Yelp Data Prediction

Shiwei Cao Jiyun Chen Jing Guo

Department of Statistics

University of Wisconsin-Madison

March 14, 2018

Our Model?

- Thesis statement: Regularized logistic regression model is a fast and accurate classifier. We spend less than 3 minutes on training a model and yields RMSE = 0.6315.
- Model: Logistic regression
- RMSE: 0.6315
- Training Time: 3 mins

What makes reviews positive or Negative?

Most Positive Words

delicious	amazing	great
excellent	awesome	fantastic
best	perfect	perfection
perfectly	incredible	outstanding
favorite	phenomenal	perfekt
love	gem	loved
superb	heaven	wonderful

Most negative Words

worst	poisoning	horrible
terrible	zero	awful
inedible	disgusting	tasteless
bland	disappointing	cockroach
waste	notrecommend	flavorless
notworth	rude	poor
unacceptable	disappointment	disgusted

Content

- Data Cleaning
- Model Fitting
  - Logistic
  - Neural Network
- Conclusions
  - Word Importance
  - Strength and Weakness

Data Cleaning Procedure

Data Cleaning

Step 1:

- Keep a-z A-Z and few punctuations(.,?!')
- Other characters -> ''

  E.g. naïve ⟶ nave

- Add space before (after) the punctuations (except ')

  E.g. I usually wouldn't review a place just to talk mess,

  ⬇

  I usually wouldn't review a place just to talk mess ,

Data Cleaning

Step 2:

- Remove meaningless English stopwords
  - e.g: 'the', 'a', 'and', 'i', 'was', 'it', 'is', 'we', 'that', 'this', 'my', 'you'

Step 3:

- Make a dictionary for text(words) and remove the words whose frequency are smaller than 250

Data Cleaning

Step 4:

- Combine the negation words with verbs and adjectives

don't, wouldn't, hasn't, weren't, shouldn't, ain't, isnt, didn't, couldn't not, cannot, never…

high, live, fresh, nice, excellent, prime, great, second, awesome, different, dry, extra, worth, busy, short, tough, last, same, good …

You guys are terrible! I honestly don't have anything nice to say about this place... I usually wouldn't review a place just to talk mess

→

guys terrible honestly notnice anything say about place usually notreview place just talk mess

Data Cleaning

Step 4:

- Combine the negation words with verbs and adjectives

don't, wouldn't, hasn't,
weren't, shouldn't, ain't,
isnt, didn't, couldn't
not, cannot, never…

like, liked, likes, love, loved, loves,
recommend, recommended,
recommends, prefer, prefers,
advocate, disappoint …

"Long wait time to get in in a half empty cafe. Long time to get drinks ... Would not recommend"

→

"long wait time in in half empty cafe long time drinks notrecommend"

Data Cleaning

Step 5:

- Remove the rest punctuations and the extra space

Model Fitting

Generating Model Matrix

- **TFIDF($TF \times IDF$)**: Scale down the impact of words with high frequency

$$TF(w,t) = \frac{\#w\ in\ t}{\#\ words\ in\ t} \qquad IDF(w,t) = log\frac{\#words\ in\ t}{\#texts\ that\ contain\ w}$$

$$\begin{array}{ccccc} \textbf{This} & \textbf{is} & \textbf{not} & \textbf{good} & \textbf{delicious} \end{array}$$
$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0.46 & 0.46 & 0.6 & 0.46 & 0 \\ 0.52 & 0.52 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

