1700-351-451-w13a-f-Classification

Roger H. French, JiQi Liu

20 November, 2018

Contents

13.1.1.1 Reading, Homeworks, Projects, SemProjects

- Homework:
  - all done
- Readings:
  - ISLR 6 on Linear Model Selection
  - ISLR 5 on Resampling: Cross-validation and Bootstrap
- Projects: We will have four 2 week EDA projects
  - Project 3 due today
  - Project 4 given out thursday
  - Project 4 due Thursday December 6th
- 451 SemProjects:
  - Report Outs 3 In Week 15a, 15b
  - December 4th and December 6th

13.1.1.2 Syllabus

13.1.1.3 ISLR Chapter 4 Classification

Unsupervised learning example

- Classifying things into two cateogories
  - eye color {brown, blue, green}
  - email {spam, ham}.

Can we use linear regression for classification problems?

For binary classification, linear regression does a decent job.

This is called linear discriminant analysis

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1,**
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

Figure 1: DSCI351/451 Syllabus

13.1.1.3.1 Expected Values

Conditional mean of Y given $X = x$.

- $E(Y|X = x) = Pr(Y = 1|X = x)$

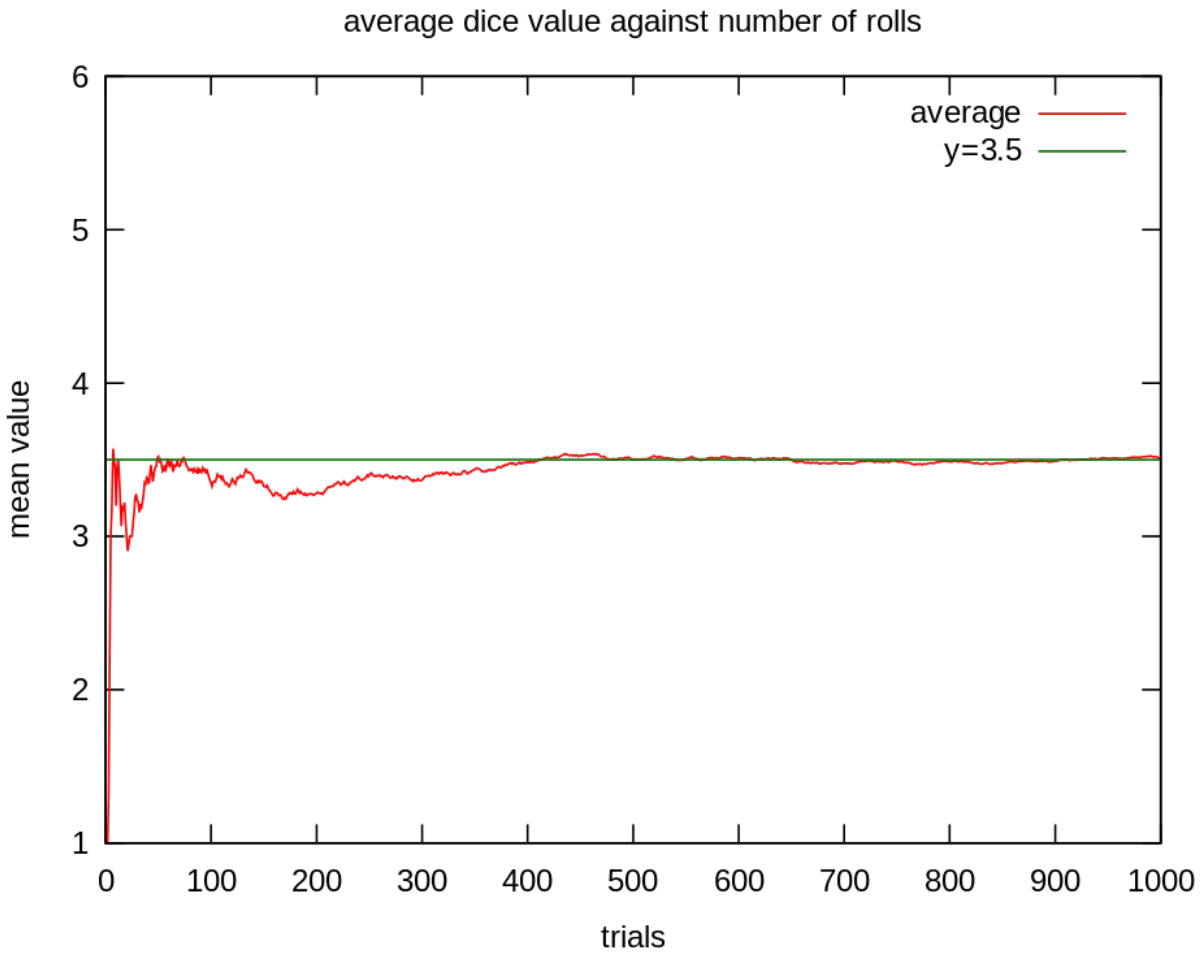
Expected Value

In probability theory, the expected value of a random variable, - intuitively,

- is the long-run average value of repetitions of the experiment it represents.

For example, the expected value in rolling a six-sided die

- is 3.5,
  - because the average of all the numbers
  - that come up in an extremely large number of rolls
  - is close to 3.5.



Caption: An illustration of the convergence of sequence averages of rolls of a die to the expected value of 3.5 as the number of rolls (trials) grows.

Less roughly, the law of large numbers states

- that the arithmetic mean of the values
- almost surely converges to the expected value
  - as the number of repetitions approaches infinity.

13.1.1.4 Logistic Regression

Linear regression can produce probabilities less than 0 or greater than 0

Instead Logistic Regression is more appropriate.

Cateogorical problems.

- Linear regression is not appropriate.
- Multi-class logistic regression is better.

Similar to encoding levels of cateogorical variables

- into a series of bits that each have only two levels.

Logistic Regression

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

Monotone transformation gives us a logarithmic ln function

- $log(p(X)/(1 - p(X))) = \beta_0 + \beta_1(X)$

[note in R, *log* is the natural log *ln*]

This is the "log odds" or the logit transformation of $p(X)$

Maximum Likelihood (Ronald Fisher)

