Lecture 10: Classification, Introduction to Networks
Modeling Social Data, Spring 2019
Columbia University

April 5, 2019

Notes from ap3772

1 Major Topics Covered

- Classification as a Regression Problem
- Introducing Logistic Regression
- Interpreting Logistic Regression
- Exploring Vowpal Wabbit
- Introduction to Networks

2 Classification as a Regression Problem

Why do we bother with Logistic Regression, when a simple regression model could be used for our classification problem? Linear regression returns a real number, which we can then convert to a class.

We can have a dependent variable 'y' and a set of independent features

$$x \in R^d \tag{1}$$

As seen in Linear Regression, we fit a model :

$$\hat{y} = w.x \tag{2}$$

where w is the vector of weights.

Analyzing the performance of this model using the previously introduced Least Square approach, we begin to see a problem. Consider fitting a regression line to this plot. Our 'y' will definitely be a real number, but we want it to be a class.

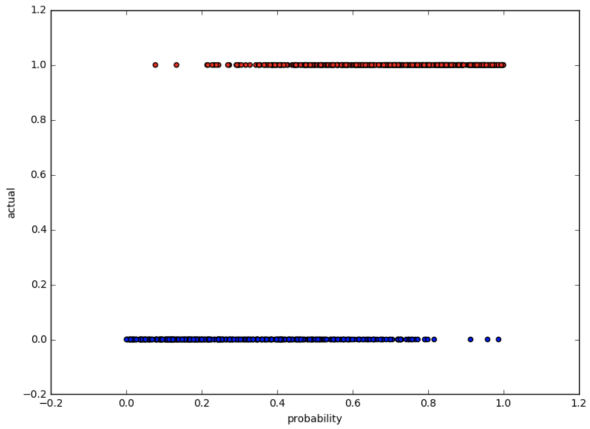


Figure 1:

Using Least Squares method to analyze our model performance on fitting a straight line, as always:

$$LSE = 1/n \sum_i (\hat{y_i} - y_i)^2 \tag{3}$$

Error rates aren't representative of actual errors, since the points that were classified correctly will also give a non-zero error.

But fret not, we can fix this with a hack - a very simple case of clipped or piecewise linear regression. This way, all predicted values above 1 will be set to 1, and all predicted values less than 0 will be set to 0.

We'll see that logistic regression is a more eloquent way of handling this issue.

3 Introducing Logistic Regression

When we've restricted our response variable to $y \in 0, 1$, logistic regression provides a solution by fitting our model to the log odds rather than the class of y.

The resulting fit equation is given by :

$$ln(\frac{p}{1-p}) = w.x \tag{4}$$

Loss is calculated as :

$$L = \prod_{i=i} p_i^{y_i}(1-p_i)^{1-y_i} \tag{5}$$

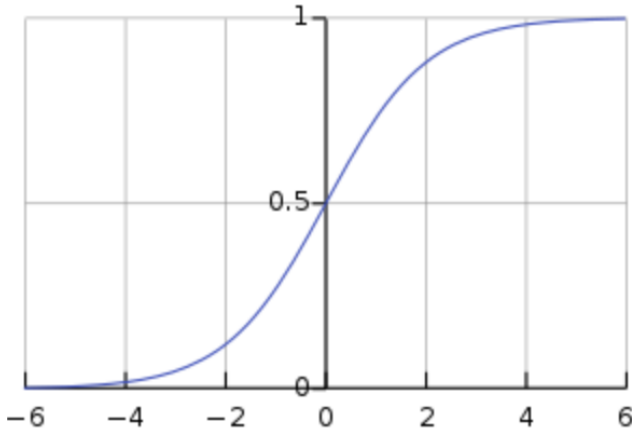


Figure 2:

The resulting sigmoid eliminates the loss problem that we saw in the previous section.

