05/22/2019

IMDB INFERENTIAL STATISTICS

1. WHY DO INFERENTIAL STATISTICS?

After exploring and visualizing the IMDB dataset, I noticed that there were some polarizing words, words rich in sentiment, being more common in either the positive reviews or negative reviews. For example, the word "love" appeared in 1,211 more reviews within the positive reviews than the negative reviews. I then wanted to test if these word differences actually showed a difference between positive and negative reviews, or if they were just a product of coincidence. I decided to test the words "hate" and "love"; below you can see some information of both.

i. "love"
    1. # of Positive Reviews: 2,825 out of 12,500
    2. # of Negative Reviews: 1,614 out of 12,500
        a. Total: 4,439 out of 25,000 reviews
        b. Difference: 1,211 reviews

ii. "hate"
    1. # of Positive Reviews: 301 out of 12,500
    2. # of Negative Reviews: 395 out of 12,500
        a. Total: 696 out of 25,000 reviews
        b. Difference: 94 reviews

2. HOW TO TEST?

The hypothesis being tested says that there is no difference between positive and negative reviews, and that they both should have equal number of reviews containing the word being tested. In order to check this hypothesis, we are going to simulate 100,000 datasets where each review in the positive and negative should have the same probability of having the tested word in the review. Each word being tested already has two lists detailing whether a given review has the word in it or not. For every review, there exists a value of either "True", if the word is in the review, or "False", if the word is not in the review. For example, the word "love" has a list with 12,500 values containing 2,825 Trues and 9,675 Falses for the positive reviews and another list for negative reviews with corresponding True and False values. We will then combine both lists into a bigger list of 25,000 and randomly shuffle those values, or in other words, permutate them. Then we will make two new lists of 12,500 values each from the permutated list. After we have obtained our two new permuted lists, we will compute the difference

between both lists of the number of reviews in each list having the word "love" in them by counting the number of Trues. The difference between both lists will be stored, and we will repeat the process 99,999 times until we have 100,000 differences. At the end of our simulations, we will compare how many of our 100,000 simulations have a difference greater than or equal to the observed difference; in the case of the word "love" the observed difference is 1,211 reviews. Below will be the results of running such test on the words "love" and "hate".

i. "love"
    1. Probability of difference of at least 1,211 reviews: 0.000%
        a. Not a single simulation had a difference of 1,211 or greater.
    2. 99%+ of simulations have a difference of 155 reviews or less
ii. "hate"
    1. Probability of difference of at least 94 reviews: 0.016%
        a. Only 16 of the 100,000 simulations had a difference of 94 or greater.
    2. 99%+ of simulations have a difference of 68 or less.

As it can be seen from the results, both words do not seem to be due to chance, and we can reject the hypothesis that positive and negative reviews have an equal likelihood of containing either word. For the word "love" we can conclude with almost 100% confidence that the difference of 1,211 reviews between the positive and negative datasets is not due to chance, since zero out of the 100,000 simulations done had a difference of 1,211 or greater; because more simulations could always be done, we can never say we are 100% confident that the difference of reviews for the word "love" will never be 1,211 reviews or larger. For the word "hate", only 16 of the 100,000 simulations had a difference of 94 reviews or higher, which lets us conclude with more than 99.9% confidence that the 94-review difference is not due to chance.

3. WHAT NOW?

After exploring and visualizing our IMBD dataset and validating some of the observations we made with inferential statistics, we can move on to the training process of our predictive model. Our predictive model will try to predict whether a review is positive or negative solely from the text of the review. The observations tested above, lets us know that our Keras deep learning model will be aided by the fact that there is indeed a difference between positive and negative reviews and the words found in each.