Assignment 1

CS 412: Introduction to Data Mining (Spring 2023) Instructor: Hanghang Tong

Release date: Jan. 17th, 2023 Due date: Feb. 2nd, 2023

* This assignment will cover the content from Chapters #1 (Introduction) and #2 (Data, Measurements, and Data Preprocessing).
* Feel free to discuss with other members of the class when doing the homework. You should, however, write down your own solution **independently**. **\*Very Important Notes\*: (1) there is a fine line between collaboration and completing the assignment by yourself and (2) aiding others to cheat would have the same consequence as the cheating itself. Please try to keep the solution brief and clear.**
* Please use Piazza first if you have questions about the assignment. Also feel free to send us e-mails and come to office hours.
* The assignment is due at 11:59 PM on the due date. We will be using Canvas for collecting homework assignments. **Please do not hand in a scan of your hand- written solution, only the typed solution (e.g., Microsoft Word, Latex, etc) will be graded**. The datasets for HW1 are in the HW1 folder on Canvas. Contact the TAs if you are having technical difficulties in submitting the assignment. We do **NOT** accept late assignment!
* The assignment should be submitted as a **single** PDF file using the name convention yourNetID HW1.pdf. If you use additional source code for solving problems, you are required to submit them and use the file names to identify the corresponding questions. For instance, ‘yourNetID HW1.py’ refers to the python source code for Problem 1, replace netid with your netid. Compress all the files (PDF and source code files) into one zip file. Submit the compressed file **ONLY**. (If you did not use any source code, submitting the PDF file without compression will be fine)
* For each question, **you will NOT get full credits if you only give out a final result.** Necessary calculation steps are required. If the result is not an integer, round your result to 4 decimal places.

**Problem 1. True or False** (20 points)

Please justify your answers with **at most** 2 sentences.

1. (2 points) Binary attribute is equivalent to discrete attribute.

Binary attribute is an attribute that can take only two possible values, while discrete attribute can take a finite number of values, but not necessarily only two. Therefore, binary attribute is not equivalent to discrete attribute. A discrete attribute can have more than two values while a binary attribute has only two values.

1. (2 points) For a unimodal curve that is not symmetric, according to the empirical formula *mean − mode* = 3(*mean − median*), its median and mode are usually on the same side of its mean.

For a unimodal curve that is not symmetric, according to the empirical formula mean — mode = 3(mean — median), its median and mode are usually on the same side of its mean is not true. This formula is an empirical rule that holds for normal distribution but it's not applicable for all unimodal curve. The position of mode, median and mean depends on the curve shape and distribution.

1. (2 points) For a group of scalars, the median is always smaller than its max and larger than its min.

For a group of scalars, the median is always smaller than its max and larger than its min is not true. The median is the middle value of a dataset when it is ordered, it is larger than or equal to the first quartile and smaller than or equal to the third quartile, it may or may not be smaller than max or larger than min.

1. (2 points) Given two real-valued vectors, if their histogram are exactly the same, these two vectors are the same.

Given two real-valued vectors, if their histogram are exactly the same, these two vectors are the same is not true. Histograms are a representation of the data, they show the frequency distribution of the data, but it doesn't give the whole information about the data, two vectors may have the same histograms but still, they can have different values.

1. (2 points) Given a similarity measure between two objects *A* and *B*, e.g., *S*(*A, B*), we can always define a valid distance measure as *D*(*A, B*) = 1 .

*S*(*A,B*)

Given a similarity measure between two objects A and B, e.g., S(A, B), we cannot always define a valid distance measure as D(A, B) = 1/S(A, B). While it is true that in some cases, a similarity measure can be transformed into a distance measure by subtracting it from 1 or taking its reciprocal, this is not always the case. Similarity and distance measures have different properties and are not interchangeable. It's important to choose the appropriate measure based on the problem and the data.

1. (2 points) Given a vector of real numbers, if a raw number in the vector is smaller than the mean, its corresponding normalized value will be negative after z-score standardization.
2. (2 points) Given any two vectors of the same length, the L2 distance between them is always smaller than or equal to the L1 distance between them.
3. The covariance matrix (of one random vector) has square shape.
4. (2 points) Let *p* and *q* be two different distributions. The KL divergence between *p*

and *q* is the same as the KL divergence between *q* and *p*.

1. (2 points) Principal Component Analysis (PCA) can generate new features.