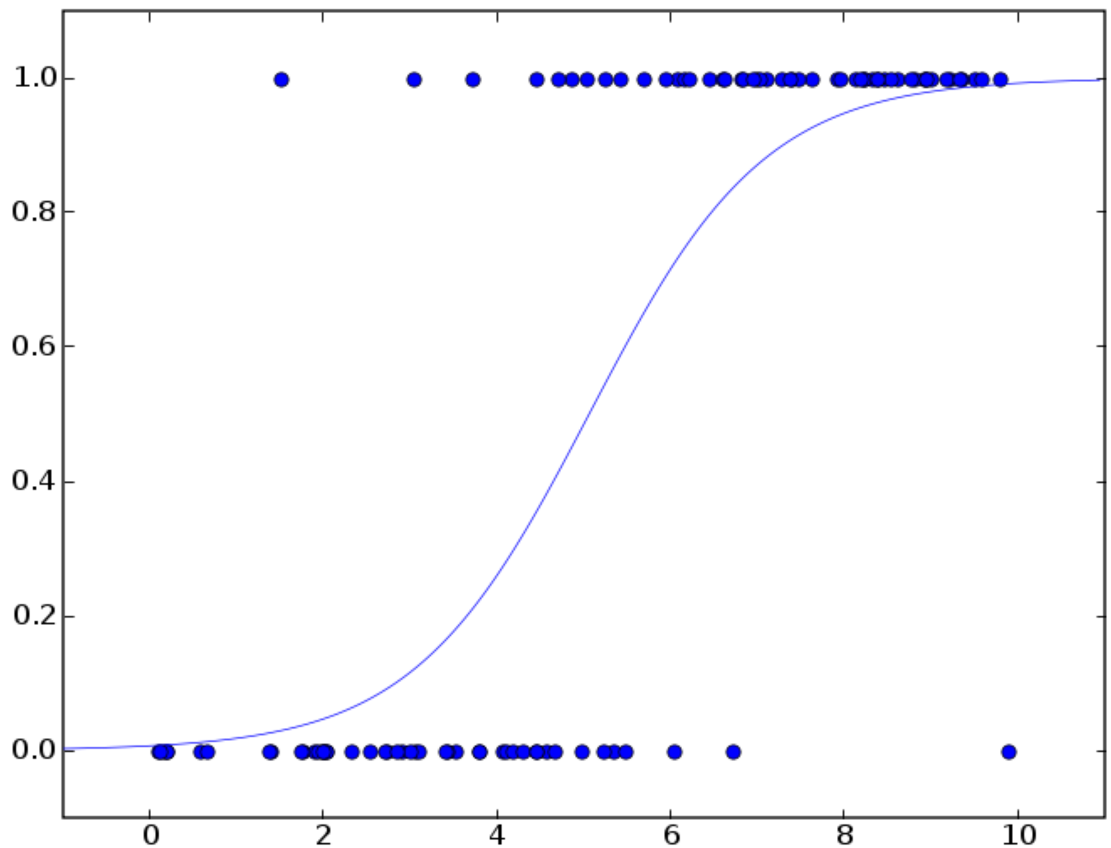
- **Generating Sparse Matrix:** combine three matrix

text, punctuation, category → final matrix

$$\begin{pmatrix} 1546379\times16542 \end{pmatrix} \begin{pmatrix} 1546379\times5 \end{pmatrix} \begin{pmatrix} 1546379\times607 \end{pmatrix} \Rightarrow \begin{pmatrix} 1546379\times17154 \end{pmatrix}$$

Logistic Regression

$$\log \frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 x_1 + \cdots + \epsilon$$



Logistics

$$P(Y_i = k) = \frac{\exp(X_i\beta_k)}{1 + \sum_{k=1}^{4} \exp(X_i\beta_k)}, k = 1, 2, 3, 4$$

$$P(Y_i = 5) = \frac{1}{1 + \sum_{k=1}^{4} \exp(X_i\beta_k)}$$

P_i: Probability of i-star rating
m: numbers of text
n : number of unique words

$$J(\beta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{5} I_{\{Y_i=j\}} \log P(Y_i = j)\right] + \frac{\lambda}{2}\sum_{k=1}^{5}\sum_{j=1}^{n} \beta_{kj}^2$$

Tuning Parameter

Train set: 20%   Validation set: 80%

1/λ	**1**	**1.5**	**2**	**2.2**	**2.5**	**3**
MSE on validation set	0.43	0.4271	0.4262	0.4261	0.4262	0.4266

```
LogisticRegression( C=2.2,
                    penalty="l2",
                    multi_class="multinomial",
                    solver="saga")
```

Neural Network

- **Word Embedding:**
  mapping words or phrases to vectors of real numbers with fixed dimensions
  - amazing ≈ awesome
  - waitress = waiter + woman – man
  - hotdog doesn't match in "He is a very gentle hotdog"

- **FastText:**
  Open-source library for text classification and representation provided by Facebook

epoch	25	50	100	200
RMSE	0.74	0.7	0.65	0.65
Training Time	5 mins	15 mins	~ 60 mins	> 2 hours

Conclusions

Comparison

	Lasso	Regularized Logistic	FastText
RMSE	0.8301	**0.6315**	0.6532
Time	60 mins	**3 mins**	60 mins

Results – Most Effective Positive Words

Results – Most Effective Negative Words

Strength

- Fast: 3 minutes
- Accurate: RMSE of 0.6315
- Simple and Interpretable: each word has an coefficient
- Foreign language can be predicted: like German.
  - “schlecht”, “nie”, “kalt”
  - “hervorragend”, “begeistert”, “klasse”, “leckeres”

Weakness

- Less accurate comparing to multi-layer neural network
- Did not adjust vocabulary roots, which may cause redundancy
  - “prefect” and “perfection”
- Multicollinearity exists
- Cannot detect slangs directly
  - “hands” → “food hands down”(means great)
  - “charts” → “off the charts”(means great)

Thank You!