Use Maximum Likelihood to estimate the parameters of the Logistic Regression model.

Using the glm package, as opposed to the lm package.

Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

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

($e \approx 2.71828$ is a mathematical constant [Euler's number.])
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

A bit of rearrangement gives

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

This monotone transformation is called the *log odds* or *logit* transformation of $p(X)$.

And if you have multiple predictors and 1 categorical response,

- you can do multiple logistic regression,
- as a simple extension.

13.1.1.4.1 Rule of Ten

A widely used rule of thumb, the "one in ten rule",

- states that logistic regression models
  - give stable values for the explanatory variables
- if based on a minimum of about 10 events per explanatory variable (EPV);
  - where event denotes the cases belonging to
  - the less frequent category in the dependent variable.

Thus a study designed to use k explanatory variables for an event

- (e.g. myocardial infarction)
  - expected to occur in a proportion p of participants in the study
  - will require a total of $10k/p$ participants.

However, there is considerable debate about the reliability of this rule,

- which is based on simulation studies
- and lacks a secure theoretical underpinning.

Logistic regression with several variables

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

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

Figure 2: Multiple Logistic Regression

13.1.1.5 Case Control Sampling

In epidemiology, you always want to use the cases of the disease,

- while then sampling from your control group.

The prevalence of disease in your study group (your sample)

- may be larger than in the population at large

So the probabilty of disease in you study sample

- (as opposed to the true population you pulled your sample from)
- might mean your logistic regression model is wrong.

Instead it turns out that only the β_0 term,

- the intercept will be wrong,
- the slopes will be right.

So you can correct the slope to represent the actual prevalence in your real population.

- $\tilde{\pi}$ is the apparent risk of disease in your study sample
- while π is the actual risk of disease in the larger population

13.1.1.5.1 Diminishing returns in unbalanced binary data

This means that you don't get alot of 1's (the disease)

- if you use unbiased sampling from the larger population.