**Problem 2. Data Measurement** (24 points)

Table [1](#_bookmark0) lists the heights and weights of 12 NBA players [1](#_bookmark1).

Table 1: Heights and Weights of 12 NBA players.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Player No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Heights (in cm) | 193 | 198 | 190 | 228 | 210 | 198 | 208 | 203 | 216 | 175 | 183 | 195 |
| Weights (in kg) | 96 | 98 | 86 | 140 | 116 | 92 | 117 | 113 | 136 | 81 | 79 | 92 |

Compute the following statistical properties for **both heights and weights**.

(6 points) Mean, mode and median.

198

92

(6 points) First quartile, third quartile and inter-quartile range.

For heights:

First quartile = (190 + 193) / 2 = 191.5

Third quartile = (208 + 120) / 2 = 209

Inter-quartile range = 209 - 191.5 = 17.5

For weights:

First quartile = (86 + 92) / 2 = 89

Third quartile = (116 + 117) / 2 = 116.5

Inter-quartile range = 116.5 - 89 = 27.5

(6 points) Suppose we use z-score normalization to normalize the data for all these NBA players. After the normalization, what are the normalized heights and weights? What are the sample variance and sample standard deviation of the normalized data.

from scipy.stats import zscore

import pandas as pd

# Create a sample df

df1 = pd.DataFrame({'heights': [175, 183, 190, 193, 195, 198, 198, 203, 208, 210, 216, 228]})

# Calculate the zscores and drop zscores into new column

df1['heights\_zscore'] = zscore(df1['heights'])

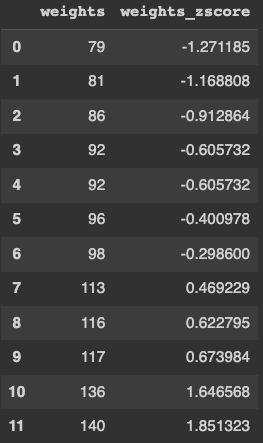
# Create a sample df

df2 = pd.DataFrame({'weights': [79, 81, 86, 92, 92, 96, 98, 113, 116, 117, 136, 140]})

# Calculate the zscores and drop zscores into new column

df2['weights\_zscore'] = zscore(df2['weights'])

display(df1, df2)

A picture containing table

Description automatically generated

import statistics

heights\_zero\_norm\_var = statistics.variance(heights\_zero\_norm)

weights\_zero\_norm\_var = statistics.variance(weights\_zero\_norm)

print(heights\_zero\_norm\_var, weights\_zero\_norm\_var)

Sample variance of heights = 1.090909090909091

Sample variance of weights = 1.0909090909090908

import statistics

heights\_zero\_norm\_stdev = statistics.stdev(heights\_zero\_norm)

weights\_zero\_norm\_stdev = statistics.stdev(weights\_zero\_norm)

print(heights\_zero\_norm\_stdev, weights\_zero\_norm\_stdev)

Sample standard deviation of heights = 1.044465935734187

Sample standard deviation of weights = 1.044465935734187

(6 points) Suppose we use min-max normalization to normalize the data for all these NBA players. After the normalization, what are the normalized heights and weights? What are the population variance and population standard deviation of the normalized data.

import pandas as pd

from sklearn.preprocessing import MinMaxScaler

data = {'heights':[175, 183, 190, 193, 195, 198, 198, 203, 208, 210, 216, 228],

'weights':[79, 81, 86, 92, 92, 96, 98, 113, 116, 117, 136, 140]}

df = pd.DataFrame(data)

scaler = MinMaxScaler()

normalized\_data = scaler.fit\_transform(df)

normalized\_df = pd.DataFrame(normalized\_data, columns=df.columns)

display(normalized\_df)

Graphical user interface

Description automatically generated

heights\_minmax\_norm = [0.000000, 0.150943, 0.283019, 0.339623, 0.377358, 0.433962,

0.433962, 0.528302, 0.622642, 0.660377, 0.773585, 1.000000]

weights\_minmax\_norm = [0.000000, 0.032787, 0.114754, 0.213115, 0.213115, 0.278689,

0.311475, 0.557377, 0.606557, 0.622951, 0.934426, 1.000000]

heights\_minmax\_norm\_var = statistics.variance(heights\_minmax\_norm)

weights\_minmax\_norm\_var = statistics.variance(weights\_minmax\_norm)

print(heights\_minmax\_norm\_var, weights\_minmax\_norm\_var)

