Models for classification

Logistic regression

A quick recap

During the last lectures, we did...

- **Focus on models for regression**.

  Linear models
  - ⤳ Ordinary least squares
  - ⤳ Ridge, Lasso and their cousins
  - ⤳ Principal component regression
  - ⤳ Partial least squares

  Nonlinear models
  - ⤳ Neural networks for regression

Today's goal

Today, we going to do...

- **Models for classification**.

  Introduce the classification problem

  Introduce logistic regression models

Reading list

Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*, Springer (2014)

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshiran. *An Introduction to Statistical Learning with Applications in R*, Springer (2017)[1]

[1] The book is available for download at: `http://www-bcf.usc.edu/~gareth/ISL/`

Introduction to the classification problem

Classification problems occur often, perhaps even more so than regression problems.

Some examples include:

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
- On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Introduction to the classification problem

The regression model assumes that the response variable is quantitative.

In many situations, the response variable is qualitative (often referred as categorical).

- ⇝ **Classification**: the process for predicting qualitative responses
- ⇝ Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class.
- ⇝ On the other hand, often the **methods used for classification first predict the probability of each of the categories of a qualitative variable**, as the basis for making the classification.
- ⇝ In this sense they also behave like regression methods.

Introduction to the classification problem

Classification models usually generate two types of predictions.

- ▸ Like regression models, classification models produce a **continuous valued prediction**, which is usually in the form of a probability
  - ⇝ The predicted values of class membership for any individual sample are between 0 and 1 and sum to 1.
- ▸ In addition to a continuous prediction, classification models generate a predicted class, which comes in the form of a discrete category.

For most practical applications, a discrete category prediction is required in order to make a decision.

- ⇝ Automated spam filtering requires a definitive judgement for each e-mail.

Introduction to the classification problem

Although classification models produce both of these types of predictions, often the focus is on the discrete prediction rather than the continuous prediction.

However, the probability estimates for each class can be very useful for gauging the model's confidence about the predicted classification.

- ⇝ An e-mail message with a predicted probability of being spam of 0.51 would be classified the same as a message with a predicted probability of being spam of 0.99.
- ⇝ While both messages would be treated the same by the filter, we would have more confidence that the second message was, in fact, truly spam.

Introduction to the classification problem

Linear regression is not appropriate in the case of a qualitative response... **Why not?**

Suppose that we are trying to predict the **medical condition of a patient in the emergency room on the basis of her symptoms**.

Suppose that there are three possible diagnoses:

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

Least squares could be used to fit a linear regression model to predict Y on the basis of a set of predictors $X_1, \ldots, X_P$.

- ▸ This coding implies an **ordering on the outcomes**
- ▸ The difference between `stroke` and `drug overdose` is the same as the difference between `drug overdose` and `epileptic seizure`.
- ▸ In practice there is **no particular reason that this needs to be the case**.

Introduction to the classification problem

Linear regression is not appropriate in the case of a qualitative response... **Why not?**

Suppose that we are trying to predict the **medical condition of a patient in the emergency room on the basis of her symptoms**.

For instance, we could choose an equally reasonable coding:

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

This would imply a totally different relationship among the three conditions.

Each of these codings would **produce fundamentally different linear models** that would ultimately lead to different sets of predictions on test observations.

Introduction to the classification problem

Linear regression is not appropriate in the case of a qualitative response... **Why not?**

If the response variable's values did take on a natural ordering,

- ⤳ For example, as `mild`, `moderate`, and `severe`.

We felt the gap between mild and moderate was similar to the gap between moderate and severe, then a 1, 2, 3 coding would be reasonable.

Unfortunately, in general there is **no natural way to convert a qualitative response variable with more than two levels into a quantitative response** that is ready for linear regression.

Introduction to the classification problem

For a **binary (two level) qualitative response**, the situation is better.

For instance, perhaps there are only two possibilities for the patient's medical condition: `stroke` and `drug overdose`.

- ▸ We could then potentially make a dummy variable and code the response as:

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke} \\ 1 & \text{if } \texttt{drug overdose} \end{cases}$$

- ▸ We could then fit a linear regression to this binary response
- ▸ We could then predict drug overdose if $\hat{Y} > 0.5$ and stroke otherwise.

In the binary case it is not hard to show that even if we flip the above coding, linear regression will produce the same final predictions.

Introduction to the classification problem

For a **binary (two level) qualitative response**, the situation is better.

In this case, with a 0/1 coding, regression by least squares does make sense:

- ⤳ It can be shown that the $X\hat{\beta}$ obtained using linear regression is in fact an estimate of $P(\texttt{drug overdose}|X)$ in this special case.
  - ▸ However, if we use linear regression, some of our estimates might be outside the [0, 1] interval, making them hard to interpret as probabilities!
  - ▸ Even so, the predictions provide an ordering and can be interpreted as crude probability estimates.
  - ▸ Curiously, it turns out that the classifications that we get if we use linear regression to predict a binary response will be the same as for the linear discriminant analysis.

