CS 412 HW 1

Pouya Akbarzadeh, pa2

February 2021

1 Question 1

1.1 Answer

A) Maximum and minimum.
Mid-Term 37 , 100
Final: 35 , 100

B) First quartile Q1, median, and third quartile Q3
Mid-Term: 68.0 , 77.0 , 87.0
Final: 82.0 , 89.0 , 96.0

C) Mean
Mid-Term: 76.715
Final: 87.084

D) Mode
Mid-Term: 77, Repeated: 37 times
Final: 97, Repeated 60 times

E) Variance
Mid-Term: 173.10577
Final: 119.11294

1.2 Explanation

What the code is doing:
Mean: $A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$
$Median$:
$Q1 : (N+1) * 1/4$
$Q2 : (N+1) * 2/4$
$Q3 : (N+1) * 3/4$
$Mode : Mostrepeatedvalue$

$$Variance : A = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

1.3 Code

```
# Reading Data
data_table = pandas.read_table('data.online.scores.txt', names = ['ID#', 'Mid-Term', 'Final'])
finals = data_table['Final']
mid = data_table['Mid-Term']
# https://numpy.org/doc/stable/reference/generated/numpy.array.html
mid_term_score_arr = numpy.array(mid)
final_score_arr = numpy.array(finals)


# Part A
# Mid-Term Data
min_mid = numpy.min(mid_term_score_arr)
max_mid = numpy.max(mid_term_score_arr)
#Final Data
min_val_final = numpy.min(final_score_arr)
max_val_final = numpy.max(final_score_arr)


# Part B
# https://numpy.org/doc/stable/reference/generated/numpy.around.html
#numpy.median(final_score_arr)
mq1 = numpy.percentile(mid_term_score_arr,25) # Q1
mq2 = numpy.percentile(mid_term_score_arr,50) # median
mq3 = numpy.percentile(mid_term_score_arr,75) # Q3
q1 = numpy.percentile(final_score_arr,25) # Q1
q2 = numpy.percentile(final_score_arr,50) # median
q3 = numpy.percentile(final_score_arr,75) # Q3

# Part C
finals_mean = numpy.mean(final_score_arr)
# https://numpy.org/doc/stable/reference/generated/numpy.around.html
finals_mean_rounded = numpy.around(finals_mean, decimals = 3)
mid_mean = numpy.mean(mid_term_score_arr)
mid_mean_rounded = numpy.around(mid_mean, decimals =3)

# Part D
# https://stackoverflow.com/questions/16330831/most-efficient-way-to-find-mode-in-numpy-array
mid_term_mode = stats.mode(mid_term_score_arr)
final_mode = stats.mode(final_score_arr)

# Part E
# https://www.geeksforgeeks.org/numpy-var-in-python/
mid_var = numpy.var(mid_term_score_arr, dtype = numpy.float32)
var_final = numpy.var(final_score_arr, dtype = numpy.float32)
```

2 Question 2

2.1 Answer

A) Compute and compare the variance of midterm-original and midterm-normalized, i.e., the midterm scores before and after normalization.
Variance Final: 119.11294
Normalized Variance: 0.99999994

B) Given an original midterm score of 90, what is the corresponding score after normalization?
Mean: 76.715
Std-Dev: 13.156966785699508
Using v' = v - Avg / std-dev
Our score of 90 is normalized to 84.16924879042922

C) Compute the Pearson's correlation coefficient between midterm-original and finals-original.?
Pearson's correlation coefficient: (0.544424742312412, 3.0298169609726267e-78)

D) Compute the Pearson's correlation coefficient between midterm-normalizedand finals-original.

E) Compute the covariance between midterm-original and finals-original.
Covariance: [[173.27905405 78.25419419] [78.25419419 119.23217618]]

2.2 Explanation

2.3 Code

3 Question 3

3.1 Answer

a) Each library has multiple copies of each book. Based on all the books (treat the counts of the 100 books as a feature vector for each of the libraries), compute the Minkowski distance of the vectors for CML and CBL with regard to different h values:
i) 6152.0
ii) 715.3278968417211
iii) 170.0

b) Compute the cosine similarity between the feature vectors for CML and CBL.
Cosine similarity: 0.8414040256623079

3.2 Explanation

3.3 Code

4 Question 4

4.1 Answer

a) Calculate the distance between the binary attributes Buy Beer and Buy Diaper by assuming they are symmetric binary variables.
Distance: 0.015691868758915834

b) Calculate the distance between the binary attributes Buy Beer and Buy Diaper by assuming they are symmetric binary variables.
Jaccard Coefficient : 0.7317073170731707

c) Compute the 2 statistic for the contingency table.
X^2 : 2450.716326822006

d) Consider a hypothesis test based on the 2 statistic where the null hypothesis is that Buy Beer and Buy Diaper are independent. Can you reject the null hypothesis at a significance level of = 0.05? Explain your answer, and also mention the degrees of freedom used for the hypothesis test.
If the value was more than 0.05 we could not reject the null hypothesis
The value based on info given was: 0.0
We were able to reject
P value: 0.0
Deg of freedom: 1

4.2 Explanation

4.3 Code