Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 3

Due: February 15, 2023(100 points)

YOUR NAME
YOUR GIT USERNAME
YOUR E-MAIL

Multisource Joins

News articles are commonly aggregated from multiple sites and companies. The landscape of news has been evolving ever since social media has amplified its effects. In politics, Congress has explored the topic of bias with the diversity of news sources. That is, news articles may cover news stories with differing perspectives and language.

The data that we will be using today comes from Kaggle, and it is available here. There are two CSV files that we wish to join in this week's homework: `data/id_titles.csv` and `data/id_publishers.csv`.

As the name suggests, there is publishing data associated with articles and there is title and description information associated with the same articles. Each table has many instances, and each instance for both tables have an associated ID, where it is possible to join the two data sources.

In this particular case, there is some missing information in the join. Your task is as follows.

1. Write out a file that has all the data columns, but where the rows are only those articles where there are titles but no publisher information. Call it titles_no_publishers.txt. In your PDF writeup, include the first 10 rows ordered by ID. This table should look something like the below (ignore the values):

ID	STORY	TITLE	PUBLISHER	CATEGORY	HOSTNAME	URL	TIMESTAMP
100126	dM3BF5lflKhsL6MQ-...	null	TheDay.com	m	www.theday.com	http://www.theday...	1397245711691
100152	dM3BF5lflKhsL6MQ-...	null	HealthLeaders Media	m	www.healthleaders...	http://www.health...	1397245718290
10021	dtBNhkt0YyqHCuM_A...	null	Android Headlines...	t	www.androidheadli...	http://www.androi...	1394714719418
10026	dtBNhkt0YyqHCuM_A...	null	The Herald | Her...	t	www.heraldonline.com	http://www.herald...	1394714720435
100374	dFxU4YSThH_gU7MT9...	null	thejournal.ie	m	www.thejournal.ie	http://www.thejou...	1397246468342
100444	dfp-Hn8YgXYtiKMx9...	null	Daily Mail	m	www.dailymail.co.uk	http://www.dailym...	1397247313815
10046	dtBNhkt0YyqHCuM_A...	null	Computerworld	t	www.computerworld...	http://www.comput...	1394714725232
100471	d0KyvrpUXPQ3XmM2h...	null	Indianapolis Reco...	m	www.indianapolisr...	http://www.indian...	1397247386697
100571	dBU-y8mnlizhV4Mzv...	null	Motley Fool	m	www.fool.com	http://www.fool.c...	1397247496216
100785	dou7Qef9Jcn7_IM4Q...	null	Today's Medical D...	m	www.onlinetmd.com	http://www.online...	1397248500577

2. Write out a file that has all the data columns, but where the rows are only those articles where there are publishers but not title information. Call it titles_no_publishers.txt. In your PDF writeup, include the first 10 rows ordered by ID. That table should look something like the below (ignore the values):

ID	STORY	TITLE	PUBLISHER	CATEGORY	HOSTNAME	URL	TIMESTAMP
100068	dJ_k5DjBr5MzK0MHf...	Networks: Kathlee...	null	null	null	null	null
100176	dM3BF5lflKhsL6MQ-...	Medicare data giv...	null	null	null	null	null
100192	dM3BF5lflKhsL6MQ-...	Medicare Records ...	null	null	null	null	null
100422	duBSqD7s8phcPsMQK...	Sales get leaner ...	null	null	null	null	null
100442	dfp-Hn8YgXYtiKMx9...	More than 100 pas...	null	null	null	null	null
100570	dBU-y8mnlizhV4Mzv...	Today's Pre-Marke...	null	null	null	null	null
100653	dwnBgdLk-3bzGBMNi...	Aid workers back ...	null	null	null	null	null
100716	dwnBgdLk-3bzGBMNi...	WHO says West Afr...	null	null	null	null	null
100850	dk_vhtrqQFe_dsMiu...	Flu Drugs Tamiflu...	null	null	null	null	null
100939	dk_vhtrqQFe_dsMiu...	Study Questions O...	null	null	null	null	null
100969	dk_vhtrqQFe_dsMiu...	Tamiflu use calle...	null	null	null	null	null
101119	dDtTmiUm0P1qeMMK8...	US close: Sell-of...	null	null	null	null	null
101301	d4p273oepCNzWtMV5...	Can Family Dollar...	null	null	null	null	null
101330	dhpby_46Ae5iB8ME-...	A Turbulent Week ...	null	null	null	null	null
10152	dOQvzWTEFn4NkVM9c...	T. rex's 'pygmy' ...	null	null	null	null	null
101704	dq4CkE5dd_NRkmMCB...	Ron Agostini: Col...	null	null	null	null	null
101839	dSAALz3Yg1Ijh5MZV...	Fitch: JPMorgan 1...	null	null	null	null	null
10191	dA0ddnisozIS59MZa...	Earth has a secre...	null	null	null	null	null
101912	dJVPX-uN99u_nuMNq...	GGG-GAME CHANGER:...	null	null	null	null	null

