Unupervised learning: k-Means, EM, Word2Vec, & WMD

Machine Learning, CS 5333/7333

This project involves developing and experimenting with several unsupervised learning approaches, where no labels are associated with dataset instances but we are interested in detecting underlying consistent patterns in the data, e.g., data sources. There are two types of data sets you will work with in this project: (a) datasets where all attributes are continuous valued, (b) text datasets.

Datasets with continuous Features

You will implement and compare the **k-Means** [2] and the **Expectation Maximization (EM)** algorithms to identify clusters in the following input data sets:

Gaussian sources: You are provided code that will sample data from S Gaussian distributions[1]. The input dataset consists of a set of SN data points, where N data instances are sampled from each of the S input Gaussian distributions. Your report must present an analysis of the experimental results, averaged over 50 runs for each experiment, for the following set of scenarios with the Gaussian input sources:

1. # of points/source, N =1000; standard deviation, σ=1, $S \in \{3, 5, 10\}$; spacing between distributions $\in \{0.5\sigma, \sigma, 1.5\sigma, 2\sigma\}$. For only $S = 3$, vary the number of clusters, C, in the k-Means or EM algorithm from 2, 3, 6, 8. For all other data sets, use $C = S$.
2. $N = 1000$, $S = 5$, Spacing = 3, $\sigma \in \{1, 2, 3\}$.
3. N =1000, $S = 5$, Spacing = 1.25, σ sampled from the uniform distribution over [0.75, 2].
4. $S = 5$, Spacing = 1.25, $\sigma = 0.75$, $N \in \{100, 1000, 5000\}$.

For this dataset, you know the source of each data instance and hence can calculate the following metric: the likelihood that two points chosen to be in the same cluster by a clustering algorithm was generated by the same source distribution, i.e., the fraction of pairs belonging to same output cluster that were generated from the same source. You should also compare the mean and standard deviations of the clusters produced by k-Means as well as the Gaussian parameter distributions estimated by EM, with that of the actual Gaussian source distribution parameters.

Real-life datasets: You are required to work on two real-life data sets:

k-means Clustering for image compression: Use RGB values of a few (4-5) images of interest to you, e.g., portraits, famous paintings, popular physical landmarks, iconic objects. You are required only to use the k-Means clustering algorithm for this data. Report on results (show images) obtained from experiments choosing $k \in \{3, 5, 10\}$ and comment on the quality of compession referring to the original image.

Dataset of your choice: Use a dataset, either from other sources or based on your own work/research containing data collected from multiple similar or related sources, e.g., prices of a given itme over time from various online vendors, on-time arrival statistics of different airlines or at different airports, consumer satisfaction ratings of similar products from different brands, etc. Use both k-Means and EM algorithm for this dataset and compare their performance using (a) Sum of Squared Errors, (b) Silhouette Index and (c) at least one other index of your choice from [1].

[1] The archive provided contains Java code to generate the artificial data sets. If you need assistance to produce the actual data sets from code provide, please bring it up in class.

Text datasets

To use the *k-Means* clustering algorithm on text datasets, you will first use the `Word2Vec` approach [4] to develop real-valued vector representations of individual words in the vocabulary. Subsequently, to find distances between two documents, you will use the `Word Mover's Distance` metric [3].

You are required to use the following text datasets for experimentation on this project:

Dataset used for the Naive Bayes project: You should use the dataset you used for supervised classification with the Naive Bayes algorithm. However, you will use the labels not during the clustering process but later for evaluating the quality of the clusters formed.

Text data set with known biases: Gather a collection of news article headlines from news sources with different bias (liberal vs conservative, international versus domestic, etc.). Again, you will use the source identities not during classification but for evaluation.

For both of text datasets, report the clustering accuracy as the fraction of data instance pairs belonging to same output cluster that were retrieved from the same source. Your report should analyze the similarlity between representation of synonyms, antonyms and any other interesting word patterns that Word2Vec produced (such as examples discussed in [4]). Also, provide examples of distances calculated between equivalent and dissmilar sentences as calculated by WMD (such as the example in [3]).

Grading

The approximate breakup of grades for this project are as following:

Implementation of k-Means and EM	30%
Appropriate use of Word2Vec and WMD	15%
Dataset Collection and Processing	15%
Experimentation	15%
Report (analysis and writing style/clarity)	25%
	—-
Total	100%

References

[1] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, JesúS M PéRez, and IñIgo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[2] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[3] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.