ECE-6143 Intro to ML

Homework1:

Haotian Yi

Net ID: hy1651

Q1:

Suppose µ is what we wish.

Prove:

for each attribute ,it has distribution of

because

for each :

to get minimum :

so,

thus, we can conclude that

Q2(a):

1)For and :

because : < ⬄ <

(do square of both side)

and: = + ,

represents cross terms of

and > 0 (when are not all equal to 0)

so, <

so, ≤

(equal when d = 1 or are all equal to 0 )

Thus, ≤

2)For and :

Inequation above is equal when d = 1 or are all 0

when d and are not all 0:

+ (d-1) ≥ + 3

represents cross terms of

for d =2, cross terms of is and

,

for d =3, cross terms of is

and

, according to two instances above, we can infer that:

(when d and are not all 0),

so < and then ≤

3) Thus: ≤

Q2(b):

1) For and :

+ +

(maxi is the subscript of the )

thus ≤ (equal when d = 1 or all 0 except )

2) For and :

+ +

(maxi is the subscript of the )

It is obvious: ≥ +

thus ≤:

3)Thus: ≤

Q2(c):

1) For and :

+

≥ =

(maxi is the subscript of the )

(equal when d = 1 or all 0 except a )

2) For and :

+

(+ and it is obviously ≥

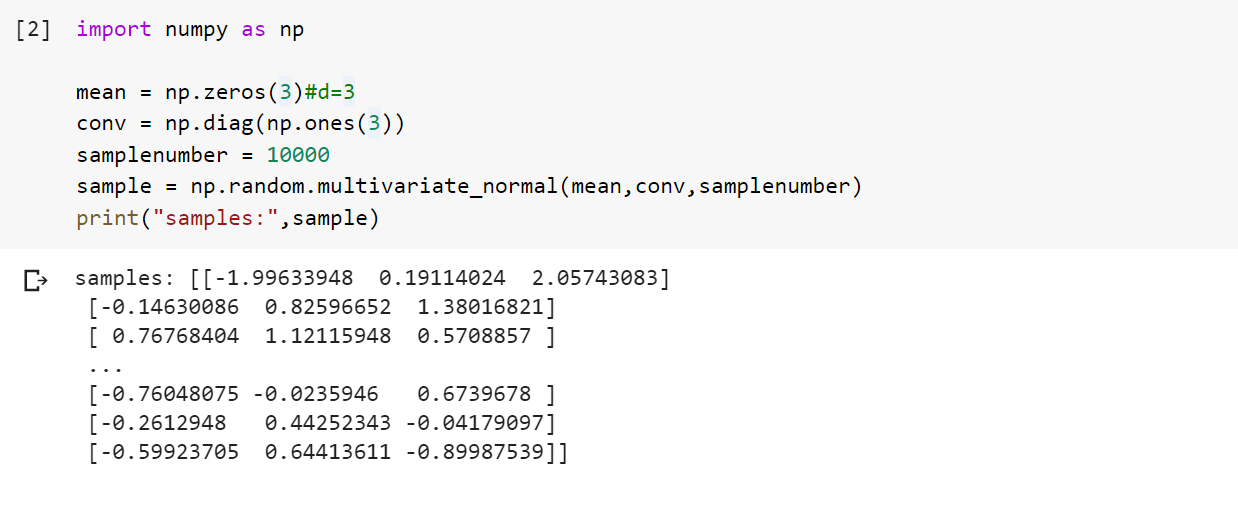
(equal when d = 1 or all attributes are equal)