So you can do "control to cases ratio"

- If you have a sparse cases,
- you can sample to account for this

13.1.1.6 Multi-class Logistic Regression

If you have multiple categorical responses

Case-control sampling and logistic regression

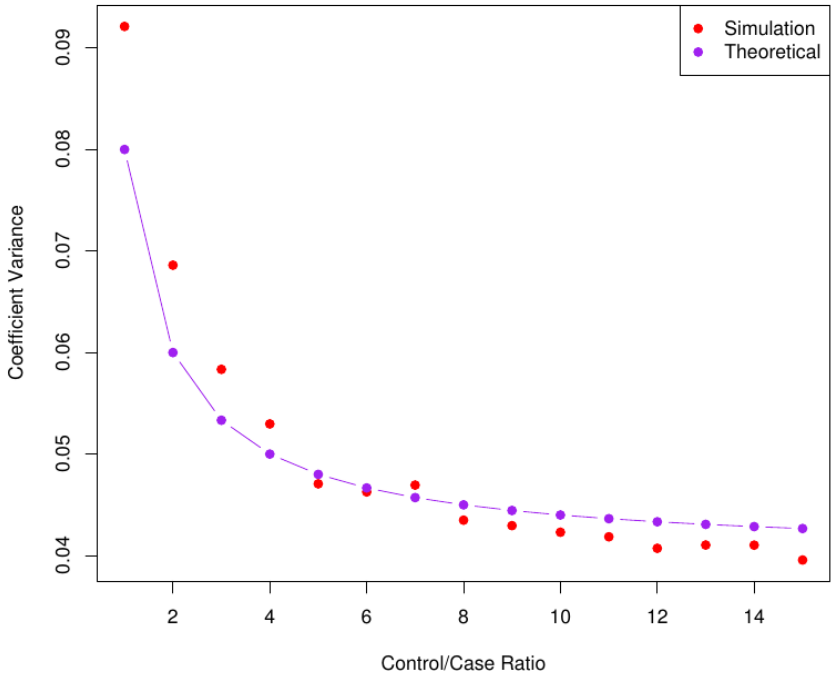
- In South African data, there are 160 cases, 302 controls — $\tilde{\pi} = 0.35$ are cases. Yet the prevalence of MI in this region is $\pi = 0.05$.
- With case-control samples, we can estimate the regression parameters β_j accurately (if our model is correct); the constant term β_0 is incorrect.
- We can correct the estimated intercept by a simple transformation

$$\hat{\beta}_0^* = \hat{\beta}_0 + \log \frac{\pi}{1-\pi} - \log \frac{\tilde{\pi}}{1-\tilde{\pi}}$$

- Often cases are rare and we take them all; up to five times that number of controls is sufficient. See next frame

Figure 3: Case Control Studies

Diminishing returns in unbalanced binary data



Sampling more controls than cases reduces the variance of the parameter estimates. But after a ratio of about 5 to 1 the variance reduction flattens out.

Figure 4: Case Control Ratio

- then you use multi-class logistic regregression.

And here you use glmnet package

Logistic regression with more than two classes

So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package `glmnet`) has the symmetric form

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

Here there is a linear function for *each* class.
(The *mathier* students will recognize that some cancellation is possible, and only $K-1$ linear functions are needed as in 2-class logistic regression.)

Multiclass logistic regression is also referred to as *multinomial regression*.

13.1.1.7 Discriminant Analysis

A different form of classification analysis.

Model the distribution of X in each of your classes Y.

Then use Bayes theorem to flip around and get $Pr(Y|X)$

You can get $Pr(Y = k|X = x)$ by knowing $Pr(X = x|Y = k)$ and adding in "priors"

- $Pr(Y = k)$ is called the marginal probability or prior probability of $Y = k$
- And you have the marginal probability of $Pr(X = x)$

Bayes theorem for classification

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probablistic modeling. Here we focus on a simple result, known as Bayes theorem:

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

One writes this slightly differently for discriminant analysis:

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

- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.

Logistic regression works well when you don't have strong predictors,

- i.e. for very complex systems with lots of predictors and interactions

Discriminant analysis is bettter used

- for cases where the classes are well separated and predictors are strong.

Why discriminant analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data.

- μ_k is the mean in class k
- σ_k^2 is the variance in class k

13.1.1.8 Notation Sidebar

We have equations

- For models The Y's X's,
- The values of predictors x_i and y_i's,
- The expected values of responses $\hat{y}$

We have the lm model notation.

We have

- μ is means,
- σ^2 are variances,
- π are probabilities