School of Engineering and Applied Science (SEAS), Ahmedabad University

B.Tech (CSE Semester VI):
Machine Learning (CSE 523)

Project Submission #2: Linear Regression

Submission Deadline: March 04, 2020 (11:59 PM)

- **Group No.: 8**
- **Project Domain: Natural Language Processing**
- **Project Title: Sentimental Analysis on movie reviews**
- **Name of the group members:**
  1. Aditya Shah (1741007)
  2. Dhruvil Shah (1741024)
  3. Nilay Patel (1741038)
  4. Varun Patel (1741080)

Contents

1 Implementation code

```
import numpy as np
from sklearn.datasets import load_files
reviews_train = load_files("aclImdb/train/")    #loading training file
text_train, y_train = reviews_train.data, reviews_train.target

print("Number of documents in train data: {}".format(len(text_train)))   #printing
     document in train folder
print("Samples per class (train): {}".format(np.bincount(y_train)))   #No.of
    samples

reviews_test = load_files("aclImdb/test/")    #loading testing file
text_test, y_test = reviews_test.data, reviews_test.target

print("Number of documents in test data: {}".format(len(text_test)))      #
    printing document in test folder
print("Samples per class (test): {}".format(np.bincount(y_test)))
from sklearn.feature_extraction.text import CountVectorizer
vect = CountVectorizer(min_df=5, ngram_range=(2, 2))        #to tokenize and to
    build new words
X_train = vect.fit(text_train).transform(text_train)   # fitting the model and
    transform into vector
X_test = vect.transform(text_test)   #transforming into vector

print("Vocabulary size: {}".format(len(vect.vocabulary_)))
print("X_train:\n{}".format(repr(X_train)))
print("X_test: \n{}".format(repr(X_test)))

feature_names = vect.get_feature_names()
print("Number of features: {}".format(len(feature_names)))  #No.of features

from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
param_grid = {'C': [0.001, 0.01, 0.1,1 , 10]}   #Taking  parameter with discrete
    value
grid = GridSearchCV(LogisticRegression(), param_grid, cv=5)  #Searching by Cross
    validation = 5
grid.fit(X_train, y_train)

print("Best cross-validation score: {:.2f}".format(grid.best_score_))
print("Best parameters: ", grid.best_params_)
print("Best estimator: ", grid.best_estimator_)

import matplotlib.pyplot as plt
import mglearn
mglearn.tools.visualize_coefficients(grid.best_estimator_.coef_, feature_names,
    n_top_features=25)  #to create figures
plt.show()

lr = grid.best_estimator_
lr.fit(X_train, y_train)   #fitting the model
lr.predict(X_test)  #predicting value
print("Score: {:.2f}".format(lr.score(X_test, y_test)))

pos = ["I've seen this story before but my kids haven't. Boy with troubled past
    joins military, faces his past, falls in love and becomes a man. "
       "The mentor this time is played perfectly by Kevin Costner; An ordinary man
     with common everyday problems who lives an extraordinary "
       "conviction, to save lives. After losing his team he takes a teaching
```

```
    position training the next generation of heroes. The young troubled "
49      "recruit is played by Kutcher. While his scenes with the local love
    interest are a tad stiff and don't generate enough heat to melt butter, "
50      "he compliments Costner well. I never really understood Sela Ward as the
    neglected wife and felt she should of wanted Costner to quit out of "
51      "concern for his safety as opposed to her selfish needs. But her presence
    on screen is a pleasure. The two unaccredited stars of this movie "
52      "are the Coast Guard and the Sea. Both powerful forces which should not be
    taken for granted in real life or this movie. The movie has some "
53      "slow spots and could have used the wasted 15 minutes to strengthen the
    character relationships. But it still works. The rescue scenes are "
54      "intense and well filmed and edited to provide maximum impact. This movie
    earns the audience applause. And the applause of my two sons."]
55 print("Pos prediction: {}". format(lr.predict(vect.transform(pos))))
56
57 neg = ["David Bryce\'s comments nearby are exceptionally well written and
    informative as almost say everything "
58      "I feel about DARLING LILI. This massive musical is so peculiar and over
    blown, over produced and must have "
59      "caused ruptures at Paramount in 1970. It cost 22 million dollars! That is
    simply irresponsible. DARLING LILI "
60      "must have been greenlit from a board meeting that said \"hey we got that
    Pink Panther guy and that Sound Of Music gal... "
61      "lets get this too\" and handed over a blank cheque. The result is a hybrid
     of GIGI, ZEPPELIN, HALF A SIXPENCE, some MGM 40s "
62      "song and dance numbers of a style (daisies and boaters!) so hopelessly old
     fashioned as to be like musical porridge, and MATA HARI "
63      "dramatics. The production is colossal, lush, breathtaking to view, but the
     rest: the ridiculous romance, Julie looking befuddled, Hudson "
64      "already dead, the mistimed comedy, and the astoundingly boring songs
    deaden this spectacular film into being irritating. LILI is"
65      " like a twee 1940s mega musical with some vulgar bits to spice it up. STAR
    ! released the year before sadly crashed and now is being "
66      "finally appreciated for the excellent film is genuinely is... and Andrews
    looks sublime, mature, especially in the last half hour....."
67      "but LILI is POPPINS and DOLLY frilly and I believe really killed off the
    mega musical binge of the 60s..... "
68      "and made Andrews look like Poppins again... which I believe was not
    Edwards intention. Paramount must have collectively fainted "
69      "when they saw this: and with another $20 million festering in CATCH 22,
    and $12 million in ON A CLEAR DAY and $25 million in PAINT YOUR WAGON...."
70      "they had a financial abyss of CLEOPATRA proportions with $77 million tied
    into 4 films with very uncertain futures. Maybe they should have asked seer "
71      "Daisy Gamble from ON A CLEAR DAY .....LILI was very popular on immediate
    first release in Australia and ran in 70mm cinemas for months but it failed "
72      "once out in the subs and the sticks and only ever surfaced after that on
    one night stands with ON A CLEAR DAY as a Sunday night double. Thank "
73      "god Paramount had their simple $1million (yes, ONE MILLION DOLLAR) film
    LOVE STORY and that $4 million dollar gangster pic THE GODFATHER "
74      "also ready to recover all the $77 million in just the next two years....
    for just $5m.... incredible!"]
75 print("Neg prediction: {}". format(lr.predict(vect.transform(neg))))
76
```

