVM and fda to classify the data. The response variable has 11 categories. There are 10 covariates* $\mathbf{x}_1, ..., \mathbf{x}_{10}$*. we use the training data to estimate the separating plane and validation set to check the methods.*

*SVM method: The error rate for the testing set is 0.3831169 (using kernel='radial', gamma = 0.3) (***(code)***)*

*FDA method: The error rate for the testing set is 0.4935065 (using method = mars, degree = 2); 0.5692641 (using method = ppr, nterms = 2); (***(code)***)*

*CART method: The error rate for the testing set is 0.6082251 (***(code)***)*

Example 2.2 *The Waveform data was designed to check the performance of classification methods. The data is generated by*

$$X_i = U * h_1(j) + (1 - U) * h_2(i) + \epsilon_j \qquad \textit{class 1}$$

$$X_i = U * h_1(j) + (1 - U) * h_3(i) + \epsilon_j \qquad \textit{class 2}$$

$$X_i = U * h_2(j) + (1 - U) * h_3(i) + \epsilon_j \qquad \textit{class 3}$$

where $j = 1, ..., 21$. U is uniformly on (0, 1) and ϵ_i are N(0, 1). The h_ℓ are shifted triangular waveforms: $h_1(i) = \max(6 - |j - 11|, 0), h_2(i) = h_1(j - 4)$ and $h_3(i) = h_1(i + 4)$.

*With 300 (***(training points)***, and 500 ***(validation points)***),*

*SVM method: The error rate for the testing set is 0.164 (using kernel='radial') (***(code)***)*

FDA method: The error rate for the testing set is 0.192 (using method = mars, degree = 2)

Some times the classification can be visualized in two dimensional space. See figure 1 for the waveform data.

References

Hastie, Tibshirani and Buja (1994) Flexible Disriminant Analysis by Optimal Scoring *J. Amer. Stat. Ass*, 1255-1270

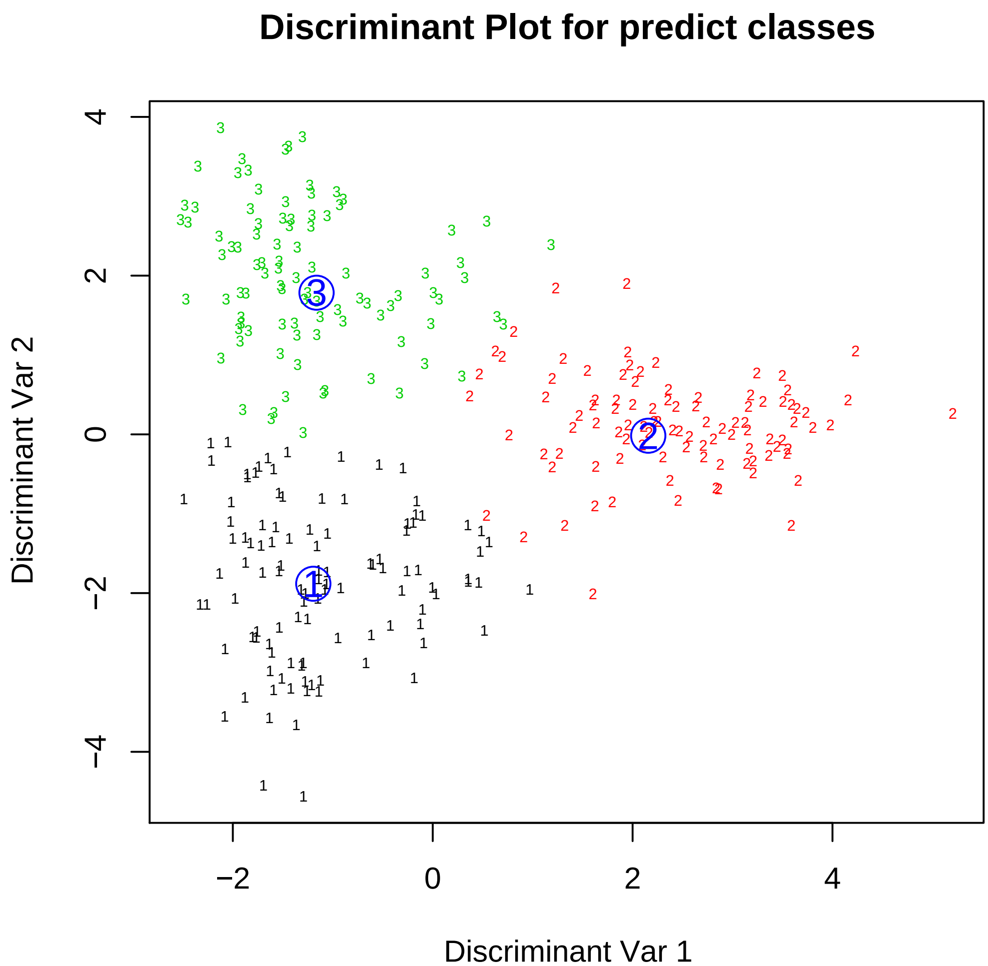

Figure 1: plot(output)