3. Explore the data further, and identify potential problems that could arise if we were to further analyze the data (e.g., apply a machine learning algorithm). Is there still missing data? That is to say, do all the columns have the correct data? What could have gone wrong in the data creation step? (You needn't code anything, but conceptually describe any issues you see and how you would remedy it.)

K-Means

The normalized automobile distributor timing speed and ignition coil gaps supplied are from production F-150 trucks over the years of 1996, 1999, 2006, 2015, and 2022. We have stripped out the labels for the five years of data. Each sample in the dataset is two-dimensional, i.e. $\mathbf{x}_i \in \mathbb{R}^2$, and there are $N = 5000$ instances in the data.

Vanilla k-Means

In this part of the homework, we'll take a look at how we can identify patterns in this data despite not having the labels. We'll start with the simplest approach, the k-Means unsupervised clustering algorithm.

4. Implement a simple k-means algorithm in Python on Colab with the following initialization:

$$\mathbf{x}_1 = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} -10 \\ -10 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} -3 \\ -3 \end{pmatrix}, \tag{0.1}$$

   You need only 100 iterations, maximum, and your algorithm should run very quickly to get the results. In order to maintain consistency between submissions, use a random seed of 27. You can do this with

```
>> numpy.random.seed(seed=27)
```

5. Scatter the results in two dimensions with different clusters as different colors. You can use **matplotlib**'s **pyplot** functionality:

```
>> import matplotlib.pyplot as plt
>> plt.scatter(<YOUR CODE HERE>)
```

6. You will notice that in the above, there are only five initialization clusters. Why is $k = 5$ a logical choice for this dataset? After plotting your resulting clusters and. What do you notice?

With Production Information

Very often, it is possible to obtain additional information about the collected data. This sometimes allows us to define a new mathematical operators (including distances). In this part of the homework, we'll look at how to use this information to improve our modeling with an understanding of how two features in each sample are related.

A common distance metric is the *Mahalanobis Distance* with a specialized covariance.

$$d(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T P^{-1} (\mathbf{x} - \mathbf{y}) \tag{0.2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two points of dimensionality m (2 in this case), and $d(\mathbf{x}, \mathbf{y})$ is the distance between them. In the case of the F150 engine components, P is a known relationship through Ford's quality control analysis each year, where it is numerically shown as below:

$$P = \begin{pmatrix} 10 & 0.5 \\ -10 & 0.25 \end{pmatrix} \tag{0.3}$$

7. Implement a specialized k-means with the above Mahalanobis Distance. Scatter the results with the different clusters as different colors. What do you notice? You may want to pre-compute P^{-1} so that you aren't calculating an inverse every single loop of the the k-Means algorithm.

8. Calculate and print out the *first* principle component of the aggregate data.

9. Calculate and print out the *first* principle components of *each cluster*. Are they the same as the aggregate data? Are they the same as each other?