2 URL links

2.1 Main folder

https://drive.google.com/open?id=1cfIl5BJaIliIsV5_aYnx3d6ygzzYa4pB

2.2 Codes

https://drive.google.com/open?id=18ea_fqozyvh5Itk6U9yIK_mxsIvK5beY

2.3 Data set

https://drive.google.com/open?id=1Rbs4SUiWbtkxpBzS7jrJ5_ep28tMCF9P

3 Inference

3.1 What we have done and why is it important?

We have done separation of positive and negative reviews by giving 1 and 0 and then we have first taken one word and then see accuracy on the training and unseen test data and then we have taken word pair and perform same operation. We have implemented tfidf model, so that large number of words and documents can be understand by computer, so we vectorize and transform it into numericals. This is important because large number of documents can be easily form into small numbers so that vectorization and transformation can be easily done. We have applied logistic regression because we are giving input data and by grid estimator and parameter of cross validation and fitting model we are predicting magnitude whether it is negative or positive.

3.2 How we have implemented?

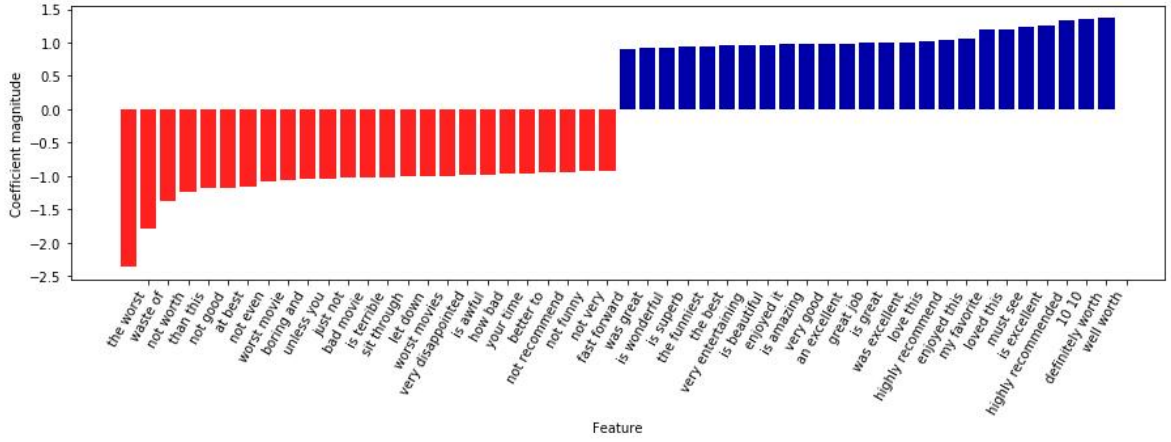
We have implemented this above method by giving path of test and train and then reading files and see whether document or text is positive or negative. If positive then append 1 and if negative then append -1 and then we vectorize the whole and transforming whole. Then we spilt data by stopwords and then taking one word or features calculating accuracy of system and then same with word pair. Then we implement tfidf to divide into more smaller into numerical so that computer can understand. The same we apply for testing data . After that as we have 1 and 0, we decided to implement logistic regression so that we can train or model with parameter by taking 5 discrete numbers as Cross vaidation points and then finding best grid estimator and then calculating magnitude and plotting graph and by best grid estimator we fit model with training or testing set and then predicting particular value and printing score prediction.

