Model Free Classification

Rebecca C. Steorts, Duke University

STA 325, Chapter 4 ISL

Agenda

- Classification revisited
- LDA revisited
- Model free classification
- K-nearest neighbors
- K-means classification

LDA versus Logistic regression

As we remarked earlier, both LDA and logistic regression model the log odds as a linear function of the predictors $x \in \mathbb{R}^p$

$$\text{Linear discriminant analysis:} \quad \log\left\{\frac{\mathcal{P}(C=1|X=x)}{\mathcal{P}(C=2|X=x)}\right\} = \alpha_0 + \alpha^T x$$

$$\text{Logistic regression:} \quad \log\left\{\frac{\mathcal{P}(C=1|X=x)}{\mathcal{P}(C=2|X=x)}\right\} = \beta_0 + \beta^T x$$

where for LDA we form $\hat{\alpha}_0, \hat{\alpha}$ based on estimates $\hat{\pi}_j, \hat{\mu}_j, \hat{\Sigma}$ (easy!), and for logistic regression we estimate $\hat{\beta}_0, \hat{\beta}$ directly based on maximum likelihood (harder)

This is what leads to linear decision boundaries for each method

Careful inspection (or simply comparing them in R) shows that the estimates $\hat{\alpha}_0, \hat{\beta}_0$ and $\hat{\alpha}, \hat{\beta}$ are different. So how do they compare?

LDA versus logistic regression

Generally speaking, logistic regression is more flexible because it doesn't assume anything about the distribution of X. LDA assumes that X is normally distributed within each class, so that its marginal distribution is a mixture of normal distributions, hence still normal:

$$X \sim \sum_{j=1}^{K} \pi_j N(\mu_j, \Sigma)$$

This means that logistic regression is more robust to situations in which the class conditional densities are not normal (and outliers)

On the other side, if the true class conditional densities are normal, or close to it, LDA will be more efficient, meaning that for logistic regression to perform comparably it will need more data

In practice they tend to perform similarly in a variety of situations (as claimed by the ESL book on page 128)

Model-free classification

The downside: these methods are essentially a black box for classification, in that they typically don't provide any insight into how the predictors and the response are related

The upside: they can work well for prediction in a wide variety of situations, since they don't make any real assumptions

These procedures also typically have tuning parameters that need to be properly tuned in order for them to work well (for this we can use cross-validation)

Classification by k-nearest-neighbors

Perhaps the simplest prediction rule, given labeled data (x_i, y_i), $i = 1 \ldots n$, is to predict an input $x \in \mathbb{R}^p$ according to its nearest-neighbor:

$$\hat{f}^{1-\mathrm{NN}}(x) = y_i \quad \text{such that } \|x_i - x\|_2 \text{ is smallest}$$

A natural extension is to consider the k-nearest-neighbors of x, call them $x_{(1)}, \ldots x_{(k)}$, and then classify according to a majority vote:

$$\hat{f}^{k-\mathrm{NN}}(x) = j \quad \text{such that } \sum_{i=1}^{k} 1\{y_{(i)} = j\} \text{ is largest}$$