4 Interpreting Logistic Regression

We considered the Titanic dataset to predict whether a person survived or drowned, based on Gender first.

term	estimate	std.error	statistic	p.value
(Intercept)	0.7420382	0.0230658	32.17052	0
Sexmale	-0.5531301	0.0286628	-19.29782	0

Figure 3:

Let us now interpret this table. The 'sexmale' estimate tells us the change in log odds of male survival per unit increase. The 'intercept' refers to the odds of a zero year old male surviving. We use the following equation to interpret the values returned by summary :

$$p(y = 1|x, w) = 1/(1 + e(-wx)) \tag{6}$$

So, for instance, we want to compute the survival odds of a person given a host of variables (with respective coefficients obtained from the summary) about them, here's the formula :

$$survival - prob = 1/(1 + exp(-(coef_1 * val_1 + coef_2 * val_2 + coef_3 * val_3 + ...))) \tag{7}$$

where $coef_i$ was obtained from the summary, and val_i are values that the feature takes.

*The discussion digressed to how the error bar can be obtained for regression model - solution is to use different samples, and generate fits for them. Get a 95% confidence interval and use convex hull.

5 Vowpal Wabbit

Here's a few takeaways from our tutorial on Vowpal Wabbits :

- Has some advantages over traditional packages - such as scalability and feature pairing.
- Uses progressive validation instead of cross validation - loss reduces over iterations.
- Accepts input in a very flexible and convenient format (esp for our news example) - just enter the list of words occuring in each row.
- Learning Algorithm is very fast as demonstrated in class.

6 Introduction to Networks

An assumption so far has been that the individual observations are Independent. With Networks we assume a relational dependence, i.e the observations are interconnected at some abstract level.

Let us go through the history of Networks :

1. An early paper analyzing social networks, by Granovetter titled "Strength of Weak Ties" argues that the "degree of overlap of friendship networks" (i.e mutual friends) is proportional to the strength of their friendship.

   We also saw the distribution of social networks can be defined as clusters of strong ties, that are connected with weak ties. The paper stresses on the power these weak ties between groups can wield.

   This also ties to a pattern often observed in academia, and in general - of "cumulative advantage" , i.e the rich getting richer. An example we observed was that of the number of citations per paper. A tiny minority of papers get cited a lot, whereas the vast majority have very few citations.

2. Erdos and Renyi worked on Random Graph Theory - probabilistic models. Each edge has a probability of existing - how high does the probability have to be for a set of users to be connected.

3. Small World Networks were introduced in the 90s, after Watts and Stogartz observed that many real world networks had small average shortest paths between any two random nodes (as observed in the ER pure Random graphs) but a higher clustering coefficient.

   This high clustering coefficient results in a bunch of high degree nodes called "hubs" that result in shorter average path lengths that are observed in real life scenarios (eg the "6 degrees of separation").

4. In the modern times, we have a lot of social network data at our disposal. We observed a visualization made using blog data that showed the polarization between Republican and Democrat supporters.

We discussed the various types of networks that are in use today - for instance "Social Media" - sites like Facebook where there is an explicit definition of the relationship between 2 users as 'friends'/'connections'. Then there's Geographical networks like Google Maps.

Representing networks can be done with varying degrees of Abstraction - for instance undirected networks indicate only a link between two nodes whereas directed networks give a direction/causation info.

We discussed different ways of visualizing networks - for instance in Facebook networks we could perhaps restrict the visualization to only the active connections, or perhaps to reciprocating connections to gather better insight. We saw an example of such 'encoding', where the emails sent within an organization were used to deduce reporting relationships (darker lines) and email communication (other lines) between the various senders and receivers.

There are a few ways of Representing Networks :

1. Edge List :
   - Simply stores a list of the connected edges in the graph/network.
   - Simple for storage

- Complex computations, finding all points adjacent to a node requires O(e) where 'e' is the number of edges.

2. Adjacency Matrix :
   - A matrix that stores a 0 or 1 indication presence of edge between all edge (n1,n2) combinations.
   - Higher space complexity
   $$O(n^2) \tag{8}$$
   - Lower computational complexity - for instance finding all neighbours of given node takes
   $$O(n) \tag{9}$$
   - Finding whether an edge exists is super fast -
   $$O(1) \tag{10}$$
3. Adjacency List :
   - An array of all nodes, where each element points to a linked list of adjacent nodes.
   - Not the best structure for checking if an edge exists, will take
   $$O(avg - degree) \tag{11}$$
   - However good for getting all the neighbours, hence traversal.

