Individual Report

Giovanni Luca Favuzzi

Contents

1 Introduction

This following report is a shallow exploration of the older **Yelp Dataset**, provided for the **Yelp Challenge**, following a mere statistical approach.

1.1 Report Goals

Two main problems are tackled:

1. Consistency of the Yelp user's ratings of Italian businesses across the three metropolitan areas of Nevada, Quebec and Arizona.

2. Investigation of the empirical probability distributions for both user's and business' reviews about Italian businesses in the Nevada metropolitan area.

1.2 Data Overview

The **Yelp Dataset** contains information about ratings of businesses around the world. It is divided into 5 json-files regarding: Businesses, Users, Reviews, Check-ins, Tips. This said, for the purpose of this report, only the Businesses and Reviews jsons are needed, in particular, from the first, using Pandas, it is rather easy to retrieve the unique business IDs and the associated metropolitan zone; from the Reviews json, using the same approach, it is possible to retrieve all the reviews containing the associated unique user ID, the given rating and the associated unique business ID.

1.3 Hypotheses

A set of assumptions are needed before proceeding:

1. The data, contained in the provided dataset, is clean and correct *enough*, meaning that even though there is some incorrect or missing information, is not enough to generate relevant deviations in the final results.

2. The data is likely to follow two main distributions: either **Gaussian** or **Power-law**, with parameters estimated with standard techniques such as **Maximum Likelihood Estimation** or Bootstrapping.

3. The data is probably positively biased. This phenomenon is well known for recommendation systems and it makes sense to expect some sort of behavior. This happens because people tend to review what they like instead of what they don't like.

4. The user's bias, affecting reviews, is assumed near *enough* the average user bias to avoid performing de-biasing for each user. This is usually present for explicit ratings systems (ratings 1 to 5).

Some of these assumptions can be considered strong, however it is reasonable to assume that the final result won't be *significantly* deviated from a much more accurate one.

2 Methodology

2.1 User Ratings' Consistency

To have a proper understanding of the ratings distribution across each metropolitan area , the users' reviews are clustered depending on the rating value and the related area.
Done with the clustering, the values are normalized with respect to the number of reviews, in order to have a correct comparison between distributions.

Finally, the moments of the empirical distribution are computed to have further information about the empirical distributions.
Of all the moments, only the *first four* are estimated. Leaving out the usual mean and standard deviation, the third moment, **Skewness**, and fourth moment, **Kurtosis**, are needed to have a better understanding of the shape of the distribution. Moreover, instead of the standard Kurtosis, this report will focus on a variation of the fourth moment, called **Excess Kurtosis**, given by the following exact formulation:

$$ExcessKurt[X] = E[(\frac{X-\mu}{\sigma})^4] - 3, \quad \mu\text{: mean, } \sigma\text{: standard deviation}$$

The Excess Kurtosis explicitely computes how much "tailed" the considered distribution is w.r.t. a Gaussian.

2.2 User's Reviews Distribution

Retrieved the reviews about Italian businesses in Nevada and the unique user IDs, the lists are *joined* against each other for matching user ID. From the joined list, is computed the number of written reviews for each user, ending up having a list containing the review-cardinality for each user.

In order to properly understand which distribution best suits the data, the first four moments are *estimated*. Both **Mean** and **Standard Deviation** are needed for a Gaussian fitting. Instead, to have a proper Power-law fitting, both the left and right tail exponent, if possible, are estimated. For the tails, the estimation criteria is Maximum Likelihood Estimation, other than a Bootstrapping, since it is accurate enough and less computationally requiring.

To effectively estimate the right tail exponent α_{right}, first, is defined the percentage of *relevant* data p, then, the reviews cardinality list is ordered and finally the relevant data r_{rel} is selected as the first *1-p* percentage.
In order to fit the left tail instead, the relevant data is selected as the last percentage p and is taken in *absolute value*.
The MLE formulation for both tail exponents is given by:

$$\alpha_{MLE} = \frac{N}{\sum_i log\frac{r_{rel,i}}{min(r_{rel})}}$$

Afterwards, the **PDF** are plotted on a on a vertical semi-logarithmic scale and the **CCDF** are plotted on a fully-logarithmic scale. This way, two different point of views are available to have a better understanding of the fitting.

