CS6220 Unsupervised Data Mining

HW1 Data Features, Similarity, KNN

Make sure you check the syllabus for the due date. Please use the notations adopted in class, even if the problem is stated in the book using a different notation.

We are not looking for very long answers (if you find yourself writing more than one or two pages of typed text per problem, you are probably on the wrong track). Try to be concise; also keep in mind that good ideas and explanations matter more than exact details.

Submit all code files Dropbox (create folder HW1 or similar name). Results can be pdf or txt files, including plots/tabels if any.

"Paper" exercises: submit using Dropbox as pdf, either typed or scanned handwritten.

DATATSET **: Kosarak : click-stream data of a hungarian on-line news portal**

DATATSET **: Aminer : public citation dataset**

DATATSET **: 20 NewsGroups : news articles**

DATATSET **: MNIST : digit images**

https://en.wikipedia.org/wiki/MNIST_database
http://yann.lecun.com/exdb/mnist/

PROBLEM 1: Aminer : basic dataset analysis

This is a large dataset (about 2 million publications – it takes about a minute just to parse!). While your notebook must successfully work on the entire dataset, you may find it useful to work on a subset while getting your code to work.

- A. Compute the number of distinct authors, publication venues, publications, and citations/references
- B. Are these numbers likely to be accurate? As an example look up all the publications venue names associated with the conference “Principles and Practice of Knowledge Discovery in Databases”13 – what do you notice?
- C. For each author, construct the list of publications. Plot a histogram of the number of publications per author (use a logarithmic scale on the y axis)
- D. Calculate the mean and standard deviation of the number of publications per author. Also calculate the Q1 (1st quartile14), Q2 (2nd quartile, or median) and Q3 (3rd quartile) values. Compare the median to the mean and explain the difference between the two values based on the standard deviation and the 1st and 3rd quartiles.
- E. Now plot a histogram of the number of publications per venue, as well as calculate the mean, standard deviation, median, Q1, and Q3 values. What is the venue with the largest number of publications in the dataset?
- F. Plot a histogram of the number of references (number of publications a publication refers to) and citations (number of publications referring to a publication) per publication. What is the publication with the largest number of references? What is the publication with the largest number of citations? Do these make sense?

- G. Calculate the so called “impact” factor for each venue. To do so, calculate the total number of citations for the publications in the venue, and then divide this number by the number of publications for the venue. Plot a histogram of the results
- H. What is the venue with the highest apparent impact factor? Do you believe this number? (http://mdanderson.libanswers.com/faq/26159)
- I. Now repeat the calculation from item b., but restrict the calculation to venues with at least 10 publications. How does your histogram change? List the citation counts for all publications from the venue with the highest impact factor. How does the impact factor (mean number of citations) compare to the median number of citations?
- J. Finally, construct a list of publications for each publication year. Use this list to plot the average number of references and average number of citations per publication as a function of time. Explain the differences you see in the trends.

https://en.wikipedia.org/wiki/IPython#Notebook
https://aminer.org
https://en.wikipedia.org/wiki/ECML_PKDD
https://en.wikipedia.org/wiki/Quartile

PROBLEM 2 : Kosarak Association Rules

Your task is to take a dataset of nearly one million clicks on a news site16 and use the Weka Explorer to identify interesting association rules. Ordinarily this would be a point-and-click task; however, the input data format is a list of transactions (each line in the file includes a list of anonymized news item id’s), whereas Weka requires a tabular format. Specifically, each distinct news item id should be represented via a column/attribute, and each row/instance should be a sequence of binary values, indicating whether or not the user visited the corresponding news item.

- A. Write a Python program which takes as its argument5 the path to a text file of data (assumed to be in the itemset format above) and produces as output to the console a sparse ARFF file.
- B. Use your program to convert the kosarak.dat file to a sparse kosarak.arff. About how long did it take to run?
- C. Load the resulting file into Weka (as described above; you should have 41,270 attributes and 990, 002 instances). About how long did it take to load this file?
- D. Use Weka’s FP-Growth implementation to find rules that have support count of at least 49, 500 and confidence of at least 99% – record your rules (there should be 2).
- E. Run the algorithm at least 5 times. Then look to the log and record how much time each took. How does the average time compare to the time necessary to convert the dataset and then load into Weka?

PROBLEM 3 MNIST, 20 NG . Parse, normalize features, Compute pairwaise similarity matrices

The parsers are very different for the two datasets (text vs images) but you are allwed to use a library/package to do so. These being very very popular research datasets, it should be easy to find appropriate parsers. You can try to normalize each column/feature separately with wither one of the following ideas. Do not normalize labels. When normalizing a column, make sure to normalize its values across all datapoints (train,test,validation, etc) for consistency

Typical options for feature values (nomalization optional):

- 20NG text row normalization TF(term,doc) / DL (doc). For text is critical to maintain a sparse format due to large number of columns; make sure any cvalue transformation retains the 0 values.
- MNIST : since these images are black and white (and some gray) the pixel values are already in a preformatted range [0-255]. They may not require normalization, but perhaps its easier to get the values to have 0 mean instead of 128 mean. Depending on what similarity/distance measure you use, computation of similarity might be easy but the size of the similarity matrix might present a challenge.
- Shift-and-scale normalization: substract the minimum, then divide by new maximum. Now all values are between 0-1
- Zero mean, unit variance : substract the mean, divide by the appropriate value to get variance=1

Options for distance/similarity. You are encouraged to use a package library available in Matlab/Java/Python/R to compute the pairwise similarity/distance matrix

- cosine or simple dot product (required)
- euclidian distance (required)
- editing distance (optional)
- jaccard similarity(optional)
- Manhattan distance(optional)

https://en.wikipedia.org/wiki/Feature_scaling
https://en.wikipedia.org/wiki/Distance_matrix
https://en.wikipedia.org/wiki/Distance
https://en.wikipedia.org/wiki/Category:Similarity_and_distance_measures
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.8446&rep=rep1&type=pdf
http://dataaspirant.com/2015/04/11/five-most-popular-similarity-measures-implementation-in-python/

PROBLEM 4: MNIST, 20 NG : Train and test KNN classification (supervised)

Some datasets might come organized into train/test in which case we respect that. Other datasets come without this organization in which case we randomly ("random" here is very important, data must be shuffled) pick about 80% of data as training , 10% as validation (also used in training) and 10% as testing data (completely unavailable to training)
For each of the two datasets, no in matrix format and with pairwise similarity computed, train and test KNN classification. Report both training performance and testing performance. You are required to implement KNN yourself, but can youse suport libraries and datastructures.