CS5990: Homework #1

Due on September 18, 2022

Prof. Ericsson

Nythi (Ned) Udomkesmalee

Bronco ID: |0|1|3|5|1|4|0|0|7|
Last Name: Udomkesmalee
First Name: Nythi

Problem 1

1. [12 points – 2 points each] Answer whether each of the following activities is a data mining task. **Justify** your answer.

   a. Dividing the customers of a company according to their gender.
   b. Monitoring seismic waves for earthquake activities.
   c. Computing the total sales of a company.
   d. Predicting the outcomes of tossing a (fair) pair of dice.
   e. Predicting the future stock price of a company using historical records.
   f. Monitoring the heart rate of a patient for abnormalities.

Solution

(a) **Not Data Mining** - No new knowledge is being gained. This is just sorting the data.

(b) **Data Mining** - We can take the data of the seismic waves and figure out if they are dangerous or not. It would be predictive classification problem.

(c) **Not Data Mining** - No new knowledge is being gained. This is simple calculation

(d) **Not Data Mining** - Assuming the die is actually fair, then prediction would not be helpful as the probabilities are known in advanced.

(e) **Data Mining** - Since the prices of stocks aren't deterministic it may be helpful to try and use past data and other analytics to create a model to describe reality as best as possible. This is prediction problem.

(f) **Data Mining** - You can create a model based on previous data to predict that a heart issue is happening before a human in the loop can diagnose it. This is an anomaly detection problem, or a classification problem.

Bronco ID: | 0 | 1 | 3 | 5 | 1 | 4 | 0 | 0 | 7 |
Last Name: Udomkesmalee
First Name: Nythi

Problem 2

2. [10 points – 2 points each] Classify the following attributes as **discrete**, or **continuous**. Also classify them as **nominal**, **ordinal**, **interval**, or **ratio**.

   a) Brightness as measured by a light meter
   b) Brightness as measured by people's judgments.
   c) Density of a substance in grams per cubic meter
   d) Time of each day in the meaning of a 12-hour clock
   e) CPP bronco IDs

Solution

(a) **Continuous, Ratio**

(b) **Discrete, Ordinal**

(c) **Continuous, Ratio**

(d) **Discrete, Interval**

(e) **Discrete, Nominal**

Bronco ID: 013514007
Last Name: Udomkesmalee
First Name: Nythi

Problem 3

3. [9 points] Explain where **visualization**, **normalization**, and **machine learning** techniques are applied during the 3 main phases of the KDD process shown below. Justify the importance of these techniques to produce useful information at the end of the process.

Solution

(a) Data Preprocessing → Normalization - Normalization is helpful to the entire process by transforming the scales of all the input data gathered into a common scale. By doing that, this helps to prevent any particular feature from skewing the results because their scale is on a different order of magnitude from other features.

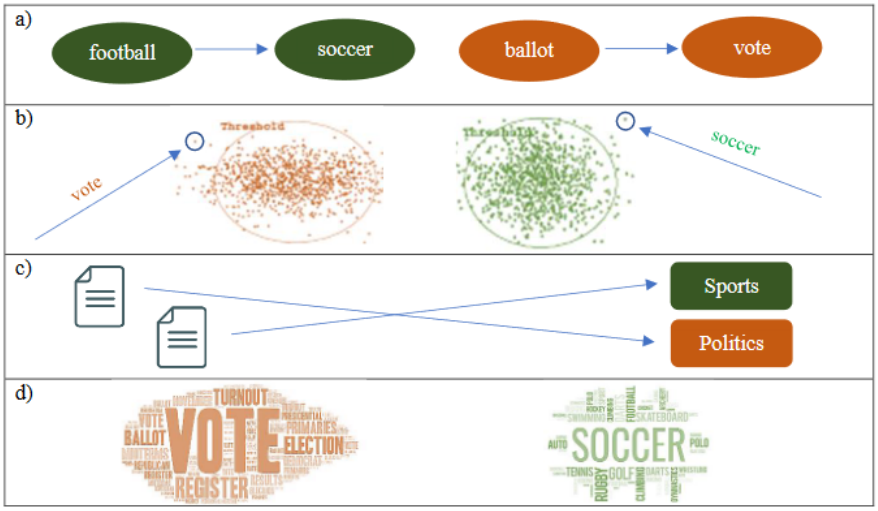
(b) Data Mining → Machine Learning - Machine Learning is a tool that is helpful in creating a model that will infer patterns from data. This model can be used to classify or predict new data coming in. This method is very important as it can process much more data and come to find patterns much better than a human can.

