Assignment 1: Bayesian Classification

Szymon Pawlica

R00187226

Task 1:

Lines [26-43] Split the dataset into training and testing sets

Split the dataset into two subsets (training and testing) using pandas, then get the count, data and labels from the respective subsets.

Return these subsets.

Task 2:

Lines [18-23] Get rid of any special characters and convert to lowercase

Convert the 'Review' column in the dataset to not contain any special characters, convert into lowercase and separate into a list containing every word in the review. I realise this could've been done using numpy late into the project, but the method works so I didn't change it.

Lines [46-64] Find the occurrence count for each word

For every word in every review in the training subset count how many times it occurs and place it into a dictionary containing the word and its total occurrence count.

Return this word:count dictionary.

Task 3:

Lines [67-86] Separate the training split into positive and negative reviews

Split the training subset into Negative and Positive Reviews.

Find in what reviews each word appears

Count the amount of positive and negative reviews in which each word found in Task 2 appears.

Return these positive:count and negative:count dictionaries.

Task 4:

Lines [88-104] Find what are the chances of a word occurring in a positive or negative review

Applying a Laplace smoothing of 1, find the probability that a word is present in a positive review

$P[word\ is\ present\ in\ review|review\ is\ positive]$

And that a word is present in a negative review

$P[word\ is\ present\ in\ review|review\ is\ negative]$

Also find the probability a review is positive

$P[review\ is\ positive]$

And the probability a review is negative

$P[review\ is\ negative]$

Return these probabilities.

Task 5:

Lines [106-123] Based on the input 'review' calculate the chances of it being a positive or negative review

Predict whether the 'review' that is input into the function is negative (0) or positive (1), add up the logs of each word in the 'review' for both positive and negative probabilities (found in Task 4) of that word respectively and compare the exponent of both results (positive – negative) to the probabilities of a review being negative – positive (prior negative – prior positive).

Task 6: Lines [132-149, 162-170] Train the classifier, running Tasks 2-5

Create a k-fold cross-validation using 6 splits (no specific reason for 6 here), make k be the word length which will be 1 – 10 as specified. Use tasks 2-5 to train the data for each word length.

Run the test data against the trained classifier and get the accuracies for each split, get the mean accuracy for each word length (k) used, extract the word length with the highest mean accuracy and train the data again using that word length using tasks 2-5 then get the confusion matrix as well as the accuracy for the test labels compared to the predictions.

Sample output (the first line is input by user):