3.3 Analysis and Results

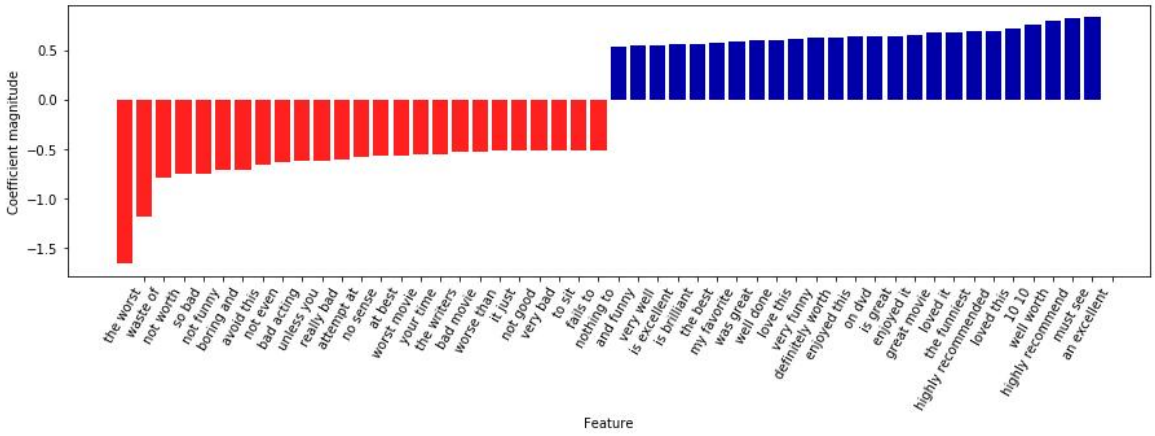
$0 \leq R_\theta(x) \leq 1$
$h_\theta(x) = g((\theta)^T \cdot x)$
$g(x) = \frac{1}{1+e^{-x}}$;g(x) is Logistic function
$h_\theta(x) = \frac{1}{1+e^{-\theta^T \cdot x}}$
$Trainingset = (x_1, y_1), ..., (x_m, y_m)$
m examples x$\epsilon[x_0, x_1, ..., x_n]$
$x_0 = 1$, y = 0 or 1
$h_\theta(x) = \frac{1}{1+e^{-\theta^T \cdot x}}$
Cost function is given by:
$J_\theta = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}(h_\theta x_i - y_i)$
$= \frac{1}{m} \sum_{i=1}^{m} cost(h_\theta x_i, y)$
where,
$cost(h_\theta(x), y) = \frac{1}{2}(h_\theta x - y)^2$
$cost(h_\theta(x), y) = -log(h_\theta(x))$ if y=1
$cost(h_\theta(x), y) = -log(1 - h_\theta(x))$ if y=0
$cost(h_\theta(x), y) = -ylog(h_\theta(x)) - (1-y)log(1 - h_\theta(x)$
If y=1,
$cost(h_\theta(x), y) = -log(h_\theta(x))$
If y=0,
$cost(h_\theta(x), y) = -log(1 - h_\theta(x))$
$J_\theta = -\frac{1}{m}[\sum_{i=1}^{m} y_i log(h_\theta(x_i)) + (1-y_i)log(1 - h_\theta(x_i))]$
To fit parameter θ,

Minimize $J_\theta w.r.t \theta and get \theta$
To make a prediction given new x:
Output: $h_\theta(x) = \frac{1}{1+e^{-\theta^T \cdot x}}$
p(y=1 — x,θ)
To minimize cost function,
Apply Gradient descent Algorithm and Repeat until convergence
$\theta_j = \theta_j - \alpha \frac{\Delta J}{\Delta \theta_j}$
Simultaneously update all θ_j
$\frac{\Delta J}{\Delta \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta x_i - y_i) x_j^{(i)}$
Repeat
$\theta_j = \theta_j - \alpha \sum_{i=1}^{m} (h_\theta x_i - y_i) x_j^{(i)}$
$\theta = [\theta_0, ..., \theta_n]$

Logistic regression for training data



Logistic regression for testing data



Parameters are used to find the best regression model which generalizes well i.e. works well with training as well as testing data sets by minimizing the mean squared error.