Intro to Data Science HW 1

Ben Green

Contents

February 23, 2015

Question 1

1a.

Feature Name	Feature Type
`producer`	nominal
`release_to_review_time`	interval
`used_real_name`	binary
`verified_purchase`	binary
`rating`	ordinal
`helpfulness`	ratio
`number_of_votes`	ratio
`length_of_review_text`	ratio

1b.

The mode of `producer` is Apple, with 4480 entries.

1c.

5077/9585, or 53% of reviewers used their real name and had a verified purchase.

1d.

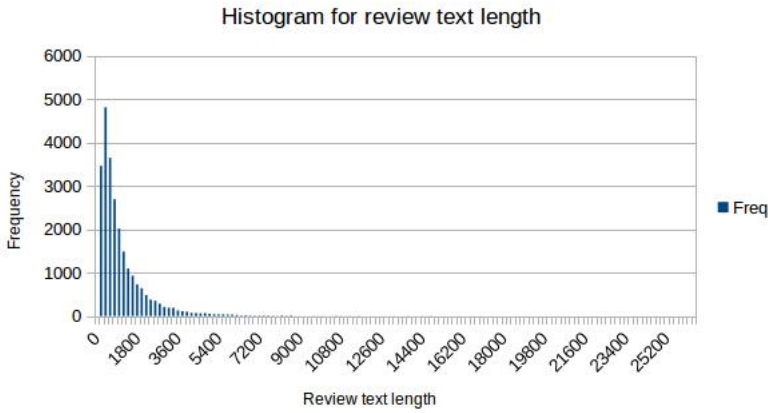
5077/13989, or 36% of reviews with a verified purchase had a reviewer who used their real name.

1e.

Measure	**Value**
min	-537
q1	74
median	144
q3	290
max	11686
interquartile	215

These numbers can be displayed conveniently in a boxplot.

1f.



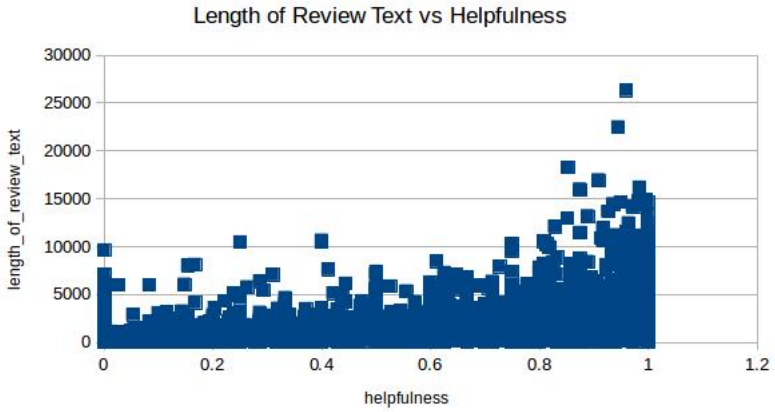
1g.

Yes, the distribution of `length_of_review_text` is heavily skewed towards the shorter end. There are outliers of 18275, 22492, and 26332.

1h.

The Pearson correlation value between `length_of_review_text` and `helpfulness` is approximately .25. This indicates that there is a slight positive correlation between these two variables.

1i.



Question 2

Regarding the AND construction, $r = 3$ being the number of hash functions and the LSH family {d1,d2,0.6,0.4}, the new family that is derived is:

$$\{d1,d2,0.6^3,0.4^3\}$$

$$w = .216$$
$$x = .064$$

Regarding the OR construction, $r = 3$ being number of hash functions and the LSH family {d1,d2,.216,0.64}, the new family that is derived is:

$$\{d1,d2,1\text{-}(1\text{-}(0.6))^3,1\text{-}(1\text{-}(0.4))^3\}$$

$$y = 0.936$$
$$z = 0.784$$

Question 3

3a.

Sketch of vector u = [1.25, -1.75, 1.75]
Sketch of vector v = [0.95, -0.95, 1.35]
Sketch of vector w = [-0.35, 1.65, 1.85]

The sketches were constructed by taking the dot products of each vector with each randomly generated vector.

3b.

$$\frac{u \cdot v}{\|u\| \cdot \|v\|} = 0.949 \tag{1}$$

$$\frac{u \cdot w}{\|u\| \cdot \|w\|} = 0.017 \tag{2}$$

Question 4

4a.

The Mahalanobis distance reduces to the Euclidian distance when the covariance matrix is the identity matrix.

4b.

The Mahalanobis distance reduces to the Euclidian distance when the covariance matrix is a diagonal matrix.

Question 5

5a.

The minihash signatures for each column are as follows:

S_1	S_2	S_3	S_4	S_5
5	5	1	1	1
5	2	2	2	2
3	0	1	4	0

5b.

Only h_2 is a true permutation.

5c.

Estimated Jaccard Similarities:

1-2	**1-3**	**1-4**	**2-3**	**2-4**	**S6**
0	0	1/4	0	1/4	1/4

True Jaccard Similarities:

1-2	**1-3**	**1-4**	**2-3**	**2-4**	**S6**
1/3	1/3	1/3	2/3	2/3	2/3

Question 6

The stop-word based shingles for the sentences are:

{and Mrs.Dursley, of number four, to say that, that they were}

{for the first, the first time, an argument had, at number four}

There are no matches between shingles of each sentence, so the Jaccard Similarity is 0/8.

Question 7

The logit function can be used to transform values obtained from a logistic equation to generate a sequence of random numbers with an (almost) normal distribution.