(c) Post Processing → Visualization - Visualization is essentially translating machine patterns about the data into something humans can understand. Humans are good at seeing obvious patterns and comparisons so graphs, charts, and reports that are created during data visualization can be digested and acted upon by the human.

Bronco ID: 010135104007
Last Name: Udomkesmalee
First Name: Nythi

Problem 4

4. [8 points] Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving examples of how techniques, such as **clustering**, **classification**, **association rule mining**, and **anomaly detection** can be applied. Requirement: use the illustrations provided below as a guide to elaborate your answer. You will need to figure out which data mining technique corresponds to which illustration and what **kind of results is being provided** to the final user (**be specific** as much as you can).



Solution

Solution

(a) **Association Rule Mining** - This technique is helpful in finding patterns of associated items. In the example we have associating words in a search to another word. This may be useful for auto-complete features, or for showing what type of content to show the user. Maybe also for which ads to show.

(b) **Anomaly Detection** - This technique is helpful for finding outliers in the general usage seen. In this example finding anomalies that fall outside of a specific clusters may be used to prune the search engine of results that are wild. Also could use outlier results to detect fraudulent sites or potentially harmful site.

(c) **Classification** - This method is the most straightforward. In the example you want to label your documents with a category. This is helpful in determining which type of content to show to the user.

(d) **Clustering** - This techniques is helpful to figure out how similiar certain attributes are to a category. In the example you have words that can be considered part of a cluster where the size of the word is the frequency/likelyhood that the word that shows up belongs to the cluster. This is helpful in categorizing the search words into cluster and then showing specific contents based on that classification.

Bronco ID: |0|1|3|5|1|4|0|0|7|
Last Name: Udomkesmalee
First Name: Nythi

Problem 5

5. [12 points – 2 points each] Analyze the dataset below and answer the proposed questions:

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

a. What is the **most likely task** that data scientists are trying to accomplish?

b. **In general**, what is a feature and how would you **exemplify** it with **this data**?

c. **In general**, what is a feature value and how would you **exemplify** it with **this data**?

d. **In general**, what is dimensionality and how would you **exemplify** it with **this data**?

e. **In general**, what is an instance and how would you **exemplify** it with **this data**?

f. **In general**, what is a class and how would you **exemplify** it with **this data**?

Solution

a. They are trying to predict if people are cheating their taxes or not.

b. A ***Feature*** is a parameter or descriptive element of the data. So the columns are the features such as Refund, Marital Sataus, and Taxible Income

c. A ***Feature Value*** is an instance value of a feature. So in the example having your Refund be "Yes" or "No" is a feature value. Similarly, having Taxible income be "125k" is also a feature value.

d. ***Dimensionality*** is the number of attributes in the data. In this example the dimensionality is 4, since there are 4 input features to take into account. I guess you could consider it 5 features if including the class for training.

e. An ***Instance*** is a singular data point where all features have a feature value. In this example each row is considered an instance.

f. A **Class** is feature with a discrete values that each instance of data can be categorized as. In the example "Cheat" is the class.

Bronco ID: |0|1|3|5|1|4|0|0|7|
Last Name: Udomkesmalee
First Name: Nythi

Problem 6

6. [6 points] Consider the dialog below between a statistician and a data miner:

Statistician: "There are so many data issues. Did you note that fields 2 and 3 are basically the same?"
Data Miner: "I guess a missed that."

Explain why the statistician concluded that fields 2 and 3 below are basically the same.

012	233.8	33.4	0	10.7
020	119.7	17.1	2	210.1
027	168.0	24.0	0	427.6

Solution

Column 2 is just 7 * Column 3. So since you can find a linear transformation between the two columns, they are representing the same underlying information and thus no new knowledge is being gained by including it.

Bronco ID: |0|1|3|5|1|4|0|0|7|
Last Name: Udomkesmalee
First Name: Nythi

Problem 7

7. [8 points] For each of the following vectors, x and y, calculate the indicated similarity or the distance measures. Show your **math** for full mark.

