Practical Machine Learning

Unsupervised Learning Algorithms for Text Classification

Ioana Chitic

January 18, 2020

1 Introduction

Sarcasm is defined as “a sharp, bitter, or cutting expression or remark”, or as the use of irony to mock or convey contempt. Given a collection of headlines, can we identify whether the headline is sarcastic or not using unsupervised learning algorithms for text classification?

I decided to explore the answer to this question by using two clustering methods: K-Means, one of the simplest and most popular unsupervised learning algorithms, and DBSCAN (which stands for density-based spatial clustering of applications with noise), a density based algorithm.

I provided a comparison of the two methods, as well as an individual performance analysis.

2 The Dataset and Preprocessing Steps

I used a dataset of headlines collected from various satirical websites, such as The Onion. The dataset consists of the article’s link, the headline, and whether the headline is sarcastic or not.

	article_link	headline	is_sarcastic
0	https://www.huffingtonpost.com/entry/versace-b...	former versace store clerk sues over secret 'b...	0
1	https://www.huffingtonpost.com/entry/roseanne-...	the 'roseanne' revival catches up to our thorn...	0
2	https://local.theonion.com/mom-starting-to-fea...	mom starting to fear son's web series closest ...	1
3	https://politics.theonion.com/boehner-just-wan...	boehner just wants wife to listen, not come up...	1
4	https://www.huffingtonpost.com/entry/jk-rowlin...	j.k. rowling wishes snape happy birthday in th...	0

I only kept the columns that contained data relevant to the task. Afterwards, I processed the headline so it would no longer contain punctuation marks, removed stop words, and stemmed the words. The resulting dataset contained only information relevant to the task:

	headline	is_sarcastic
0	former versac store clerk sue over secret blac...	0
1	the roseann reviv catch up to our thorni polit...	0
2	mom start to fear son web seri closest thing s...	1
3	boehner just want wife to listen not come up w...	1
4	jk rowl wish snape happi birthday in the most ...	0

Since I need numeric data for my machine learning algorithm, I have to convert text into numbers. In order to achieve this, I have to extract featuresd using the TF-IDF

Vectorizer method that attributes to each words a value that is inversely proportional to its frequence. In other words, it convert a collection of documents to a matrix of TF-IDF features

3 K-Means Clustering

Each clustering algorithm in scikit comes in two variants: a class, that implements the fit method to learn the clusters on train data, and a function, that, given train data, returns an array of integer labels corresponding to the different clusters. The k-means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares.

The k-means algorithm divides a set of N samples X into K disjoint clusters C, each described by the mean μ_j of the samples in the cluster. The means are commonly called the cluster "centroids".

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

This algorithm requires the number of clusters to be specified. In our case, that number is two (corresponding to a sarcastic or a non-sarcastic headline).

4 DBSCAN Clustering

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex shaped. There are two parameters to the algorithm, minsamples and epsilon, which define formally what we mean when we say dense. Higher minsamples or lower epsilon indicate higher density necessary to form a cluster.

5 Clustering performance evaluation

- Adjusted Rand index: a function that measures the similarity of the true and predicted label assignments, ignoring permutations.
- Mutual Information based scores, like Adjusted Mutual Information: a function that measures the agreement of the the true and predicted assignments, ignoring permutations.
- Homogeneity: each cluster contains only members of a single class.
- Completeness: all members of a given class are assigned to the same cluster.
- V-measure: the harmonic mean of homogeneity and completeness

```
Homogeneity: 0.020
Completeness: 0.059
V-measure: 0.030
Adjusted Rand Index: -0.011
Adjusted Mutual Information: 0.020
```

Figure 1: k-means results

```
Number of clusters: 2
Estimated number of noise points: 26603
Homogeneity: 0.001
Completeness: 0.016
V-measure: 0.001
Adjusted Rand Index: -0.001
Adjusted Mutual Information: 0.001
```

Figure 2: DBSCAN results

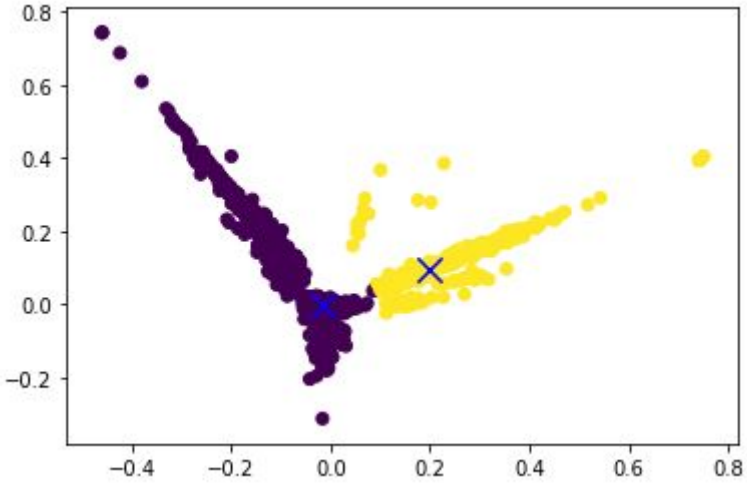

Figure 3: Plot of k-means clusters