Sample variance of heights = 0.07470309383808334

Sample variance of weights = 0.11188743159633334

heights\_minmax\_norm\_stdev = statistics.stdev(heights\_minmax\_norm)

weights\_minmax\_norm\_stdev = statistics.stdev(weights\_minmax\_norm)

print(heights\_minmax\_norm\_stdev, weights\_minmax\_norm\_stdev)

Sample standard deviation of heights = 0.2733186671965223

Sample standard deviation of weights = 0.33449578711298195

**Problem 3. Data Cleaning** (16 points)

Suppose a big tech company holds a database about employees’ performance for year 2022 in Table [2.](#_bookmark2) Please answer the following questions.

Table 2: Employee Performance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Level** | **Year of Experience** | **Department** | **Rating Scores** | **Performance** |
| 43218 | Senior | 3.5 | Privacy | 83 | A |
| 26520 | Senior |  | Ads | 88 | 88 |
| 70645 | Junior | 1.5 | Privacy | 96 | A |
| -08002 | Staff | 8.0 | Retail | 92 | A |
| 31466 |  | 4.0 | Ads | 78 | A |

1. (4 points) (True or False) Does the ID column contain noisy data? Please justify your answer.

Yes, since negative values cannot be considered as ID in database system, -08002 is noisy data.

1. (4 points) For the ‘Level’ column, if we fill in the missing value with the value with highest probability, what would it be? Please justify your answer.

It should be “Senior”; ID 43218’s year of experience is 3.5, and its level is Senior, and there are 2 Senior Level in there. Therefore, “Senior” is the value with highest probability.

1. (4 points) For the ‘Year of Experience’ column, if we fill in the missing value with the attribute mean, what would it be? Please justify your answer.

(3.5 + 1.5 + 8.0 + 4.0) / 4 = 17 / 4 = 4.25

Therefore, missing value in this column will be 4.25.

1. (4 points) Does the *Performance* column contain any inconsistent data? Why?

Yes, most of the value of Performance are Character type, but ID 26520’s Performance is 88, which make the performance column inconsistent.

**Problem 4. Data Distribution** (20 points)

Table [3](#_bookmark3) shows 2*,* 100 student grades and their preferences on courses collected from 10 years of a university’s data.

(6 points) Let ***p*** = [*pA pB pC pD pF* ]*T* be the probability distribution of the course grade. Find ***p***.

PA = (290 + 212) / 2100 = 251 / 1050 = 0.23904

PB = (428 + 347) / 2100 = 31 / 84 = 0.36904

PC = (361 + 236) / 2100 = 199 / 700 =0.28428

PD = (103 + 78) / 2100 = 181 / 2100 = 0.08619

PF = (17 + 28) / 2100 = 3 / 140 = 0.02142

(4 points) Assuming the course grade and student’s preference on courses are independent, what is the expected number of students who both select CS412 and gets A?

P(A | CS412) =

Table 3: Student grades and their preference on courses.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | F |
| CS412 | 290 | 428 | 361 | 103 | 17 |
| CS512 | 212 | 347 | 236 | 78 | 28 |

(10 points) Calculate the *χ*2 correlation value for “CS412” and “CS512”

**Problem 5. Principal Component Analysis (PCA)** (20 points)

1. (8 points) Suppose we have 5 data points in a 2-dimensional Euclidean space: **x**1 =

[1 *−*2]*T* , **x**2 = [*−* 3 6 ]*T* , **x**3 = [*−*3 6]*T* , **x**4 = [ 6 *−* 12 ]*T* , **x**5 = [ 20 *−* 40 ]*T* . What are the first

7 7 5 5 3 3

and second principal components? (Please do not use code to solve this sub-problem.)

1. (12 points) Consider a dataset of automobiles’ information (file: automobile.csv [2](#_bookmark4)), we will analyze it through PCA. Use these five feature columns: curb-weight, horsepower, city-mpg, highway-mpg, price to obtain a subdataset, and answer the following ques- tion: what are the first principal component and second principal component of this subdataset?

If you want to use Python for this problem, please note that only standard Python li- brary such as Numpy, Scikit-learn, Pandas are allowed. Your source code is required to be submitted and please make sure it is bug-free.

2Modified based on the original dataset from h[ttps://www.kaggle.com/datasets/toramky/automobile-](http://www.kaggle.com/datasets/toramky/automobile-) dataset