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.