DATA605

Week11 Lecture

Week 10 Review: Issues with PageRank

The assignment from Week 10 asked you to calculate the PageRank of nodes in a graph in three ways:

1. Power Iteration
2. Eigenvector of modified transition matrix
3. PageRank algorithm as provided by a graph package

We had some issues with part 1 and part 2.

Week 10 Review: Issues with PageRank

First issue seems to be that the Power Iteration method doesn’t work with the matrix as described in the notes. The solution was to modify the matrix to make sure node 2 has values other than 0, to make sure the rows sum to 1.

The second issue is the order of multiplication in the Power Iteration method.

The third issue was in part 2, where finding the eigenvector does not get the same value as part 1. The solution here is to think about the order of multiplication above.

Week 10 Review: Issues with PageRank

Review code in page_rank.Rmd

What is a Hypothesis? What is Hypothesis Testing?

A hypothesis is a formal statement that presents an expected relationship between a dependent variable and a set of independent variables. Might be formed by guessing, or having a hunch as to a relationship.

Hypothesis testing is a collection of formal methods for how we can determine how confident we should be that a given hypothesis holds, based on data you have collected.

For a link outlining other kinds of hypothesis testing that can be done, check here: http://facweb.cs.depaul.edu/sjost/csc423/documents/test-descriptions.htm

https://www.slideshare.net/17somya/hypothesis-16733191

What should be in a hypothesis?

Formally, a hypothesis has these components:

- Variables
- A population these variables come from
- A stated relationship between those variables

Let us look at some examples and see how these components show up

Examples of Hypothesis Testing

- Out of 100 people, 80 are given a vaccine for a particular disease and 20 are given a placebo. Of the 80 who were administered the drug, how do we know that those who did not suffer was due to their being vaccinated? How confident are we in this conclusion?
  - Our hypothesis is “the vaccine is effective at decreasing the disease prevalence”
- 80 / 100 customers are sent a coupon in the mail. 20/100 are not. Of those who purchased something with the coupon, how many did so because of the coupon? How confident are we in our conclusion?
  - Our hypothesis is “the coupon increases sales among our customers”

What are the variables, population and relationship for each hypothesis?

Examples of Hypothesis Testing

- Out of 100 people, 80 are given a vaccine for a particular disease and 20 are given a placebo. Of the 80 who were administered the drug, how do we know that those who did not suffer was due to their being vaccinated? How confident are we in this conclusion?
  - Our hypothesis is “the vaccine is effective at decreasing the disease prevalence”
    - **Variables:** administered the vaccine, prevalence of disease
    - **Population:** general populace that can get said disease
    - **Relationship:** administered the vaccine should result in **lower** prevalence of disease

Examples of Hypothesis Testing

- 80 / 100 customers are sent a coupon in the mail. 20/100 are not. Of those who purchased something with the coupon, how many did so because of the coupon? How confident are we in our conclusion?
  - Our hypothesis is “the coupon increase sales among our customers”
    - **Variables:** sent coupon, whether or not a customer made a purchase
    - **Population:** customers of our business
    - **Relationship:** coupons **increase** sales

Null Hypothesis Significance Testing

One **specific** form of hypothesis testing is null hypothesis significance testing. In the above examples the Null Hypothesis would be:

- The vaccine has no effect
- The coupon has no effect

A null hypothesis assumes that there is **no relation** between variables; and the alternative / research hypothesis is that there **is a relation** between variables. Our job is to either **reject** or **fail to reject** the null hypothesis.

Factors that govern whether the Null Hypothesis is rejected or not:

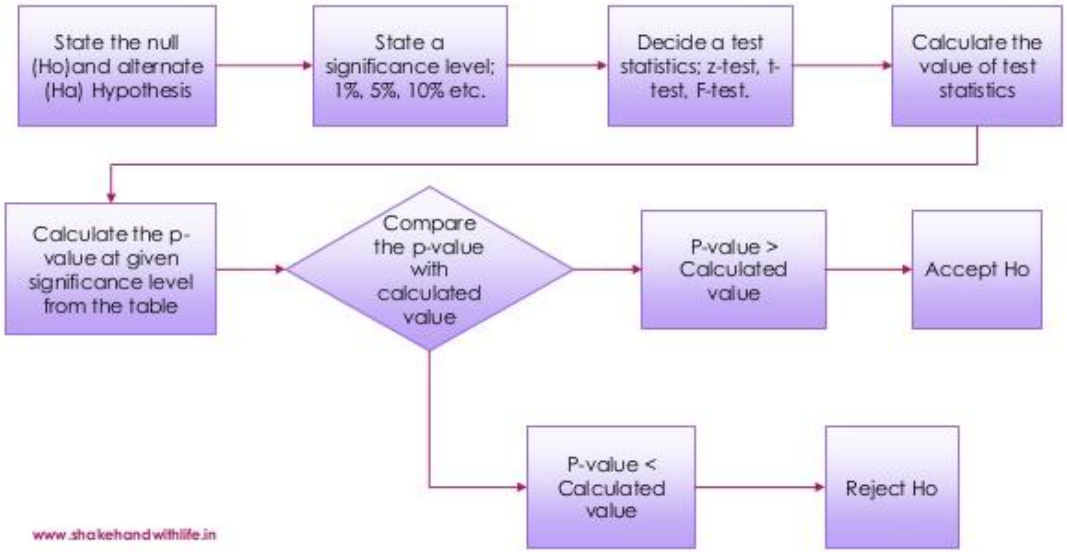
- The size of the effect is so large we are fully confident in rejecting it
- The sample size: The larger the sample size, the closer the estimator is to the true value of the metric.

Null Hypothesis Significance Testing

The general process:

1. Create a null (H_0) and alternative hypothesis (H_a)
2. Assume the null hypothesis is true
3. Determine a level of confidence you need in order to overturn H_0
4. Calculate the likelihood of seeing a particular data sample, given that the hypothesis is true. Mathematically this is:
   P(data | hypothesis)
   a. In this step, make sure to account for ‘direction’. If your hypothesis is that the sample mean is greater / less than the population mean, then use a one-sided test; otherwise use a two sides test
5. Reject H_0 if P(data | hypothesis) < significance level; otherwise fail to reject

Procedure for Hypothesis Testing



https://image.slidesharecdn.com/hypothesistestingz-testt-test-150831182208-lva1-app6892/95/hypothesis-testing-z-test-ttest-ftest-13-638.jpg?cb=1441045393

Types of Errors

1) **ERROR TYPE I: False Positive**
   This occurs when we incorrectly reject a true null hypothesis**
   (think about an alarm going off when there is no fire)

2) **ERROR TYPE II: False Negative**
   Failing to correctly reject a false null hypothesis
   (think about no alarm going off when there is a raging fire ongoing)

	H_0 True	H_0 False
Reject H_0	Type I Error (False Positive)	Correct (True Positive)
Accept H_0	Correct (True Negative)	Type II Error (False Negative)

Significance

Alpha (α) is the probability of committing a Type I error, which is rejecting the null and accepting the alternative when you should not. This is going to be the value you compare to in the hypothesis test workflow to determine if we reject (or fail to reject) the null hypothesis.

There is another number we use to describe the probability of committing a Type II error, but this is not covered here.

Discussion Topic

Why is it important to have significance level at step 2?

Discussion Topic

Why is it important to have significance level at step 2?

Significance should be chosen before hand based on experimental design, the domain you are working in, business considerations, etc.

Discussion board: I will post an article about ‘p hacking’ and would love to see everyone discuss it. Optional in the sense it is not graded, but this is an important topic for data scientists to be familiar with.

P-Values and Rejection the Null Hypothesis

The *p*-value is defined as the probability, under the assumption of the null hypothesis , of obtaining a result equal to or more extreme than what was actually observed. We need to be careful with the phrase “more extreme”, as this could mean different things in different contexts:

- If your hypothesis is testing for an increase, then your p-value = $Pr(X >= x \mid H_0)$ which is a one-sided (right-tail) test
- If your hypothesis is testing for a decrease, then your p-value = $Pr(X <= x \mid H_0)$ which is a one-sided (left-tail) test
- If your hypothesis is testing for any change, then your p-value = $2 * Pr(X >= x \mid H_0)$ which is a two-sided test

The lower the p-value, the better: this means we are less likely to have seen this data given the null hypothesis. But when do officially say the null hypothesis is rejected? When the p-value falls below your α ! Otherwise, we fail to reject.

