6.3: Classification

Navona Calarco

The University of Toronto

Intro

Classification involves predicating a qualitative response by a assigning it to a category. The methods that are used to classify observations are called **classifiers** and most of them work by following two steps:

- Compute the probability that an observation belongs to a category.
- Classify the observation based on some probability threshold (i.e. if the probability that an observation belongs to some category is greater than 0.5 then assign the observation to that category)

Why not use linear regression?

Suppose we are trying to diagnose a patient with either a *stroke*, *drug overdose*, or *epileptic seizure* based on their symptoms. We can code this response as follows

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

At this point we could use linear regression to predict Y based on a set of predictors. However there are several problems with this coding

- Implies an ordering of the outcomes.
- The difference between epileptic seizure and stroke versus stroke and drug overdose is assumed to be the same.

A different ordering would give completely different results for the linear regression. There is no convenient way to code a qualitative response with more than two levels so that linear regression can be used.

Why not use linear regression?

The 0/1 coding for a binary qualitative response variable does not suffer the same problems. However the probabilities we obtain will be difficult to interpret

- negative probabilities
- probabilities above 1

So, linear regression only able to give crude estimates of the probabilities for a binary response.

In summary, we don't use linear regression for classification since:

- It does not work for a qualitative response variable with more than 2 classes.
- With 2 classes, the probability estimates are not meaningful.

Logistic Regression

Logistic regression models the probability that the response Y belongs to a particular category. Suppose we have a qualitative response Y that has two levels, coded as 0 and 1, and one predictor variable. We want to model

$$p(X) = \Pr(Y = 1 \mid X)$$

The logistic function keeps the probabilities between 0 and 1. For one predictor, the function is

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

As with linear regression, we are trying to fit β_0, β_1.

Estimating the regression coefficients

β_0 and β_1 are estimated using the training data using a method called **maximum likelihood**. This involves maximizing the likelihood function, but we will not cover the details of this function.

Odds

The **odds** compares the probability of a particular outcome to the probability of all the other outcomes.

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

- takes values between $(0, \infty)$
- odds close to $0 \Rightarrow$ very low probability of the outcome in question
- odds much greater than $0 \Rightarrow$ very high probability of the outcome in question.

Log Odds

The **log odds** (or logit) is obtained by taking the logarithm of the odds

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

- Increasing X by one unit changes the log odds by β_1.
- If β_1 is positive, increasing X is associated with increasing $p(X)$
- If β_1 is negative, increasing X is associated with decreasing $p(X)$

Making Predictions

Once the coefficients have been estimated predictions can be made for any value of the predictor. Logistic regression will give the probability of the outcome and the classification will be according to some threshold which depends on the problem or how conservative the predictions should be.

Multiple Predictors

Simple logistic regression can be extended to include multiple predictors

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

The log odds in this case becomes

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

As before, the maximum likelihood is used to estimate the coefficients.

Exercise: Logistic Regression

Open the Classification Exercises R Markdown file.

- Go over "Getting Started" together as a class.
- Go through the "Logistic Regression" as a class.
- 5 minutes for students to complete the questions from "Logistic Regression".
- Questions should be completed at home if time does not allow.

Multinomial logistic regression

We can extend to two-class logistic regression to accommodate K classes. We need to select one class to serve as the **baseline**, so we will choose the Kth class. Then the model becomes

$$\Pr(Y = K \mid X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1} x_1 + \cdots + \beta_{lp} x_p}} \quad \text{and,}$$

$$\Pr(Y = k \mid X = x) = \frac{e^{\beta_{k0} + \beta_{k1} x_1 + \cdots + \beta_{kp} x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1} x_1 + \cdots + \beta_{lp} x_p}} \quad \text{for } k = 1, \ldots, K-1$$

The interpretation of the coefficients is tied to the choice of the baseline.