3)Thus: ≤

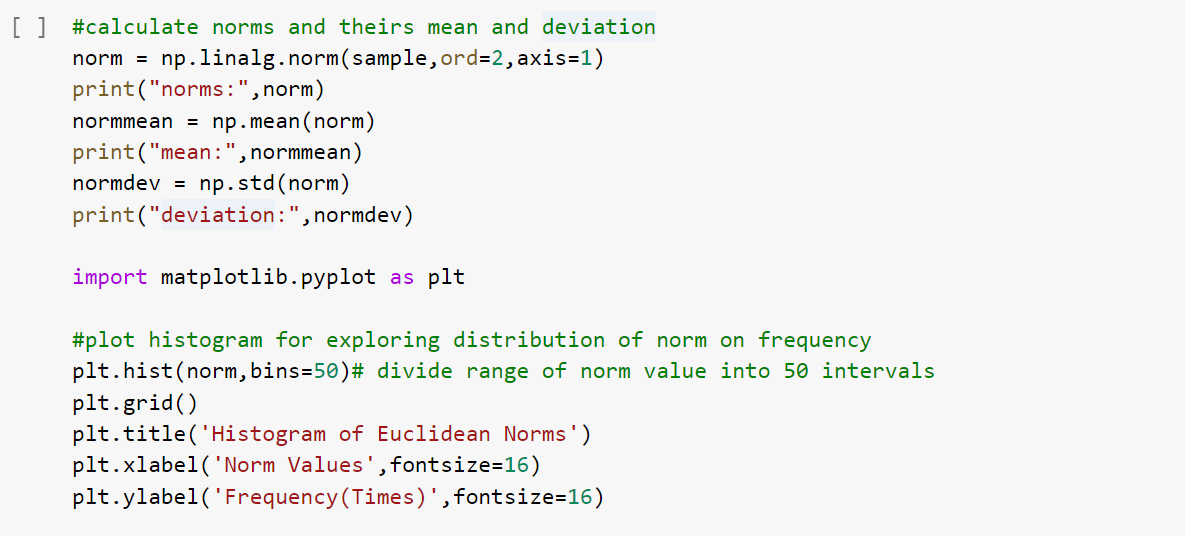
For question 3 and question 4, there are more comments and interpretation in details in .ipynb files.

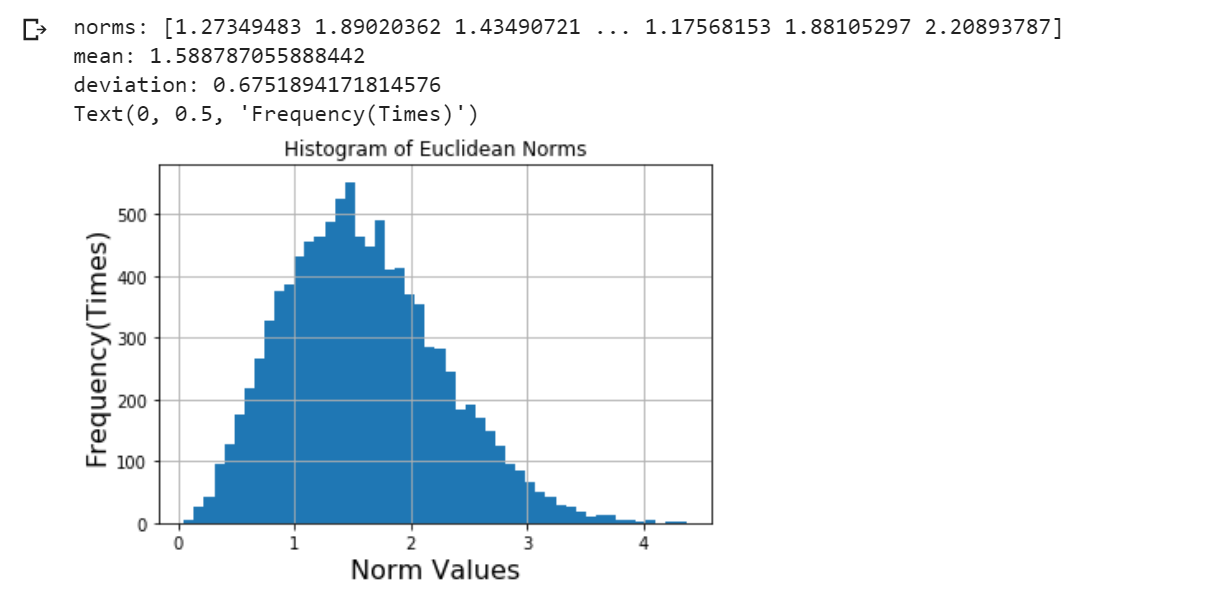
Q3:

a. Fix d = 3 and generate 10,000 random samples from the standard multi-variate Gaussian distribution defined in Rd.



b. Compute and plot the histogram of Euclidean norms of your samples. Also calculate the average and standard deviation of the norms.

