CS 412 Assignment 1

Yayi Ning

September 2017

Notes for graders:

- For a cleaner version of all answers please see: $hw1.output.txt$
- Python coding are in QuestionX.yning4.py for question X = 1,2,3.
- Question 4 does not have code.

1 Question 1

Python code is placed in Question1.yning4.py

(a)

- Maximum mid-term score = 100
- Maximum mid-term score = 37

  Used python function : [max(np.array), min(np.array)]

  code:
  $max_score = max(self.scores)$
  $min_score = min(self.scores)$

(b)

- The first quartile = 68.0
- The median = 77.0
- The third quartile = 87.0

  The first quartile, median, and third quartile are the [25percen$^t h$, 50percen$^t h$, 75percen$^t h$] scores after arranged from small to large.

  Used python function : np.percentile(np.array, n$^t h$-percentile)

code:
$quartile_1 = np.percentile(self.scores, 25)$
$median = np.percentile(self.scores, 50)$
$quartile_3 = np.percentile(self.scores, 75)$

(c)

- The mean = 76.715

  Mean = $\frac{\sum_i Xi}{N}$
  Used python function : np.mean(np.arraay)

  code:
  $mean = np.mean(self.scores)$

(d)

- The mode = [83, 77]

  Mode = the number(s) in an array have highest frequency.

  Used python packages : collections, itertools

  Used python function : $groupby(Counter(np.array).most_common(), lambdax : x[1])$
  To pick the highest frequency group, where the frequency is calculated by $Counter().most_common()$.

  code:
  $freqs = groupby(Counter(self.scores).most_common(), lambdax : x[1])$
  $mode_=np.array([valforval, countinnext(freqs)[1]])$

(e)

- Empirical Variance = 173.279

  Empirical Variance : $s^2 = \frac{\sum_i (xi-\bar{x})^2}{N-1}$

  Used python function : $statistics.variance(np.array)$

  code:
  $var = statistics.variance(self.scores)$

2 Question 2

Python code is placed in Question3.yning4.py

(a)

- Variance before normalization: 173.279
- Variance after normalization: 1.0

Normalization scores : $score_{normalized}$ = array($\frac{(xi-\bar{x})}{\sigma}$)

Variance after normalization : $\sigma^2 = \frac{\sum_i (xi_{normalized}-\bar{x}_{normalized})^2}{N-1}$

Used python function: $var\_norm = statistics.variance(np.array)$

code:
$z_s cores = (self.scores - mean)/(var) * *0.5$
$var_n orm = statistics.variance(z_s cores)$

(b)

- Score of 90 after normalization: 1.009

Normalize 90 $= \frac{(90-\bar{x})^2}{\sigma} = \frac{(90-76.715)^2}{173.279^{0.5}} = 1.009$

Used python function : $statistics.variance(np.array)$

code:
$normalized\_90 = (90 - mean)/var * *0.5$

(c)

- Person's correlation coefficient between midterm scores and final scores is: 0.544

$r = \frac{n(\sum xy)-\sum x \sum y}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$

Used python function : $np.corrcoef(np.arrayone, np.arraytwo)$

code:
$corr\_coeff = np.corrcoef(midterm, final)[0,1]$

Notes: function np.corrcoef() return us an matrix, where M[0,0] is the correlation coefficient of (x,x), M[1,1] is the correlation coefficient of (y,y), and M[0,1] and M[1,0] are the correlation coefficient of (x,y) and (y,x) [they are the same].

(d)

- The covariance between midterm and final is: 78.254

COV[X,Y] $= \frac{\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y})}{n-1}$

Used python function : $np.cov(np.array\_one, np.array\_two)$

code:
$covariance = np.cov(midterm, final)[0, 1]$

Notes: function np.cov() return us an matrix, where COV[0,0] is the covariance of (x,x), COV[1,1] is the covariance of (y,y), and COV[0,1] and COV[1,0] are the covariance of (x,y) and (y,x) [they are the same].

3 Question 3

Python code is placed in Question3.yning4.py

(a)

- The Jaccard coefficient of Citadel's Maester Library (CML) and Castle Black's library(CBL) is: 0.322

$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{n_{in\_both\_A\_and\_B}}{N_{total\_books}}$

$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{58}{120+2+58} = 0.322$

code:
$Jaccard = count_1 1/(count\_01 + count\_10 + count\_11)$

(b)

- The Minkowski Distance of CML and CBL for h = 1 is: 6152.0
- The Minkowski Distance of CML and CBL for h = 2 is: 715.328
- The Minkowski Distance of CML and CBL for h = infinity is: 170.0

For $X\_CML = (x_1, x_2, ..., x_n)$ and $Y\_CBL = (y_1, y_2, ..., y_n)$

Minkowski_Distance$_{CML_and_CBL}$ =$(\sum_{i=1}^{n} |x_i - y_i|^h)^{\frac{1}{h}}$

Used python package : scipy

Used python function : $scipy.spatial.distance.minkowski(array\_1, array\_2, h\_orderofthenorm)$

code:
$mink_1 = scipy.spatial.distance.minkowski(self.cml, self.cbl, 1)$
$mink_2 = scipy.spatial.distance.minkowski(self.cml, self.cbl, 2)$
$mink_infi = scipy.spatial.distance.minkowski(self.cml, self.cbl, math.inf)$

(c)

- The cosine similarity of CML and CBL is : 0.841

For A_CML = $(a_1, a_2, ..., a_n)$ and B_CBL = $(b_1, b_2, ..., b_n)$

Cosine similarity $= \frac{A \cdot B}{\|A\|^2 \|B\|^2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$

Used python package : scipy

Used python function : $scipy.spatial.distance.cosine(array\_1, array\_2)$

code:
$cos_similarity = 1 - scipy.spatial.distance.cosine(self.cml, self.cbl)]$

Notes: The function scipy.spatial.distance.cosine() calculate the cosine distance. To get the cosine similarity, use 1 - cosine distance = cosine similarity.

(d)

- Kullbac-Leibler divergence between Citadel's Maester Library (CML) and Castle Black's library(CBL) is : 0.207

  For P(i) = the probability of p(x = book_i) in CML and Q(i) = the probability of p(y = book_i) in CBL

  $D_{KL}(P \parallel Q) = \sum_i^n log \frac{P(i)}{Q(i)}$

  code:
  $p_cml = self.cml/sum(np.array(self.cml)[:])$
  $p_cbl = self.cbl/sum(np.array(self.cbl)[:])$
  $D_cml_cbl = sum(p_cml * np.log(p_cml/p_cbl))$

  Notes:

4 Question 4

Calculate the chi-square correlation value

- Total count = 150 + 40 + 15 + 3300 = 3505
- E__Buy diaper&Buy beer = (150 + 15) * (150 + 40) / 3505 = 8.944
- E__ Not Buy diaper&Buy beer = (40 + 3300) * (150 + 40) / 3505 = 181.056
- E__ Buy diaper& Not Buy beer = (150 + 15) * (15 + 3300) / 3505 = 156.056
- E__ Not Buy diaper& Not Buy beer = (40 + 3300) * (15 + 3300) / 3505 = 3158.944

$\chi^2 = \frac{(150-8.944)^2}{8.944} + \frac{(40-181.056)^2}{181.056} + \frac{(15-156.056)^2}{156.056} + \frac{(3300-3158.944)^2}{3158.944}$

$\chi^2 = 2468.183$