Reducing Variance With Logistic Regression

We see from this example that logistic regression can also suffer from "overfitting", due to high variance in the estimated βs.

The possible solutions are much the same as for ordinary regression:

- Select a subset of variables.
- Select a small set of linear combinations of variables (eg, PCs).
- Maximize the log likelihood minus a penalty.
- Use Bayesian methods (not yet covered).

I tried introducing a penalty for the vowel example, but it seems it doesn't help. Perhaps performance is limited by bias from the linearity restriction, not by variance.

Reducing Variance With LDA and QDA

In this example, LDA seems to suffer more from high bias than from high variance — the class densities seem to be nothing like Gaussians with the same covaraince. QDA seems to overfit a lot, however.

To reduce variance with LDA or QDA, we could select a subset of variances, or use some number of PCs. Two other possibilities are discussed in the book.

A compromise between LDA and QDA uses class covariance matrices of the form

$$\hat{\Sigma}_k(\alpha) \;\; = \;\; \alpha\hat{\Sigma}_k \;+\; (1-\alpha)\hat{\Sigma}$$

This is a weighted average of the QDA and LDA matrices. We'd need to choose α by cross validation. We'd hope that in this way we'd get lower bias than LDA, without all the variance from QDA.

In the opposite direction, we can restrict LDA to try to reduce its variance (presumably at the cost of more bias).

Reduced Rank LDA

Note: This slide isn't quite right. It needs to be revised...

The common covariance matrix of LDA can be used to "sphere" the data, by multiplying $\mathbf{X}$ (assumed centred) by $\hat{\Sigma}^{-1/2}$. The resulting sample covariance of $\mathbf{W} = \mathbf{X}\hat{\Sigma}^{-1/2}$ will be

$$(1/N)\mathbf{W}^T\mathbf{W} \;=\; (1/N)\hat{\Sigma}^{-1/2}\mathbf{X}^T\mathbf{X}\hat{\Sigma}^{-1/2} \;=\; \hat{\Sigma}^{-1/2}\,\hat{\Sigma}\,\hat{\Sigma}^{-1/2} \;=\; \mathbf{I}$$

Once we've done this, the linear discriminants are just finding the class whose mean is closest to the test point. With K classes, the class means lie in a $K-1$ dimensional subspace. We can project into this space and get the same results.

What's more, we can do PCA on the means, and use the first $M < K-1$ of them. This may reduce variance.

When $M = 2$, we can visualize the situation in a scatterplot, as done in Figure 4.4 in the book.