Lecture 10: Finishing Up Classification and Networks
Modeling Social Data, Spring 2019
Columbia University

Brigid Lynch

April 5, 2019

1 Classification as Regression?

$$y \subseteq \{0,1\} y \subseteq R \tag{1}$$

Recall from last time Logistic Regression:

$$\log \frac{p}{1-p} = w \cdot x \tag{2}$$

$$L = \prod_{i=i} p_i^{y_i} (1-p_i)^{1-y_i} \tag{3}$$

What if we just used linear regression to predict, something simple such as gender based on height?

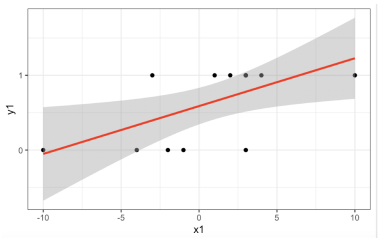


Figure 1: The points represent height and gender (1 being male 0 being female and x representing height), the line is the regression line

Since regression won't just be predicting 0 or 1, you're error rate may be incorrectly penalizing you, causing for an inaccurate line. Look at the graph, although line values above y=1 are technically you will still get an error of y-1. Logistic regression prevents this.

2 Reading Regression Tables

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7804879  0.0394345  19.792   <2e-16 ***
Sexmale     -0.5469036  0.0323428 -16.910   <2e-16 ***
Age         -0.0009206  0.0010730  -0.858    0.391
---
```

Figure 2: The output for a logistic regression table from the glm() function

The key here is to remember that these are logistic regression coefficients and therefore one must do a transformation on all of these values (p being the different estimates):

$$\frac{1}{1+\exp(-p} \tag{4}$$

Examples: Likelihood of a 20 year old male surviving: (recall that .009 means for each year your likelihood of survival decreases)

$$\frac{1}{1+\exp(-.74204-.546-.009*20)} \tag{5}$$

- What does the Estimate,Intercept value mean?
  - The Likelihood of a 0 year old female surviving
- This is why using predict() is useful, these coefficients are difficult to interpret

We also looked at graphically examining predictions with ggplot

- Emphasized bin size when looking at results
- ex: if you have less data points for 80 year olds than for 50 year olds then you should consider that when looking at your prediction results

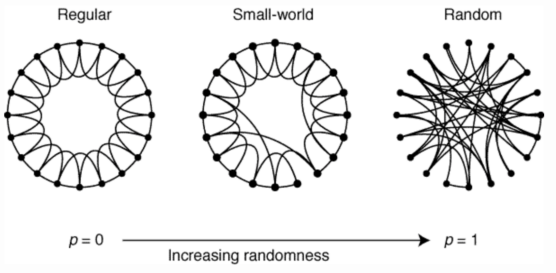
3 History of Networks

Main idea of Networks: things are not completely independent!

- 1930s: Relationships as Networks,socio-grams first network diagrams,big deal NYT headline, examined runaway girls through a school social network (Moreno)
- 1960s: Random graph theory:

$$p > \frac{(1+\epsilon)\ln_n}{n} \tag{6}$$

- 1970s: Clustering weak ties, theories motivated by intuition and smaller data sets
  - Granovetter, The Strength of Weak Ties:"The degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another" Relationship exists and matters!
  - Forbidden Triad: Mutual friends are connected in strong relationships
  - Also examined whether one finds jobs through strong/weak ties, without weak ties strong groups would not relate to one another.
  - Cumulative advantage: de Solla Price, looked at citations of other academic papers, citations are highly concentrated among popular papers and then a "long tail" of lesser known papers
- 90s: internet, got a lot more data! Less interview based
  - Small-world networks Watts and Strogatz, used probability and randomness to prove features of small-world networks



Watts & Strogatz (1998)

Figure 3: (From MSD Lecture 10 Slides)

- 2000s: Homophily, contagion and all that
  - Adamic and Glance, looked at political viewpoints in blogposts and connections based on political orientation
  - homophily: because you are similar you are friends
  - contagion: because you are friends you are similar
  - Attempts to figure out which is the cause

4 Types of Networks

Useful for many different types of data

- Social Networks (Facebook)
- Information Networks (web)
- Activity networks (email)
- Biological networks (protein interactions)
- Geographical networks (roads)

There are also many different levels of abstraction for representing networks.(Directed, weighted, metadata). They can also vary depending on how detailed you want your representation to be.

Which Network? Imagine a person's Facebook:

- ego network: person in middle all friends represented
- Maintained relationships: friends that actually interact
- One way communications
- Mutual communications

Important:

- What is the network?
- What are you counting?
- What are you encoding?
- *Simple is often better, don't get too crazy

5 Adjacency Matrix

Adjacency matrix

	0	1	2	3	4	5	6	7	8	9
0	0	1	0	0	0	0	1	0	1	0
1	1	0	0	0	1	0	1	0	0	1
2	0	0	0	0	1	0	1	0	0	0
3	0	0	0	0	1	1	0	0	1	0
4	0	1	1	1	0	1	0	0	0	1
5	0	0	0	1	1	0	0	0	0	0
6	1	1	1	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	1	1
8	1	0	0	1	0	0	0	1	0	0
9	0	1	0	0	1	0	0	1	0	0

Figure 4: (From MSD Lecture 10 Slides)

Time complexity if there is an edge: constant (index by row and column), find neighbors O(n), also good for linear algebra

6 Adjacency list

- Good for graph transveral checking neighbors and going through the graph.
- Time complexity to check if an edge exists: O(average degree))

7 Descriptive statistics

- Degree: How many connections does a node have?
- Path length: Shortest path between two nodes?
- Clustering: How many friends of friends are also friends?
- Components: How many disconnected parts does the network have?

This is where your text goes. If you're new to LaTeX, check out Overleaf[1], an online LaTeX environment where you can edit and render your documents. They also have a very useful getting started guide.

[1] http://overleaf.com