Bayes Classifier

Suppose that we have a qualitative response variable Y with K distinct and ordered classes. The Bayes classifier use a less direct approach using Bayes' theorem to estimating the probabilities

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

- π_k is the **prior** probability that a random observation belongs to the kth class.
  - estimated as the fraction of the training observation that belong to the kth class.
- $f_k(X) \equiv \Pr(X \mid Y = k)$ is the **density function** of X for an observation from the kth class.

There are several methods we will discuss that attempt to approximate the Bayes classifier using different approaches for estimating $f_k(x)$.

Why Use Bayes Classifier?

- When there is a lot of separation between two classes logistic regression does not does not provide stable coefficient estimates.
- If the distribution of each of the predictors is approximately normal and the sample size is small, these approaches are more accurate.

The methods that attempt to estimate the Bayes classifier that we will cover are:

- Linear discriminant analysis,
- Quadratic discriminant analysis, and
- Naive Bayes.

Linear Discriminant Analysis

Suppose we only have one predictor, so $p = 1$. In order to estimate $f_k(x)$ we make the following assumptions:

- $f_k(x)$ is normal
- the variance is the same across all K classes.

Linear discriminant analysis (LDA) then approximates the Bayes classifier using the estimates:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \qquad (k\text{th mean})$$

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

$$\hat{\pi}_k = n_k/n \qquad (k\text{th prior probability}).$$

where n = number of training observations, and n_k = number of observations in the kth class.

Linear Discriminant Analysis

The LDA classifier uses the estimates for π_k, μ_k, and σ^2 to assign an observation $X = x$ to the class that has the largest

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

Linear Discriminant Analysis for $p > 1$

Now suppose we have $X = (X_1, \ldots, X_p)$ predictors. Assumptions: X is multivariate Gaussian (i.e. each predictor is normally distributed with some correlation between them)

- class-specific mean vectors
- common covariance matrix across classes.

We classify observations to the class for which $\hat{\delta}_k(x)$ is the largest.

Binary Classifiers

Binary classifiers, similarly to such as tests for diseases (positive versus negative), can make two types of errors:

- Incorrectly assign an individual as positive when they are negative (False positive).
- Incorrectly assign an individual as negative when they are positive (False negative).

Confusion Matrix

A confusion matrix helps to summarize the two types of errors of binary classifiers. They compare the LDA predictions to the true outcomes of the training observations. In the case of medical tests this looks like:

		True class		
		–	+	Total
Predicted class	–	True –	False –	N*
	+	False +	True +	P*
	Total	N	P	

- N and P are the number of actual negatives and positives respectively in the training data.
- N* and P* are the number of predicted negative and positives in the training data.

The red text is where the numbers are filled in.

Threshold

Recall that the Bayes classifier assigns observations to the class for with the posterior probability is the greatest. Since probabilities sum to 1, for a binary classifier this means that a test will come back positive if:

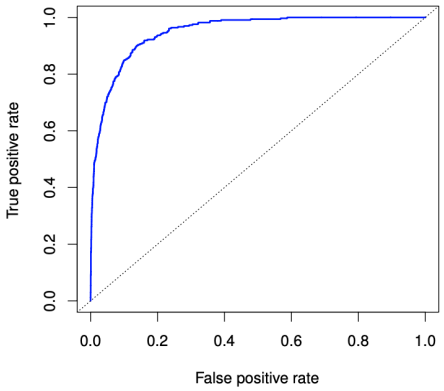
$$\Pr(\text{positive} \mid X = x) > 0.5$$

That is, the binary classifier uses a threshold of 50%. Depending on the classification problem, one may want to specify a different threshold level. For example:

$$\text{positive if } \Pr(\text{positive} \mid X = x) > 0.2$$

ROC

The ROC (receiver operator characteristics) curve is a method for visualising the errors previously discusses for all possible thresholds.



- Performance of a classifier is given by the area under the ROC curve, called the AUC.
- The larger the AUC the better.
- Ideal ROC curve is as close to the top left corner as possible.

Exercise: Linear Discriminant Analysis

Open the Classification Exercises R Markdown file.