2.3 Business' Reviews Distribution

The same process is repeated for the number of reviews received by each Italian business in Nevada.

3 Results

3.1 User Ratings' Consistency

The aforementioned procedure leads to the following distributions, plotted on a linear plane. The connection between edges is pointed out by the dashed lines to make it easier spotting the difference between the rating progressions.

The following table displays the first four moments for each area.

FIRST FOUR MOMENTS				
AREA	MEAN	STANDARD DEVIATION	SKEWNESS	EXCESS KURTOSIS
Nevada	3.74	1.35	-0.81	-0.59
Quebec	3.71	1.25	-0.82	-0.33
Arizona	3.85	1.33	-0.96	-0.30

3.2 User's Reviews Distribution

Considering only the top $p = 10\%$ as relevant, the estimate of the exponent of the right tail is:

$\alpha_{right} = 2.018$.

Coherently with the high values of Skewness and Excess Kurtosis, the estimate of the left tail, instead, is *not feasible*, since its value tends to *infinite*.

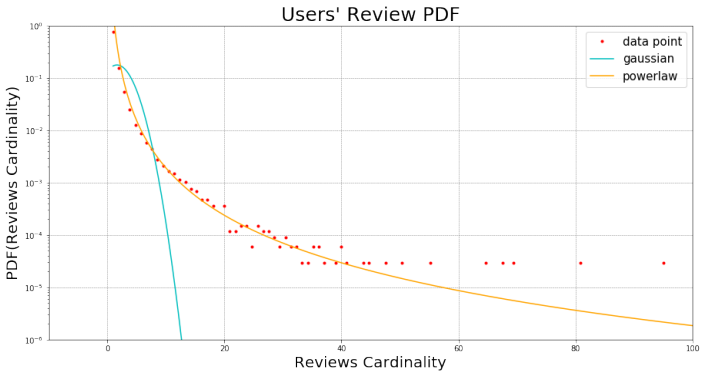
The resulting estimated Gaussian and Power-law PDF and CCDF are shown in the two plots below, respectively on the upper left and right corners.

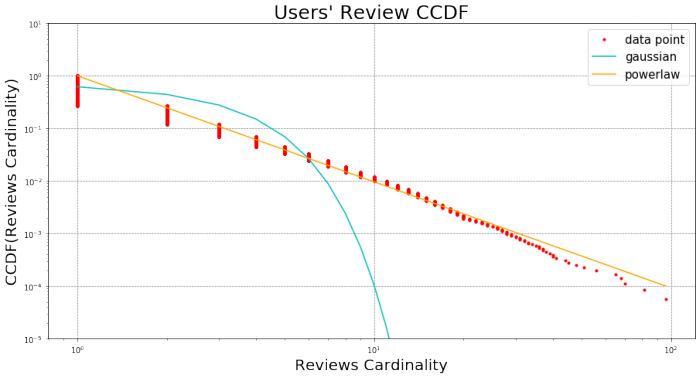
3.3 Businesses' Reviews Distribution

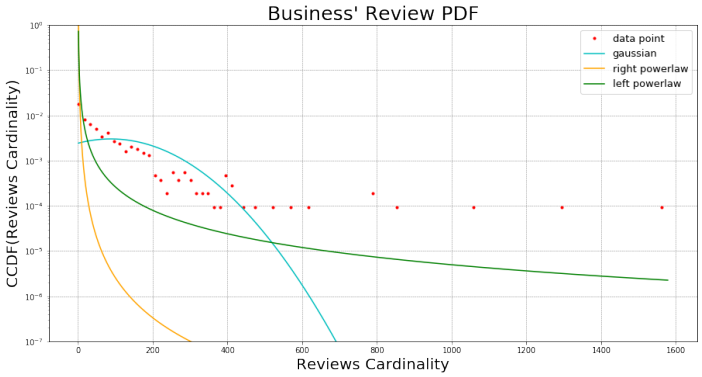
The top 10% relevance has been used for the business reviews distribution too.
This time however, it is possible to estimate both tail exponents, their estimates are:

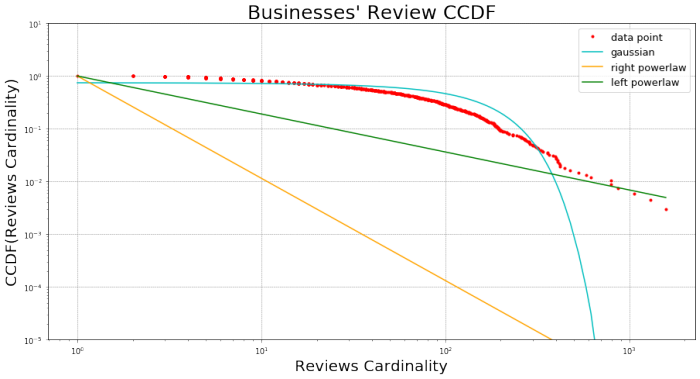
$\alpha_{right} = 1.94$
$\alpha_{left} = 0.72$.

The data-points are finally plotted against the resulting estimated Gaussian and Power-law PDFs and CCDFs, in the two plots respectively on the lower left and right corners.









4 Discussion and Conclusion

4.1 User Ratings' Consistency

The results concerning the user's ratings show clearly how there is *correlation* between the two American areas of Nevada and Arizona: both their distributions seem to mimic an *exponential* distribution in the leftmost part but look *linear* in the rightmost part.
Quebec reviews follow a sort of exponential distribution, but right before the end, inflect rather unexpectedly, given the first two cases. Still, considering the zone difference, this is pretty coherent and likely to have some sort of *correlation* with *latitude* and *longitude* due to the cultural differences between the Canadian and American situation.

These considerations, however, are much more hardly spotted by only looking at the first four moments. In fact, it would make sense to have similar moments for the metropolitan areas of Nevada and Arizona and different from the Quebec area. It is easy to notice, however, that this is not the case: the moments are rather similar between all three areas, without showing any sharp difference between the American and Canadian distributions. One last mention on how the *third hypothesis*, that was previously made, *is correct*: the mean for all three areas is higher than the actual score mean of 3. This shows how, on average, the positiveness of the reviews is shared between every user.

4.2 User's Reviews Distribution

The retrieved *35520* user reviews distribute in an *exponential* fashion. This is observable by just looking at the data-points and confirmed by the estimated Power-law distribution that *fits* the dataset nicely both in the PDF and CCDF plot, showing fitting uncertainty only at the very end of the right tail. This consideration is strengthened by the value of the right tail exponent, which is near 2.

The complete asymmetry of the distribution is coherent with a pretty high positive $Skewness = 11.87$. For such a unimodal distribution, it shows how fitting the dataset with a Power-law distribution parametrized by an estimation for the right tail exponent, is a good idea.
On the other hand, the $ExcessKurtosis = 273.58$ confirms that a Gaussian fitting cannot work. The Gaussian distribution, in fact, is completely misfitting the dataset starting from the very beginning of the distribution.

4.3 Businesses' Reviews Distribution

The reviews relates to the *680* retrieved Italian businesses in Nevada follow a much *less defined* distribution than the Nevada users' review distribution. This makes it harder to find a distribution that fits this data-points properly.

Even though in this case it is possible to have an estimation for both left and right tail exponent, by plotting the PDFs and CCDFs of the two estimated Power-law and Gaussian, *none* of those is fitting the data properly. By looking at the PDF plot, it seems that the left Power-law is fitting the points better than the alternatives, however, by looking at the CCDF plot, the best fit seems to be given by the Gaussian for most of the distribution up until the rightmost part that, instead, distributes as a Power-law. In fact, the right tail exponent, which is smaller than 2, indicates how the right tail is rather *fat*, and cannot be fit by a mere Gaussian.

It is safe to say that the *hypothesized* distributions do *not fit* the data properly.

5 References and Word Count

5.1 References

- **matplotlib**: Python library used to *plot* distributions.
- **pandas**: Python library used to *retrieve* information.
- **numpy, scipy**: Python libraries used to perform *computations*.

5.2 Word Count

1500 words