Introduction to Statistical Learning Chapter 4

Cheat-sheet

Classification

Classification refers to problems where the response is qualitative or categorical. These terms are interchangeable.

Some classification methods compute the probabilities that a given observation belongs to a certain class to determine its predicted response.

Logistic regression, linear discriminant analysis, quadratic discriminant analysis, naive Bayes, K-nearest neighbours, generalized additive models, trees, random forests, boosting and support vector machines are some of the classification methods.

An Overview of Classification

Like regression, we have a set of training observations we can use to construct our classifier.

Why Not Linear Regression?

When we attempt to encode the classes, we are placing an ordering on the classes and deciding that the difference between any set of classes is equal.

In a binary (two-level) qualitative response, we can in fact encode the response classes (for example, as 0 and 1).

Performing linear regression on a two-level response gives us the same classifications as the linear discriminant analysis method.

Logistic Regression

This method models the probability Y belongs to a given class.

Logistic function: $p(X) = \frac{e^{\beta_0+\beta 1 X}}{1+e^{\beta_0+\beta 1 X}}$

LHS is odds: $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$

LHS is log odds or logit: $\log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$

Maximum likelihood is used to fit the model since it has better statistical properties.

Maximum likelihood function is maximised when finding coefficient estimates:
$\ell(\beta_0, \beta_1) = \prod p(x_i) \prod (1 - p(x_{i'}))$

A large absolute value of the z-statistic indicates the probability that the parameter is not 0 is high.

Dummy variables can be implemented into the model.

An extension to multiple logistic regression is straight-forward, where we simply extend the model for p predictors.

The phenomenon where a predictor has influence on both the predictor and the response is known as confounding (we detect a correlation between predictors).

Multinomial logistic coding is a method for being able to perform logistic regression on more than two classes. Softmax coding is equivalent to the above (used extensively in machine learning literature).

Generative Models for Classification

In the new approach, we model the distribution of the predictors separately for each of the response classes.

These methods can be preferable as logistic regression can be surprisingly unstable for classes that have substantial separation, they can be more accurate if the distributions of the predictors are approximately normal in each of the classes, and they can more easily implement 3 or more class responses.

Bayes theorem: $Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum \pi_l f_l(x)}$ where π_k is the overall or prior probability and $f_k(x)$ is a density function; $Pr(Y = k|X = x)$ is the posterior probability (also denoted $p_k(x)$).

Estimating π_k is easy: simply determine the fraction of training data observations in a random sample belong to the k^{th} class.

Linear Discriminant Analysis

To estimate $f_k(x)$, we first assume its form is normal or Gaussian. In 1-d, the normal density takes the form $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} exp(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2)$

We assign an observation to the class for which $p_k(x)$ is highest, which can be reduced to the class for which the following is greatest:
$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$.

x = $\frac{\mu_1+\mu_2}{2}$ are the set of points for which $\delta_1 = \delta_2$, or the Bayes decision boundary.

The linear discriminant analysis method approximates the Bayes classifer by estimating π_k, μ_k and σ^2.

In LDA, the estimates are $\hat{\mu}_k = \frac{1}{n_k}\sum x_i$, $\hat{\sigma}_k = \frac{1}{n-K}\sum\sum(x_i - \mu_k)^2$ and $\pi_k = \frac{n_k}{n}$.

The discriminant functions $\hat{\delta_k}$ are linear in x.

LDA for $p > 1$

We assume that X is drawn from a multivariate Gaussian distribution, with a class-specific mean vector and a common covariance matrix.

The multivariate Gaussian distribution assumes each predictor follow a one-dimensional normal distribution and there is correlation between every pair of predictors.

The Bayes classifier assigns an observation to the class for which $\delta_k(x) = x^T \sum^{-1} \mu_k$ - $\frac{1}{2}\mu_k^T \sum^{-1} \mu_k$ + log(π_k) is largest.

A confusion matrix is useful for identifying which category is undergoing more incorrect classification.

Sensitivity: in the example, the number of true defaulters correctly identified

Specificity: in the example, the number of non-defaulters correctly identified

Quadratic Discriminant Analysis

QDA assumes each class has its own covariance matrix.

Bayes classifier assigns an observation to the class for which $\delta_k(x)$ is highest: equation on pg 156.

x appears as a quadratic quantity in the discriminant function.

LDA requires estimating Kp parameters whereas QDA requires estimating Kp(p+1)/2 parameters.

Naive Bayes

Naive Bayes operates under the assumption that under the k^{th} class, the p predictors are independent.

Ways to estimate the one dimensional density function f_{kj}: assume that the predictor within each class follows a uni-variate normal distribution, construct a histogram and determine the fraction of training observations that belong in the k^{th} histogram bin or determine the proportion of training observations for the j^{th} predictor corresponding to each class.

Analysis of Methods

LDA is a special case of QDA.

LDA is a special case of Naive Bayes.

Naive Bayes can be a special case of LDA.

Because KNN is non-parametric, we can infer that it will dominate LDA and Naive Bayes when the true decision boundary is highly non-linear, when n is large and p is small.