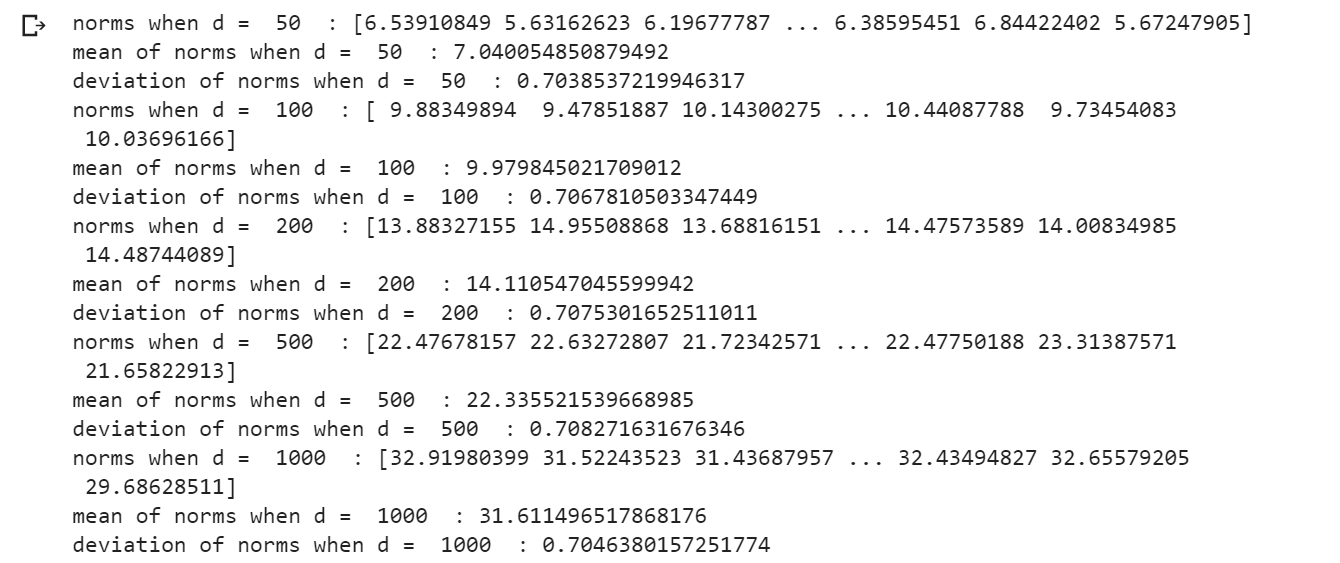


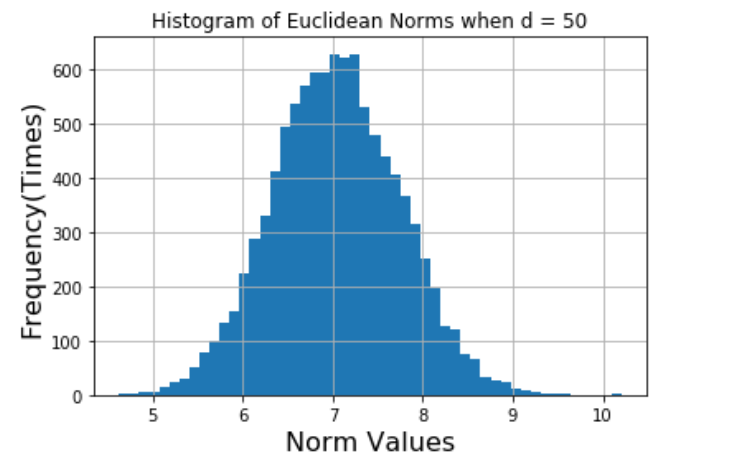


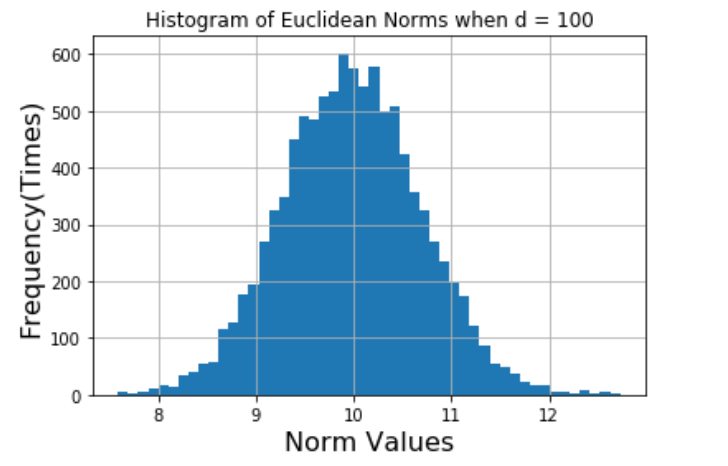
c. Increase d on a coarsely spaced log scale all the way up to d = 1000 (say d =