Describing our Network can use many parameters, for instance degree distributions - say there are lots of 3-4 way intersections. We've discussed some others, like the distance between two nodes (in history of networks), groups (tightly knit clusters).

Notes from bml2133

1 Introduction

In this lecture we'll finish going over classification and begin our discussion on networks. Namely we will look at more subtle aspects of classification such as the difference between logistic regression and linear regression and why one may be preferable. Later we'll do some coding which is where the R code will come in handy, looking at how we can use ggplot to think critically about what our classification models are saying.

2 Classification as Regression?

$$y \subseteq \{0,1\} y \subseteq R \tag{12}$$

Recall from last time Logistic Regression:

$$\log \frac{p}{1-p} = w \cdot x \tag{13}$$

$$L = \prod_{i=i} p_i^{y_i} (1-p_i)^{1-y_i} \tag{14}$$

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

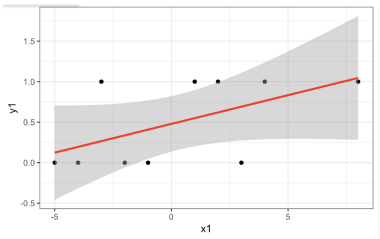


Figure 4: The points represent height and gender (1 being male 0 being female and x representing height), the line is the regression line

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.

3 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 5: 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):

$$\frac{1}{1+\exp(-p} \tag{15}$$

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

$$\frac{1}{1+\exp(-.74204-.546-.009*20)} \tag{16}$$

- 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

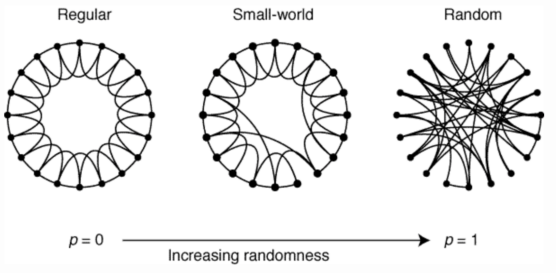
4 History of Networks

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:

$$p > \frac{(1+\epsilon)\ln_n}{n} \tag{17}$$

- 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 6: (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

5 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

6 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 7: (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

7 Adjacency list

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

8 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?

Notes from gia2105

1 Outline

- Classification
- Logistic Regression, with examples in R
  Vowpal Rabbit Example
- Networks
  History
  Structures and Applications

2 Classification

Question: why not solve Classification as a regression problem?

More precisely, say you have a set of response y variables that take values of either 0 or 1, and a set of x values that take any values on the real line.

$$x \in \mathbf{R}$$
$$y \in \{0, 1\}$$

Why not model a linear regression solution using OLS? That is, fit a model like:

$$\hat{y} = w \cdot x$$
$$\text{where } w = \frac{\mathbf{x} \cdot \mathbf{y}}{\mathbf{x} \cdot \mathbf{x}}$$
$$\text{, minmizing the loss function:}$$
$$L = \frac{1}{n} \sum_i \left(y_i - \hat{y_i}\right)^2$$

Look at the example plot (on the next page)

We can see that the point at the top right is giving a non zero loss, even though the line is above 1, and so it would predict the right label, but would still give some loss.

One possible fix to this, that is used in practice, is to set up a piece-wise linear model that is 1 if the prediction is greater than 1, and zero if the prediction in less that zero.

