Your Name: Yash Hatekar

1. Transaction data (with each transaction represented as a set of items) can also be represented as a matrix (transaction-item matrix, similar to the document-term matrix we showed in the class). List one advantage and one disadvantage of using such a matrix. Also discuss why a transaction-item matrix is an example of a data set that has asymmetric discrete features.

Answer –

Transaction data when represented as a matrix is easy to manipulate and transform using standard matrix operations and each object can be thought of as vectors in a multidimensional space, where each dimension represents a distinct attribute about the object. The disadvantage of such representation would be the size it takes to store all this data.

Transaction data consists of collection of sets of items. For example – a grocery shopping list, for a customer there will be some collections of items. When such data is converted into its matrix equivalent form creates discrete data points for each customer. And since any item only has information about whether it was purchased or not, this makes them discrete asymmetric features.

2. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there many be some ambiguity.
(a) Brightness as measured by a light meter.
(b) Brightness as measured by people's judgments.
(c) Number of customers in a grocery store.
(d) Letter grades (A, B, C, D and F).
(e) Distance from the Monroe County Courthouse.

Answer -

a) Brightness measured by a light meter would give an accurate reading so it should be **continuous** and **quantitative** and **ratio.**

b) Brightness as measured by people's judgments would not be very accurate so it should be **discrete** and **qualitative** and **ordinal.**

c) Number of customers in a grocery store would be **discrete** and **quantitative** and **ratio.**

d) Letter grades should be **discrete** and **qualitative** and **ordinal.**

e) Distance from the Monroe County Courthouse should be **continuous** and **quantitative** and **ratio.**

3. Distinguish between noise and outliers. Be sure to consider the following questions.

• Is noise ever interesting or desirable? Outliers?

• Can noise objects be outliers?

• Are noise objects always outliers?

• Are outliers always noise objects?

• Can noise make a typical value into an unusual one, or vice versa?

Answer -

Noise values are modified original values. Noise can be readings which are incorrect due to errors. Example – Voice changes during calls due to poor connection.

Outliers are original readings, but which do not follow the pattern of other readings in the same group.

Example – In a class of students with heights in the range of 180-195 cm if there is someone with a height 120 cm then that student with 120cm height becomes an outlier.

a) Is noise ever interesting or desirable? Outliers?

Noise is erroneous data which is not desired and should be removed. But outliers are interesting and sometimes desired data points like in case of credit card fraud detection.

b) Can noise objects be outliers?

Yes, noise objects can be outliers.

c) Are noise objects always outliers?

No, noise objects are not always outliers. Since noise is a modified version of the original value, they can be a mix of any values.

d) Are outliers always noise objects?

No, since outliers are original data points, we cannot classify them as noise every time.

e) Can noise make a typical value into an unusual one, or vice versa?

Yes, noise modifies the original value and transforms it into a new value which can be an unusual one or a typical value.

5. Learn about bitcoin using Google Trends. Write a brief summary including four highlights about what you have learned.
Answer -

- The search term 'bitcoin' gained initial popularity during early 2014.
- The next big explosion was during December 2017 – January 2018, this might be due to people realizing the uses of bitcoin. The explosion in digital banking might also be one of the reasons why the term bitcoin became popular during this time.
- We observe a spike during late 2021 as at that time a bitcoin was worth more than $60,000.
- The bitcoin search trend and the actual price of bitcoin follow very different trends.
- If we compare the trend of the term 'bitcoin' vs 'bitcoin price' we can observe that whenever 'bitcoin' search spiked, there was a spike in the search query 'bitcoin price prediction.' This tells us that people were trying to learn and use prediction models to find more information about the trend of the price of the bitcoin.
- When we look at the trend on a worldwide scale, we can see that developed or developing countries which had started digitizing banking solutions were observing spikes in 'bitcoin' search term.

6. Write a summary for this paper: Fair Labeled Clustering (KDD 2022). It is a math-heavy research paper and it is perfectly fine if you don't understand some parts of the paper. Focus on the problem it tries to solve.