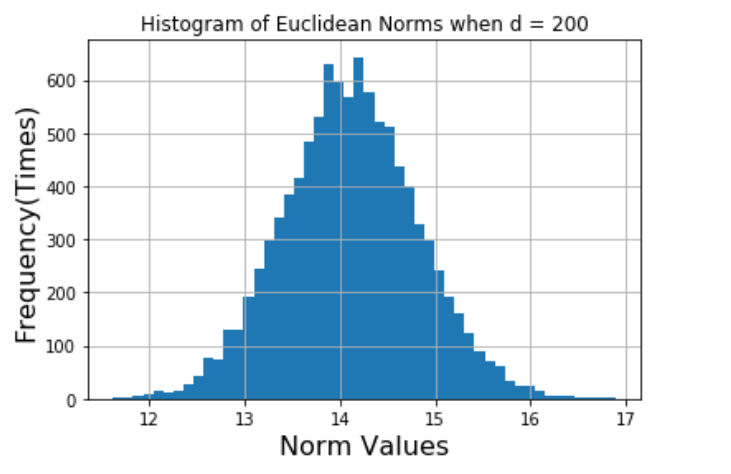
50; 100; 200; 500; 1000), and repeat parts (a) and (b). Plot the variation of the average and the standard deviation of Euclidean norm of the samples with increasing d.

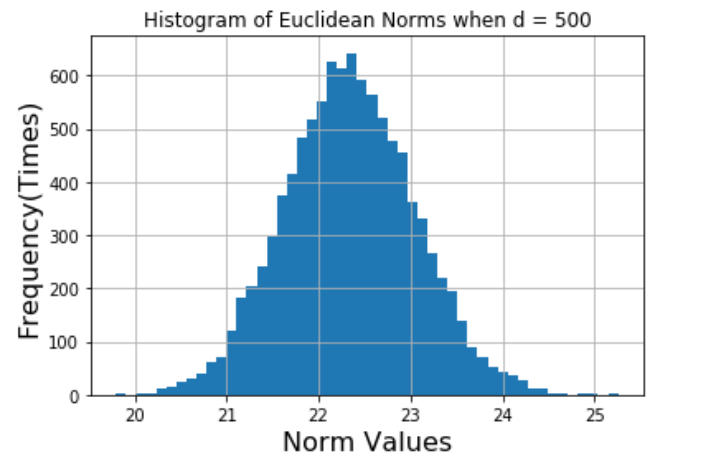
Write a for loop, achieve goals according to the method of part a and b, here are results below:

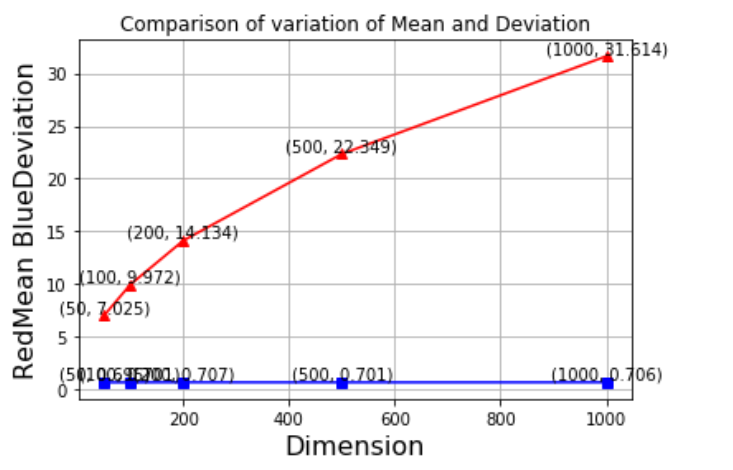
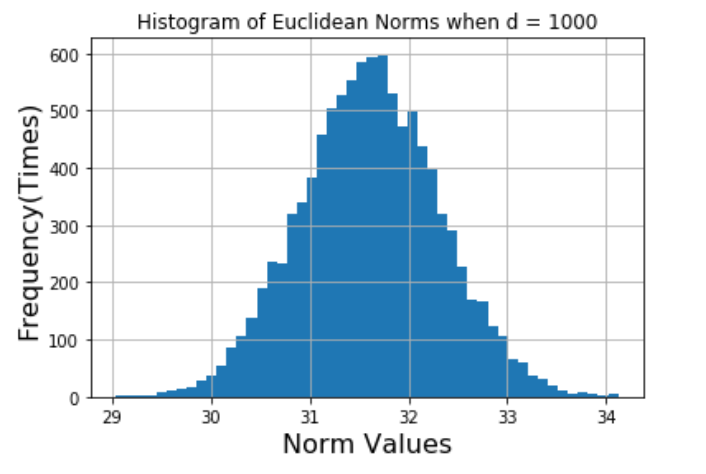