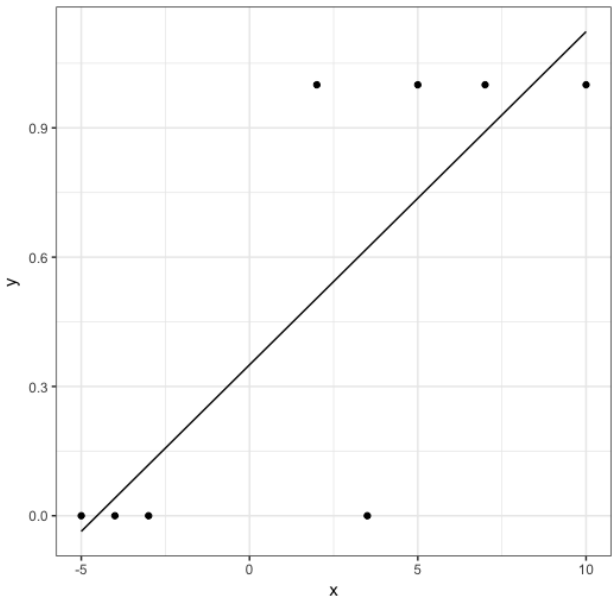
$$\hat{y} = 1 \text{ if } w \cdot x > 1$$
$$0 \text{ if } w \cdot x < 0$$
$$w \cdot x \text{ otherwise}$$

Sort of like a piece-wise-linear, jagged version of logistic Regression

3 Logistic Regression

Review: the model make a linear fit to the log odds:

$$\ln(\frac{p}{1-p}) = w \cdot x$$
$$\text{This minimizes the loss function:}$$
$$L = \prod_i p_i^{y_i} \cdot (1 - p_i)^{1 - y_i}$$



We then went through an example analyzing the chances of survival for passengers on the Titanic. (See class github page for the R file). The main takeaways are:

- Coefficients returned from logistic regression models can be difficult to interpret, as they relate to log-odds, although it is possible to make the tranformations to regular probabilities
- Instead, it can be much easier to analyze predictions against observations
- As more variables are added into the model, the results are harder to interpret

Next, we looked at machine-learning based models:

- goals are less about model interpretability
- prioritizes ability fro large-scale data processiong
- solely for prediction, quickly

One such machine-learning based approach we looked at was Vowpal Rabbit, which has the following advantages:

- Input formal is very easy, fast (Uses Stochastic Gradient Descent)
- Speed is fast
- It is scalale, can deal with a large number of features
- Feature pairing, progressive validation
- can do logistic regression, but can also do much more!

4 Networks

In our previous models, there has been an assumption of independence between datapoints. With networks, this assumption does not hold, and instead the datapoints are relational. Networks have both mathematical and social applications. First, a brief look at the history of Networks:

- first looked at Granoveter paper, modeling two individuals' number of mutual friends as a function of the strength of their frienship. (stronger friendship = more mutual friends)
- analyzing the success of academic papers. We could see the "long tail" that we've discussed in previous classes... most papers don't get cited much, if at all.
- Next, we looked at Watts and Strogat's Small-World Networks. A regular network sees no random connections between nodes, and a random network has only random connections between nodes. A small-world network is somewhere in between, with some randonm connections and some nonrandom connections. This model has the ability to capture long-ranging ties, i.e. the "6 degrees of separation" phenomenon.
- we saw another plot illustration the connections that blogs have. The plot observed a disconnect between different sides of the political spectrum.

Next, we Looked at types of Networks. There can be many applications for networks, such as modeling social, info, activity, biological or geographical networks. There is no single way to represent networks, but there are usually connections illustrated between nodes. Features of networks can also be directed, weighted...

There are many ways to store network data. Here are a few examples (see the lecture notes for graph visualizations):

1. Edge List
    - Is a list made up of every edge, with each entry containing the two nodes the edges connects
    - simple storage
    - bad for computation, computational complexity is O(number of edges)
2. Adjacency Matrix
    - a square matrix, where each row and column index is a node
    - grid of 1's and 0's. 1 represents a connection between nodes, 0 means no connection
    - this is quick to check edges
    - good for linear algebra
    - matrix is often sparse
3. Adjacency List
    - for each node, there is a list of what other node is connected to
    - this is good for graph traversal
    - if directed graph, needs 2 lists per node

There are many features that can be used to desctibe networks:

- degree - number of connections a node has
- path lenght - shortest past between nodes
- clustering
- components - how many disconnected parts does the network have?