The paper talks about algorithmic fairness in unsupervised learning clustering algorithms. The authors describe how traditional clustering algorithms inspect the center of each cluster and decide an outcome(label) for each cluster. Typically, clustering is applied to obtain *k* many clusters and the centers of these clusters are put into focus. Naturally, more than one cluster can be assigned the same label. In such cases strict approach of group fairness in each cluster causes large degradation in clustering quality. For such scenarios they have provided examples of how the traditional clustering algorithms would yield sub-optimal solutions like in cases of job screening and targeted advertising. Thus, the authors propose a less stringent, less costly and a different approach which tries to satisfy the fairness criteria for an existing group of clustering algorithms.

The authors have provided the solution for two types of scenarios **labeled clustering with assigned labels** (LCAL) where the center labels are decided based on their position and **labeled clustering with unassigned labels** (LCUL) where we are free to select the center labels subject to some constraints. Finally, the authors claim that their solution provides fairness at a lower cost than fair clustering and that it can scale for larger datasets. For LCAL (assigned labels) the max distance function between a point and its assigned center is calculated by putting constraints which are to be satisfied in fair labeled clustering. The formulas are mentioned in section 3.1.1b .They have introduced new additional constraints in fair clustering from the fact

that since positive and negative outcomes could be associated with different labels, an upper bound is set on the total number of points assigned to a positive label. Similarly, a lower bound may be set to avoid trivial solutions in case of negative outcomes. For LCUL (unassigned labels) since the labels for centers are unknown in such scenarios, a new constraint is introduced which bounds the subset of centers that have been assigned a label by an additional optimization function. The formulas for the same are mentioned in section 3.2.2d.

The authors have provided theoretical proofs for LCAL being polynomial time solvable (4.1 theorem 1) and efficient algorithms (4.2 theorem 2) for LCAL for the two label case problems. For LCUL theorem 3 says that LCUL is NP-hard problem with two labels and two colors and if any one of the constraints (2b ,2c or 2d) is dropped. Theorem 4 explains how LCUL problem solvable in polynomial time when fixed parameters are used for constant number of labels.Algorithms for both the methods are provided in their respective proofs section. Theorem 5 explains that the LCUL problem is solvable in polynomial time under constraint (2b or 2c) alone if number of labels is super constant but is still NP-hard under constraint (2d). Theorem 6 explains that for special case of LCUL called as color and label proportional case (CLP) is a NP-hard problem and theorem 7 states that algorithm 2 (randomized LCUL algorithm) gives an optimal clustering when all the constraints are satisfied and constraint 4d being satisfied deterministically at a violation at most 1.

For experimentation the authors have used Python3 using the NumPy library and functions from the Scikit-learn library. They have used two datasets called the **Adult** and **CreditCard** from the UCI repository, containing 32,561 and 30,000 points respectively. For the group membership attribute, they have used race for **Adult** which takes on 5 possible values (5 colors) and marriage for **CreditCard** which takes on 4 possible values (4 colors). For the **Adult** dataset they have used the numeric entries of the dataset (age, final-weight, education, capital gain, and hours worked per week) as coordinates in the space. Whereas for the **CreditCard** dataset they have used age and 12 other financial entries as coordinates. LCAL provides smaller PoF than fair clustering algorithm in case of both the datasets. LCUL also shows a lower PoF value versus Nearest Center with Random Assignment (NCRA) and Fair Clustering (FC) algorithms. For the scalability of the algorithm the authors have taken Census1990 dataset which contains 2,458,285 points. They noticed that for 500,000 values their approach takes less than 90 seconds whereas it would take fair clustering algorithm around 30 minutes to solve.

In my opinion the paper is thorough and provides good insight with various examples and approaches on how the proposed solution is efficient, scalable and fair. With the help of the mentioned theorems and algorithms we can try to implement their approach in our use cases while understanding the math behind it.