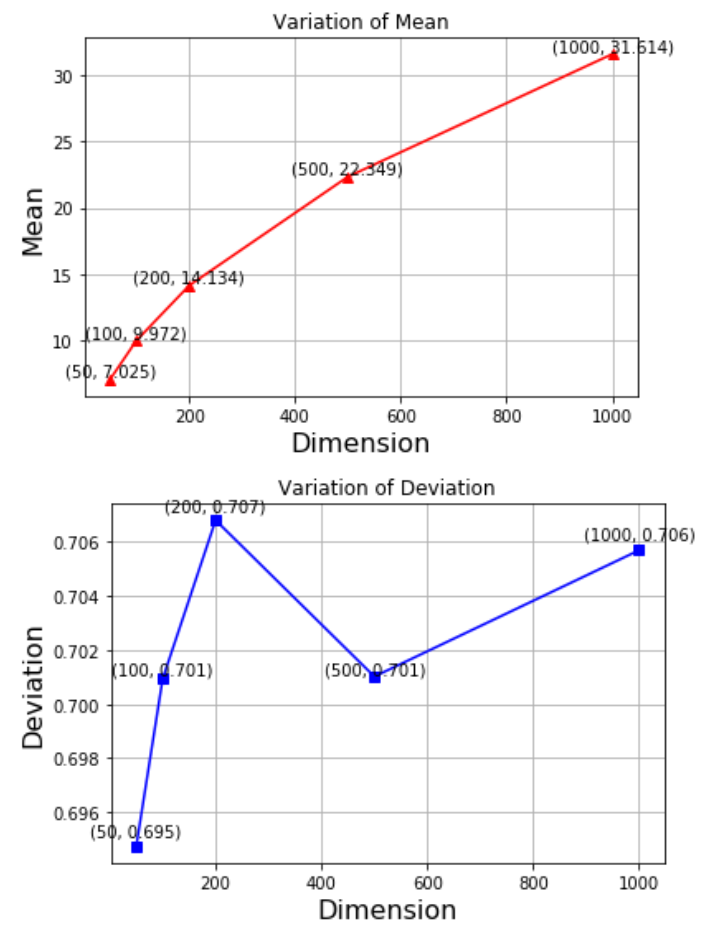












d. What can you conclude from your plot from part (c)?

According to diagram above, I discover that:

1.in high dimension, most of mass is concentrated near the mean of Euclidean norms,

2.mean of norms increase with dimension,

3.deviation is approximately not changed with dimension.

Q4:

a. Write a small parser to read each document and convert it into a vector of words.

define an function to read each document and convert it into a vector of words and obtain appearance time of each word, tf of each word and a list of all kinds of word in each .txt document saving as dictionaries

b. Compute tf-idf values for each word in every document as well as the query.

b1. write a function to obtain result of function "preprocess" on each document from paths provided in a list and put appearance counter of words, tf of words in every document and a collection of all kinds of words appear in all documents respectively into three lists : counter, tfdic, all\_words

b2.write a function to unify the format of counter and tfdic to make every dictionary have same attributes and dimension, put them respectively into two lists: counters, tfdics

b3. write a function to obtain idf of each words in every document and put them together into one list: idfdics , actually they are all the same but I save them as the same format and container as tf

b4. write a function to obtain tf-idf of each words in every document and put them together into one list: tfidfdics

c. Compute the cosine similarity between tf-idf vectors of each document and the query.

define a function that can calculate cosine similarity between two dictionaries of tf-idf

d. Report the document with the maximum similarity value.

define a recognition function that can use function defined above to compare cosine similarity among documents and return sequence number of result document and maximum value of similarity

results:

d 4 is the document with maximum cosine similarity value of 0.19616295821215432

Q5: approximately 21 hours for whole homework