(a) x = (1 1 0 0 0), y = (0 0 0 1 1) Jaccard, Cosine, Euclidean, Correlation

(b) x = (0 1 0 1 1), y = (1 0 1 0 0) Jaccard, Cosine, Euclidean, Correlation

Solution

a. i. Jaccard -

$$\frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{2+2+0} = \boxed{0}$$

ii. Cosine -

$$cos(\theta) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\,|\vec{y}|} \Rightarrow$$

$$\frac{\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}}{\sqrt{1^2+1^2}\sqrt{1^2+1^2}} = \frac{0+0+0+0+0}{2} = \boxed{0}$$

iii. Euclidian -

$$d = \sqrt{\sum_{i=0}^{n}(x_i-y_i)^2} = \sqrt{(1-0)^2+(1-0)^2+(0-0)^2+(0-1)^2+(0-1)^2} = \sqrt{4} = \boxed{2}$$

iv. Correlation -

$$\bar{x} = \frac{1+1+0+0+0}{5} = \frac{2}{5}$$

$$\bar{y} = \frac{0+0+0+1+1}{5} = \frac{2}{5}$$

$$n = 5$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x}^2)} = \sqrt{\frac{1}{4}\left[(1-\frac{2}{5})^2+(1-\frac{2}{5})^2+(0-\frac{2}{5})^2+(0-\frac{2}{5})^2+(0-\frac{2}{5})^2\right]} = \sqrt{\frac{3}{10}}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y}^2)} = \sqrt{\frac{1}{4}\left[(0-\frac{2}{5})^2+(0-\frac{2}{5})^2+(0-\frac{2}{5})^2+(1-\frac{2}{5})^2+(1-\frac{2}{5})^2\right]} = \sqrt{\frac{3}{10}}$$

$$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y}) = \frac{1}{4}\left[(\frac{3}{5})(-\frac{2}{5})+(\frac{3}{5})(-\frac{2}{5})+(-\frac{2}{5})(-\frac{2}{5})+(-\frac{2}{5})(\frac{3}{5})+(-\frac{2}{5})(\frac{3}{5})\right] = -\frac{1}{5}$$

Problem 7 continued on next page...

$$corr(x,y) = \frac{s_{xy}}{s_x s_y} = \frac{-\frac{1}{5}}{\sqrt{\frac{3}{10}}\sqrt{\frac{3}{10}}} = \boxed{-\frac{2}{3}}$$

b. i. Jaccard -

$$\frac{f_{11}}{f_{10}+f_{01}+f_{11}} = \frac{0}{2+3+0} = \boxed{0}$$

ii. Cosine -

$$cos(\theta) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\,|\vec{y}|} \Rightarrow$$

$$\frac{\begin{bmatrix} 0 & 1 & 0 & 1 & 1 \end{bmatrix}\begin{bmatrix}1\\0\\1\\0\\0\end{bmatrix}}{\sqrt{1^2+1^2+1^2}\sqrt{1^2+1^2}} = \frac{0+0+0+0+0}{\sqrt{6}} = \boxed{0}$$

iii. Euclidian -

$$d = \sqrt{\sum_{i=0}^{n}(x_i-y_i)^2} = \sqrt{(0-1)^2+(1-0)^2+(0-1)^2+(1-0)^2+(1-0)^2} = \boxed{\sqrt{5}}$$

iv. Correlation -

$$\bar{x} = \frac{0+1+0+1+1}{5} = \frac{3}{5}$$
$$\bar{y} = \frac{1+0+1+0+0}{5} = \frac{2}{5}$$
$$n = 5$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x}^2)} = \sqrt{\frac{1}{4}\left[(0-\frac{3}{5})^2+(1-\frac{3}{5})^2+(0-\frac{3}{5})^2+(1-\frac{3}{5})^2+(1-\frac{3}{5})^2\right]} = \sqrt{\frac{3}{10}}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y}^2)} = \sqrt{\frac{1}{4}\left[(1-\frac{2}{5})^2+(0-\frac{2}{5})^2+(1-\frac{2}{5})^2+(0-\frac{2}{5})^2+(0-\frac{2}{5})^2\right]} = \sqrt{\frac{3}{10}}$$

$$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y}) = \frac{1}{4}\left[(-\frac{3}{5})(\frac{3}{5})+(\frac{2}{5})(-\frac{2}{5})+(-\frac{3}{5})(\frac{3}{5})+(\frac{2}{5})(-\frac{2}{5})+(\frac{2}{5})(-\frac{2}{5})\right] = -\frac{3}{10}$$

$$corr(x,y) = \frac{s_{xy}}{s_x s_y} = \frac{-\frac{3}{10}}{\sqrt{\frac{3}{10}}\sqrt{\frac{3}{10}}} = \boxed{-1}$$