What is the minimum word occurrence: 1000

Positive reviews in the training set: 12500

Negative reviews in the training set: 12500

Positive reviews in the testing set: 12499

Negative reviews in the testing set: 12500

Split 1 - Using minimum word length of 1 the accuracy is: 0.7149

Split 2 - Using minimum word length of 1 the accuracy is: 0.7024

Split 3 - Using minimum word length of 1 the accuracy is: 0.7087

Split 4 - Using minimum word length of 1 the accuracy is: 0.7204

Split 5 - Using minimum word length of 1 the accuracy is: 0.7040

Split 6 - Using minimum word length of 1 the accuracy is: 0.7002

The mean accuracy for this split is 0.7084

Split 1 - Using minimum word length of 2 the accuracy is: 0.7132

Split 2 - Using minimum word length of 2 the accuracy is: 0.7264

Split 3 - Using minimum word length of 2 the accuracy is: 0.7192

Split 4 - Using minimum word length of 2 the accuracy is: 0.7384

Split 5 - Using minimum word length of 2 the accuracy is: 0.7232

Split 6 - Using minimum word length of 2 the accuracy is: 0.7105

The mean accuracy for this split is 0.7218

Split 1 - Using minimum word length of 3 the accuracy is: 0.7391

Split 2 - Using minimum word length of 3 the accuracy is: 0.7331

Split 3 - Using minimum word length of 3 the accuracy is: 0.7187

Split 4 - Using minimum word length of 3 the accuracy is: 0.7418

Split 5 - Using minimum word length of 3 the accuracy is: 0.7446

Split 6 - Using minimum word length of 3 the accuracy is: 0.7681

The mean accuracy for this split is 0.7409

Split 1 - Using minimum word length of 4 the accuracy is: 0.7346

Split 2 - Using minimum word length of 4 the accuracy is: 0.7540

Split 3 - Using minimum word length of 4 the accuracy is: 0.7403

Split 4 - Using minimum word length of 4 the accuracy is: 0.7473

Split 5 - Using minimum word length of 4 the accuracy is: 0.7578

Split 6 - Using minimum word length of 4 the accuracy is: 0.7391

The mean accuracy for this split is 0.7455

Split 1 - Using minimum word length of 5 the accuracy is: 0.7514

Split 2 - Using minimum word length of 5 the accuracy is: 0.7703

Split 3 - Using minimum word length of 5 the accuracy is: 0.7528

Split 4 - Using minimum word length of 5 the accuracy is: 0.7583

Split 5 - Using minimum word length of 5 the accuracy is: 0.7508

Split 6 - Using minimum word length of 5 the accuracy is: 0.7518

The mean accuracy for this split is 0.7559

Split 1 - Using minimum word length of 6 the accuracy is: 0.7291

Split 2 - Using minimum word length of 6 the accuracy is: 0.7144

Split 3 - Using minimum word length of 6 the accuracy is: 0.7027

Split 4 - Using minimum word length of 6 the accuracy is: 0.7135

Split 5 - Using minimum word length of 6 the accuracy is: 0.7069

Split 6 - Using minimum word length of 6 the accuracy is: 0.7108

The mean accuracy for this split is 0.7129

Split 1 - Using minimum word length of 7 the accuracy is: 0.7005

Split 2 - Using minimum word length of 7 the accuracy is: 0.7156

Split 3 - Using minimum word length of 7 the accuracy is: 0.6974

Split 4 - Using minimum word length of 7 the accuracy is: 0.7142

Split 5 - Using minimum word length of 7 the accuracy is: 0.6947

Split 6 - Using minimum word length of 7 the accuracy is: 0.7132

The mean accuracy for this split is 0.7059

Split 1 - Using minimum word length of 8 the accuracy is: 0.6664

Split 2 - Using minimum word length of 8 the accuracy is: 0.6604

Split 3 - Using minimum word length of 8 the accuracy is: 0.6621

Split 4 - Using minimum word length of 8 the accuracy is: 0.6664

Split 5 - Using minimum word length of 8 the accuracy is: 0.6779

Split 6 - Using minimum word length of 8 the accuracy is: 0.6702

The mean accuracy for this split is 0.6672

Split 1 - Using minimum word length of 9 the accuracy is: 0.6220

Split 2 - Using minimum word length of 9 the accuracy is: 0.6175

Split 3 - Using minimum word length of 9 the accuracy is: 0.6223

Split 4 - Using minimum word length of 9 the accuracy is: 0.6273

Split 5 - Using minimum word length of 9 the accuracy is: 0.6104

Split 6 - Using minimum word length of 9 the accuracy is: 0.6159

The mean accuracy for this split is 0.6192

Split 1 - Using minimum word length of 10 the accuracy is: 0.5700

Split 2 - Using minimum word length of 10 the accuracy is: 0.5604

Split 3 - Using minimum word length of 10 the accuracy is: 0.5978

Split 4 - Using minimum word length of 10 the accuracy is: 0.5781

Split 5 - Using minimum word length of 10 the accuracy is: 0.5655

Split 6 - Using minimum word length of 10 the accuracy is: 0.5665

The mean accuracy for this split is 0.5730

Highest accuracy minimum word length: 5, with a mean accuracy of: 0.7559

True Positive: 45.0578

False Positive: 4.9442

True Negative: 31.9013

False Negative: 18.0967

Accuracy score: 0.7696