Counting Stars

Final Project Presentation
By Daniel Conde

Background

Goal

Create a model that predicts the star rating of an Amazon product.

- Use basic ML techniques studied in class
- Sentiment analysis on customer's reviews to predict his/her rating

Motivation

Sometimes, an Amazon product can become subject to low reviews for reasons unrelated to the actual product.

- Poor shipping and handling
- Wrong sizing for clothing and apparel
- Vague/unclear product description

→ Will results of model be affected?


Process

Source(s) Used

Julian McAuley, Associate Professor from UCSD

- http://jmcauley.ucsd.edu/data/amazon/
- Datasets (20gb raw) including reviews and metadata from 1996-2014
- Sample data collected:


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

Data Preprocessing

Python Pandas

- Read JSON and filter columns
  - ASIM
  - Summary
  - Review
  - Star Rating
- Replace punctuation and odd symbols with “”

Amazon AWS & Zepl

- AWS: Upload finalized csv dataset to S3 storage service
- Zepl: cloud-based platform designed to process larger amounts of data → more favorable than running locally on my laptop

Data Preprocessing

Zepl → Pyspark

Added columns for length of review and transformations

- Tokenize reviews (lowercase and separate w commas)
- Filter stop words (unimportant)
- Term frequency (hashingTF)
- IDF

Add feature vector (results of IDF) at the end of pipeline

Data Preprocessing

Sample Data Before Test:

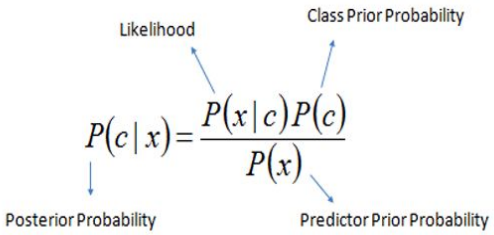
```
+----------+-----------+--------------------+--------------------+--------------+--------------------+--------------------+--------------------+--------------------+
|      ASIN|Star_Rating|             Summary|        Clean_Review|length(Summary)|          token_text|         stop_tokens|          hash_token|           idf_token|            features|
+----------+-----------+--------------------+--------------------+--------------+--------------------+--------------------+--------------------+--------------------+
|B000H00VBQ|          2|A little bit bori...|I had big expecta...|            26|[a, little, bit, ...|[little, bit, bor...|(262144,[16332,53...|(262144,[16332,53...|(262145,[16332,53...|
|B000H00VBQ|          5|Excellent Grown U...|I highly recommen...|            21|[excellent, grown...|[excellent, grown...|(262144,[72090,11...|(262144,[72090,11...|(262145,[72090,11...|
|B000H00VBQ|          1|Way too boring fo...|This one is a rea...|            21|[way, too, boring...|       [way, boring]|(262144,[16332,53...|(262144,[16332,53...|(262145,[16332,53...|
|B000H00VBQ|          4|Robson Green is m...|Mysteries are int...|            27|[robson, green, i...|[robson, green, m...|(262144,[15889,11...|(262144,[15889,11...|(262145,[15889,11...|
```

Data Preprocessing

Sample Data Before Test:

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|    2|(262145,[16332,53...|
|    5|(262145,[72090,11...|
|    1|(262145,[16332,53...|
|    4|(262145,[15889,11...|
|    5|(262145,[52805,91...|
|    5|(262145,[24417,33...|
|    3|(262145,[2711,501...|
|    3|(262145,[83656,19...|
|    5|(262145,[21427,26...|
|    3|(262145,[10287,84...|
|    4|(262145,[102787,1...|
|    4|(262145,[10287,17...|
|    3|(262145,[122516,2...|
|    3|(262145,[16332,25...|
|    5|(262145,[33933,10...|
```

Model



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Naive Bayes model

- python.ml.classification
- Simple, commonly-used machine learning classifier; popular for text-classification and categorization
- Used to classify text in customer reviews

Results

Accuracy Score

7.2%

0.7:0.3 Training/Test Split

8.6%

0.8:0.2

7.9%

0.95:0.05

Analysis

Common Words in Both Good and Bad Reviews:

	★	★★	★★★	★★★★	★★★★★
“late”	•	•		•	•
“very”, “extremely”	•	•			•
“bright”, “warm”		•	•	•	
“break”	•		•	•	

Analysis

- Lack of consistency with reviews
- Too many categories (star ratings) as opposed to “positive/negative” classification

What Could Have Been Done Better

Different Data - Amazon Instant Video

- Eliminates the problem of “shipping and handling,” “damaged” products, etc.
- Words are more consistently associated with the rating given
    - “Dull” will almost never be part of a 5-star rating

Different Model - Regression

Naïve Bayes → has a naive assumption of conditional independence for every feature, which means that the algorithm expects the features to be independent which not always is the case.

Regression → linear classification method that learns the probability of a sample belonging to a certain class.

NEW MODEL: Logistic Regression / ~~Linear Regression~~

Logistic regression tries to find the optimal decision boundary that best separates the classes.

Thank You.