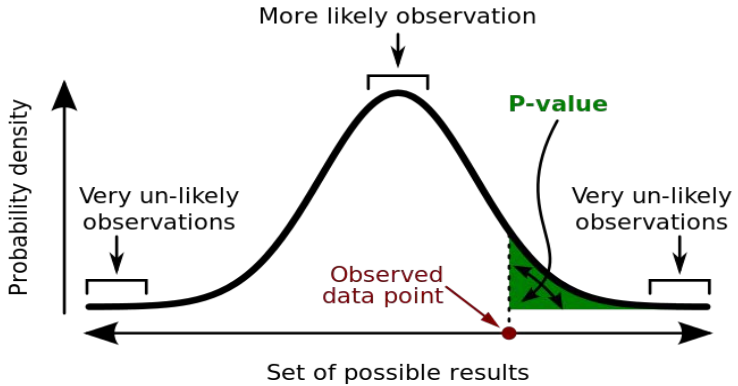
Important Note

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a “score” is committing an egregious logical error: **the transposed conditional fallacy.**



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

https://en.wikipedia.org/wiki/P-value#Definition_and_interpretation

Hypothesis Testing Example in R

Review code from hypothesis_testing.Rmd

Inspired by these links but please note, I made a correction in the code from these pages (in that they code a one sided test but claim they are doing a two sided test)

1. http://stats.seandolinar.com/calculating-z-scores-with-r/
2. http://stats.seandolinar.com/one-sample-t-test-with-r-code/

Hypothesis Testing Example: Regression

Regression analysis allows us to shed some light on the relationship between two variables, the dependant variable (or outcome of interest) in this case ‘y’, and the dependant variable(s) in this case ‘x’. If the relation between these variables is linear, we can model the relationship by:

$$y = b_0 + b_1 x + e$$

You can extend this formula for more than one independent variable. The terms are:

b_0 is the ‘y intercept’
b_1 is the slope of the regression line, or the ‘size of the effect’ of x on y
e is an error term

Regression Assumptions (Optional, Not in Notes)

We make certain assumptions when doing linear regression. While not a topic in this class, you will be reviewing this in your next class, Mathematical Modeling. I present it here so you can be familiar with the topic, but it is not required for this class:

- https://www.coursera.org/learn/regression-modeling-practice/lecture/ZUFlh/lesson-4-linear-regression-assumptions
- http://r-statistics.co/Assumptions-of-Linear-Regression.html

Hypothesis Testing Example: Regression

We can use Ordinary Least Squares to estimate the best b_0 and b_1 model parameters that reduce the error to the minimum. After fitting, we can evaluate the values to see if they are significant.

For example, we can evaluate the null hypothesis that b_1 is 0 (i.e. there is no ‘effect’) and reject / fail to reject it based on the evidence from your data.

Hypothesis Testing on Linear Regression Example in R

Since the homework assignment asks about this directly, I will provide links to some great resources on lm(), summary() and other R functions needed to build and evaluate linear regression models. These links go into a lot more detail than needed for this class, and introduces a lot of topics we go into in the next course, Mathematical Modeling (like how to access the quality of a regression):

- http://www.learnbymarketing.com/tutorials/linear-regression-in-r/
- http://blog.yhat.com/posts/r-lm-summary.html

Example From Industry: A / B Testing

A/B testing, also known as two-sample hypothesis testing, is big in the web industry. In this kind of hypothesis testing, you have two samples and you are trying to see if there is a significant difference in some metric of those samples. An example, many ecommerce sites want to know if teaking their website will create more conversions (visitors who decide to register). The company will setup two samples of visitors:

- Some visitors see the normal website (the control, A)
- Some visitors see the updated website (the experimental group, B)

The company will measure the metric in question and will test a hypothesis like “Group B will have higher conversion rate than Group A” for a chosen significance level. Given enough interactions in these groups, the website can determine if the tweaks ended up helping the site in a significant way.

Resources

- Great overview of hypothesis testing: https://www.analyticsvidhya.com/blog/2015/09/hypothesis-testing-explained/