- Go over the "Linear Discriminant Analysis" section together as a class.
- 5 minutes for students to complete the questions from "Linear Discriminant Analysis".
- Questions should be completed at home if time does not allow.

Quadratic Discriminant Analysis

The Quadratic discriminant analysis (QDA) classifier assumes that:

- observations are drawn from a class-specific Gaussian distribution
- each class has its own covariance matrix (unlike LDA)

The QDA uses estimates for the class-specific means (μ_k), covariance matrices ($\mathbf{\Sigma}_k$), and prior probability (μ_k) to assign an observation x to the class for which

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log \pi_k$$

is the largest. Unlike the LDA, the function for $\delta_k(x)$ is quadratic which gives the QDA it's name.

LDA vs QDA

When to use the LDA versus the QDA:

- LDA is better than QDA when there are few training observations since it requires fewer parameters to be estimated.
- QDA is best when there are many training observations or when the assumption of a common covariance matrix in the LDA is clearly wrong.

Exercise: Quadratic Discriminant Analysis

Open the Classification Exercises R Markdown file.

- Go over the "Quadratic Discriminant Analysis" section together as a class.
- 5 minutes for students to complete the questions from "Quadratic Discriminant Analysis".
- Questions should be completed at home if time does not allow.

Naive Bayes

The naive Bayes classifier assumes: *within each class, the p predictors are independent*. This allows us to disregard any association between the p predictors and gives the form of $f_k(x)$ as

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

where f_{kj} is the density function of the jth predictor for observations in the kth class.

To estimate f_{kj} using the training data there are several options:

- For quantitative X_j:
  - assume that for each class, the jth predictor is drawn from a normal distribution, or
  - estimate it as the fraction of the training observations in the kth class that belong to the same histogram bin as x_j.
- For quantitative X_j:
  - count the proportion of training observations for the jth predictor that belong to each class.

Exercise: Naive Bayes

Open the Classification Exercises R Markdown file.

- Go over the "Naive Bayes" section together as a class.
- 5 minutes for students to complete the questions from "Naive Bayes".
- Questions should be completed at home if time does not allow.

K-Nearest Neighbours

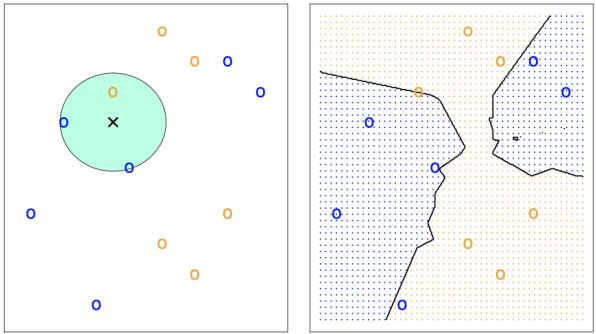
The K-nearest neighbors (KNN) classifier works very differently than any of the previous classification methods. For a test observation x_0, it identifies K training data points that are closest to x_0 (represented by $\mathcal{N}_0$) and estimates the conditional probability for class j as

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

where $I(y_i = j)$ if an **indicator variable** that equals 1 is $y_i = j$ and 0 otherwise. The KNN classifier classifies the test observation x_0 to the class for which the above probability is the largest.

K-Nearest Neighbours

These figures illustrate the KNN approach with $K = 3$. To the left we see the 3 closest points to x are 1 orange and 2 blue so this observation will be classified as blue. The right figure shows the decision boundaries where an observation will be classified as blue or orange.



Exercise: K-Nearest Neighbours

Open the Classification Exercises R Markdown file.

- Go over the "K-Nearest Neighbours" section together as a class.
- 5 minutes for students to complete the questions from "K-Nearest Neighbours".
- Questions should be completed at home if time does not allow.

How to choose the classification method

The choice of classification method depends on two things:

- the true distribution of the predictors in each of the K classes, and
- the number of training observations (n) compared to to the number of predictors (p).

References

Chapter 4 and section 2.2.3 of the ISLR2 book:

James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R, 2nd ed., Springer, 2021.