Boxi Huang's Reading Note bh563@cornell.edu

Page Covered: PP 127-154

Classification

1. Qualitative variables are referred to as categorial.
2. **Classification** is a process to predict qualitative responses.
3. Predicting a qualitative response for observation can be referred to as *classifying* that observation since it involves assigning the observation to a category, or class.

In a binary classification situation, we can use the dummy variable approach to code the response variable as 0/1 and fit a linear regression to this binary response and predict whether the response variable is above a certain threshold.

Logistic Regression

Logistic regression models the probability that Y belongs to a particular category. P(default = YES | balance) is one example.

The Logistic Model

In order to set a boundary to squeeze all predicted responses Y to be between 0 and 1, inclusively, for all values of X. In logistic regression, we use the logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

To fit the model, we use a method called *maximum likelihood*. Note that logistical function will always produce an S-shaped curve of this form.

With manipulation of the function above, we can get

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}. \tag{4.3}$$

The quantity $p(X)/[1-p(X)]$ is called the *odds*, and can take on any value between 0 and ∞. Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively. For example, on average

We then continue to take the logarithm of both sides to obtain the following:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X. \tag{4.4}$$

The left-hand side is called the *log-odds* or *logit*. We see that the logistic regression model (4.2) has a logit that is linear in X.

The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for B0 and B1 such that the predicted probability $\widehat{P}$(Xi) of default for each individual, corresponds as closely as possible to the individual's observed default status.

Commonly, we also use the sigmoid function as a logistic function to squeeze Y values into a [0,1] boundary, inclusively. The sigmoid function is as follows:

$$\phi(z) = \frac{1}{1+e^{-z}}$$

To make a prediction, we can fit the linear function to the maximum likelihood function to produce a probability:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576,$$

In the case of multiple logistic regression,

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Linear Discriminant Analysis

Reasons to use LDA (Or when to use):

1. When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. LDA does not suffer from this problem.

2. When the sample size is small and the distribution of the predictors X is approximately **normal**, LDA is more stable than logistic regression. LDA classifiers assume that observations from each class come from a normal distribution with a class-specific mean vector and a common variance σ^2.

3. LDA also assumes that all variance across K classes is identical.

4. LDA is great for classifying with more than 2 response classes.

Cons of using LDA:

1. Low performance in sensitivity: Since LDA tries to approximate the Bayes classifier, which has the lowest total error rate out of all classifiers, therefore, Bayes Classifier will yield the smallest possible total number of misclassified observations, irrespective of which class the errors come from.
2. Low performance in classifying data points produced by a non-linear function. (QDA can be a solution)

Using Bayes' Theorem for Classification

Suppose we wish to classify an observation to one of the K classes, where K >= 2. Let πk represents the overall or prior probability that a randomly chosen observation comes from the kth class. This is the probability that a given observation is associated with kth category of the response variable Y. $f_k(x) \equiv \Pr(X = x|Y = k)^1$ is the density function of X for an observation that comes from the kth class.

$f_k(x)$ is large if there is a high probability that the observation in the kth class has X ~= x and $f_k(x)$ is small if it is unlikely that the observation in the kth class has X ~= x.

Bayes' theorem states that,

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

To estimate the prior probability πk, we simply compute the fraction of the training observation that belongs to kth class. The function above is the posterior probability that an observation X = x belongs to the kth class. That is, it is the probability that the observation belongs to the kth class, given the predictor value for that observation.

Let's say, we assume $f_k(x)$ is normal or *Gaussian,* in a 1-dimensional setting, the normal density takes the forms:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right),$$

where μ_k and σ_k^2 are the mean and variance parameters for the kth class. For now, let us further assume that $\sigma_1^2 = \ldots = \sigma_K^2$: that is, there is a shared variance term across all K classes, which for simplicity we can denote by σ^2. Plugging (4.11) into (4.10), we find that

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}. \tag{4.12}$$

(Note that in (4.12), π_k denotes the prior probability that an observation belongs to the kth class, not to be confused with $\pi \approx 3.14159$, the mathematical constant.) The Bayes classifier involves assigning an observation

$X = x$ to the class for which (4.12) is largest. Taking the log of (4.12) and rearranging the terms, it is not hard to show that this is equivalent to assigning the observation to the class for which

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{4.13}$$

is largest. For instance, if $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier assigns an observation to class 1 if $2x\,(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, and to class 2 otherwise. In this case, the Bayes decision boundary corresponds to the point where

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}. \tag{4.14}$$

The LDA method approximates the Bayes classifier by plugging estimates for πk,μk, and σ^2. The following is used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

ties $\pi_1, \ldots, \pi_K$, which can be used directly. In the absence of any additional information, LDA estimates π_k using the proportion of the training observations that belong to the kth class. In other words,

$$\hat{\pi}_k = n_k/n. \tag{4.16}$$

The LDA classifier plugs the estimates given in (4.15) and (4.16) into (4.13), and assigns an observation $X = x$ to the class for which

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \tag{4.17}$$

is largest. The word *linear* in the classifier's name stems from the fact that the *discriminant functions* $\hat{\delta}_k(x)$ in (4.17) are linear functions of x (as opposed to a more complex function of x).

ROC curve is a popular graphic for simultaneously displaying the two types of eros for all possible thresholds. The ROC name is actually "Receiver Operating Characteristics". The overall performance of a classifier is given by the **area under the curve (AUC)**. An ideal ROC curve will hug the top left corner and so the larger the AUC the better the classifier.

Sensitivity (True positive rate):
The fraction of observations that are correctly identified, using a given threshold value.

1- Specificity (False positive rate):
The fraction of false observation that we classify incorrectly as a true observation

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

.

QDA (Quadratic Discriminant Analysis):

Assume that observations from each class are from Gaussian distribution (Normal), but it also assumes that each class has its own covariance matrix.

Why use QDA instead of LDA?
→ Bias-Variance Tradeoff

LDA estimates a covariance matrix of p(p+1)2 parameters, but QDA estimates a total of Kp(p+1)/2 parameters. In this way, LDA is a much less flexible classifier than QDA and so has significantly **low variance**, but it can also suffer from **high bias** if the assumption of K classes share a common covariance matrix is off.

LDA is better than QDA when dealing with a small training set and so reduce variance is the key.

QDA is better than LDA when the training set is large and the variance of the classifier is not a major concern.

Classification Model Comparision (***) :

LDA assumes that observations are from a normal distribution, so it will work better than logistic regression. Conversely, LR works better when it is not in the normal distribution.

KNN is a completely non-parametric approach: no assumptions are made about the shape of the decision boundary. So it will work better than logistic regression and LDA when the decision boundary is highly non-linear.

QDA assumes a quadratic decision boundary, so it can model with more situations. When the dataset is small, QDA is better than KNN.