Bronco ID: |0|1|3|5|1|4|0|0|7|
Last Name: Udomkesmalee
First Name: Nythi

Problem 8

8. [8 points] Given vectors u = (2, k) and v = (3, -2), find the value of k such that vectors are

   (a) perpendicular (b) parallel

   Show your **math** for full mark.

Solution

a. So perpendicular implies that the angle between the vectors is 90°. Thus from the dot product formula we have:

$$cos(\theta) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}|\,|\vec{v}|} \Rightarrow$$

$$cos(90°) = 0 = \begin{bmatrix} 2 & k \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} \Rightarrow$$

$$6 - 2k = 0 \Rightarrow$$

$$\boxed{k = 3}$$

b. So parallel implies that the angle between the vectors is 0°. Thus from the dot product formula we have:

$$cos(\theta) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}|\,|\vec{v}|} \Rightarrow$$

$$cos(0°) = 1 = \frac{\begin{bmatrix} 2 & k \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix}}{\sqrt{2^2 + k^2}\sqrt{3^2 + (-2)^2}} \Rightarrow$$

$$\sqrt{4 + k^2}\sqrt{9 + 4} = \begin{bmatrix} 2 & k \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} \Rightarrow$$

$$\sqrt{13}\sqrt{4 + k^2} = 6 - 2k \Rightarrow$$

$$\sqrt{52 + 13k^2} = 6 - 2k \Rightarrow$$

$$52 + 13k^2 = (6 - 2k)^2 \Rightarrow$$

$$52 + 13k^2 = 36 - 24k + 4k^2 \Rightarrow$$

$$9k^2 + 24k + 16 = 0 \Rightarrow$$

$$k = \frac{-24 \pm \sqrt{(24)^2 - 4 * 9 * 16}}{18} = \frac{-24 \pm \sqrt{576 - 576}}{18} \Rightarrow$$

$$\boxed{k = \frac{-24}{18} = -\frac{4}{3}}$$

Bronco ID: |0|1|3|5|1|4|0|0|7|
Last Name: Udomkesmalee
First Name: Nythi

Problem 9

9. [12 points] Given the raw and processed datasets below, **explain** which technique has been used to perform data preprocessing: **feature selection**, **sampling**, **aggregation**, or **standardization**.

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	$10
2	Shoes	0	09/10/22	$15
3	TV	1	09/09/22	$20

a)

Item	Downtown	Date	Price
Electronics	1	09/09/22	$30
Shoes	0	09/10/22	$15

b)

ID	Item	Date	Price
1	Watch	09/09/22	$10
2	Shoes	09/10/22	$15
3	TV	09/09/22	$20

c)

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	0.1
2	Shoes	0	09/10/22	0.15
3	TV	1	09/09/22	0.2

d)

ID	Item	Downtown	Date	Price
1	Watch	1	09/09/22	$10
2	Shoes	0	09/10/22	$15

Solution

a. **Aggregation** - "Watch" and "TV" have been combined (aggregated) into "Electronics" and aggregated the price as well of those sub-units.

b. **Feature Selection** - The boolean feature "Downtown" has been removed as it may not be important to the model.

c. **Standardization** - The scale of "Price" has been mapped down to a range of [0, 1] to limit the effects of scale on training.

d. **Sampling** - The third data point has been removed. So taking a subset of the total amount of data is called sampling.

Bronco ID: 013514007
Last Name: Udomkesmalee
First Name: Nythi

Problem 10

10. [15 points] Complete the python program (similarity.py) that will output the two most similar documents according to their cosine similarity. Add the link to the online repository as your answer.

Important Note: Answers to all questions should be written clearly, concisely, and unmistakably delineated. You may resubmit multiple times until the deadline (the last submission will be considered).

Solution

Code can be found at: https://raw.githubusercontent.com/NinjaNed/CS5990_HW/master/HW1/similarity.py
Repo link: NinjaNed/CS5990_HW/HW1