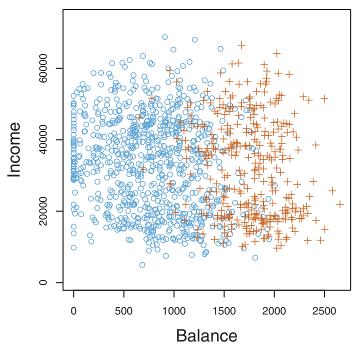
The dummy variable approach cannot be easily extended to accommodate qualitative responses with more than two levels.

For these reasons, **it is preferable to use a classification method that is truly suited for qualitative response values**.

Logistic regression

Consider the simulated `Default` data set[1].

- **Objective**: Predict whether an individual will default on his/her credit card payment, on the basis of annual income and monthly credit card balance.



We have available the following **data**:

- `annual income`
- `monthly credit card balance`

for a subset of 10000 individuals.

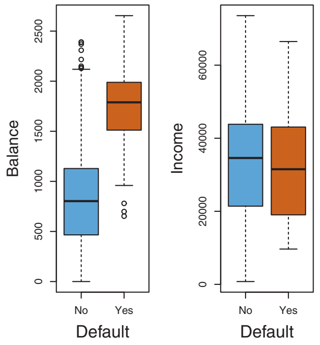
Individuals who defaulted tended to have higher credit card balances than those who did not.

[1] Available in R as: `library(ISLR); data(Default)`

Logistic regression

Consider the simulated `Default` data set.

- **Objective**: Predict whether an individual will default on his/her credit card payment, on the basis of annual income and monthly credit card balance.



- The first boxplot shows the distribution of `balance` split by the binary default variable.
- The second boxplot is a similar plot for `income`.

We want to build a model to predict `default` (Y) for any given value of `balance` (X_1) and `income` (X_2).

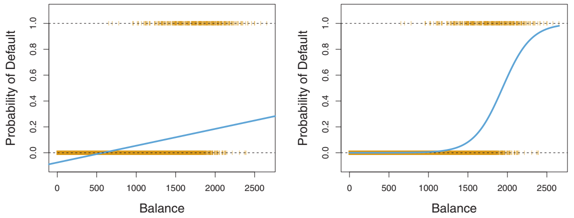
Since Y is not quantitative (`Yes` or `No`), the simple linear regression model is not appropriate.

Logistic regression

Consider the simulated `Default` data set.

- **Objective**: Predict whether an individual will default on his/her credit card payment, on the basis of annual income and monthly credit card balance.

Rather than modeling this response Y directly, **logistic regression models the probability that Y belongs to a particular category**.



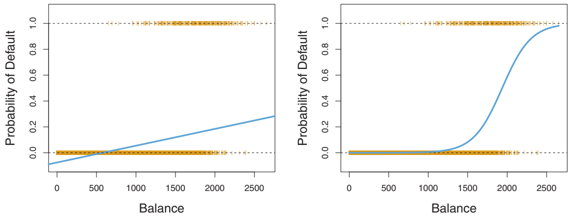
- **Left**: Estimated probability of default using linear regression. Some estimated probabilities are negative!
- **Right**: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.

Logistic regression

Consider the simulated `Default` data set.

- **Objective**: Predict whether an individual will default on his/her credit card payment, on the basis of annual income and monthly credit card balance.

Rather than modeling this response Y directly, **logistic regression models the probability that Y belongs to a particular category**.



- **Left**: Estimated probability of default using linear regression. Some estimated probabilities are negative!
- **Right**: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.

Logistic regression

Consider the simulated `Default` data set.

- **Objective**: Predict whether an individual will default on his/her credit card payment, on the basis of annual income and monthly credit card balance.

Rather than modeling this response Y directly, **logistic regression models the probability that Y belongs to a particular category**.

For example, the probability of `default` given balance can be written as

$$P(\texttt{default} = \texttt{Yes}|\texttt{balance})$$

- The values of $P(\texttt{default} = \texttt{Yes}|\texttt{balance})$ (abbr. $p(\texttt{balance})$) will range between 0 and 1.
- For any given value of `balance`, a prediction can be made for `default`.

  We might predict `default` = `Yes` for any individual for whom $p(\texttt{balance}) > 0.5$.

  Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for `default`, then they may choose to use a lower threshold, such as $p(\texttt{balance}) > 0.1$.

Logistic regression

How should we model the relationship between $p(X) = Pr(Y = 1|X)$ and X?

Logistic regression takes the form:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad 0 \le p(X) \le 1 \quad \text{for every value of } \beta_0, \beta_1 \text{ or } X$$

- To fit this model, we use a method called **maximum likelihood**.
- The logistic function will always produce an S-shaped curve of this form, and so regardless of the value of X, we will obtain a sensible prediction.
- The logistic model is better able to capture the range of probabilities than is the linear regression model in the left-hand plot.

Logistic regression

How should we model the relationship between $p(X) = Pr(Y = 1|X)$ and X?

After a bit of manipulations, we get:

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

- The quantity $p(X)/[1-p(X)]$ is called the **odds**.
- They 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.
- Odds are traditionally used instead of probabilities in horse-racing, since they relate more naturally to the correct betting strategy.

Logistic regression

How should we model the relationship between $p(X) = Pr(Y = 1|X)$ and X?

By taking the logarithm of both sides, we get:

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

- The left-hand side is called the **log-odds** or **logit**.
- The logistic regression model has a logit that is linear in X.
- Increasing X by one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}.
- However, because the relationship between $p(X)$ and X is not a straight line, β_1 does not correspond to the change in $p(X)$ associated with a one-unit increase in X.
- The amount that $p(X)$ changes due to a one-unit change in X will depend on the current value of X.
- But regardless of the value of X, if β_1 is positive then increasing X will be associated with increasing $p(X)$, and if β_1 is negative then increasing X will be associated with decreasing $p(X)$.

Logistic regression

Estimating the coefficients

To fit the model parameter, we could use the least squares as we did for regression. However, the more general method of **maximum likelihood** is preferred, since it has better statistical properties.

- **Maximum likelihood parameter estimation** is a technique that can be used when we are willing to make assumptions about the probability distribution of the data.
- Based on the theoretical probability distribution and the observed data, the likelihood function is a probability statement that can be made about a particular set of parameter values.
- If two sets of parameters values are being identified, the set with the larger likelihood would be deemed more consistent with the observed data.

Logistic regression

Estimating the coefficients

To fit the model parameter, we could use the least squares as we did for regression. However, the more general method of **maximum likelihood** is preferred, since it has better statistical properties.

We fit a logistic regression model using the maximum likelihood:

- We seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of `default` for each individual corresponds as closely as possible to the individual's observed `default` status.
- That is, we are trying to find $\hat{\beta_0}$ and $\hat{\beta_1}$ such that plugging these estimates into the model for $p(X)$ yields a number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not.
- This intuition can be formalized using the **likelihood function**:

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i=1} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

⇝ The estimates β_0 and β_1 are chosen to maximize this likelihood function.

Logistic regression

Estimating the coefficients

To fit the model parameter, we could use the least squares as we did for regression. However, the more general method of **maximum likelihood** is preferred, since it has better statistical properties.

Below are the coefficient estimates and related information that result from fitting a logistic regression model on the `Default` data in order to predict the probability of `default=Yes` using `balance`.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -10.65133    0.36116   -29.5   <2e-16 ***
Default$balance   0.00550    0.00022    24.9   <2e-16 ***
```

⇝ $\beta_1 = 0.0055$ indicates that an increase in `balance` is associated with an increase in the probability of default.

To be precise, a one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

Logistic regression

Making prediction

Once the coefficients have been estimated, it is a simple matter to compute the probability of `default` for any given credit card balance.

For example, we can predict the `default` for an individual with a `balance` of $1000:

$$\hat{p}(X) = \frac{e^{\hat{\beta_0} + \hat{\beta_1}X}}{1 + e^{\hat{\beta_0} + \hat{\beta_1}X}} = 0.00576 \quad \text{where } \beta_0 = -10.65133 \text{ and } \beta_1 = 0.0055$$

Multiple logistic regression

We can generalize the logistic regression to consider the problem of predicting binary responses using **multiple predictors**.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_P X_P$$

Equivalently:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_P X_P}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_P X_P}}$$

$\rightsquigarrow$ We use the maximum likelihood method to estimate $\beta_0, \beta_1, \ldots, \beta_P$.

Logistic regression for >2 response classes

We sometimes wish to classify a response variable that has more than two classes.

- For example, we had three categories of medical condition in the emergency room: stroke, drug overdose, epileptic seizure.
- In this setting, we wish to model both $P(Y = \texttt{stroke}|X)$ and $Pr(Y = \texttt{drug overdose}|X)$, with the remaining

$$Pr(Y = \texttt{epileptic seizure}|X) = 1 - Pr(Y = \texttt{stroke}|X) - Pr(Y = \texttt{drug overdose}|X)$$

- The two-class logistic regression models have multiple-class extensions, but in practice they tend not to be used all that often.

One of the reasons is that the **discriminant analysis method**, is popular for multiple-class classification.

However, multiple-class logistic regression is possible and software for it is available.