CS573-HW0

Yifan Fei

January 2018

Q0

1

Yes, I discussed the homework with another person but came up with my own answers. His name(s) is Yupeng Han.

2

I have used online resources to help me answer this question, but I came up with my own answers. Here is a list of the websites I have used in this homework:

Bayes Rule: https://en.wikipedia.org/wiki/Bayes%27_theorem

Maximum likelihood estimation: https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

Python scatter 2D: https://matplotlib.org/gallery/shapes_and_collections/scatter.html

Python scatter 3D: https://matplotlib.org/gallery/mplot3d/scatter3d.html

MIT webcourse about SVD: https://ocw.mit.edu/resources/res-18-009-learn-differential-equations-up-close-with-gilbert-strang-and-cleve-moler-fall-2015/differential-equations-and-linear-algebra/applied-mathematics-and-ata/singular-value-decomposition-the-svd/

Inner Product: https://stackoverflow.com/questions/11033573/difference-between-numpy-dot-and-inner

Q1

a

What is the distribution of (Y1,Y2)? What are the parameters of this distribution?

Since the linear combination of multiple i.i.d random variables with normal distributions is also the variable with normal distribution, both Y1 and Y2 are normal distributed.

$$Y_1 \sim Normal(1 \cdot 0 + 2 \cdot 0, 1^2 \cdot 1^2 + 2^2 \cdot 1^2), Y_1 \sim Normal(0, 5) \tag{1}$$

$$Y_2 \sim Normal(2 \cdot 0 + 1 \cdot 0, 2^2 \cdot 1^2 + 1^2 \cdot 1^2), Y_2 \sim Normal(0, 5) \tag{2}$$

(Y_1, Y_2) is a multivariate normal distribution.

$$Cov(Y_1, Y_2) = E[Y_1 Y_2] - E[Y_1]E[Y_2] = E[(X_1 + 2X_2)(2X_1 + X_2)] = 2E[{X_1}^2] + 2E[{X_2}^2] = 4$$

Similarly,

$$Cov(Y_2, Y_1) = 4$$

So the covariance matrix $\mathbf{\Sigma}$ is:

$$\begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) \\ Cov(Y_2, Y_1) & Var(Y_2) \end{bmatrix}$$

In conclusion, the distribution of (Y_1, Y_2) is with:

$$\boldsymbol{\mu} = [0, 0]$$

$$\mathbf{\Sigma} = [5, 4; 4, 5]$$

b

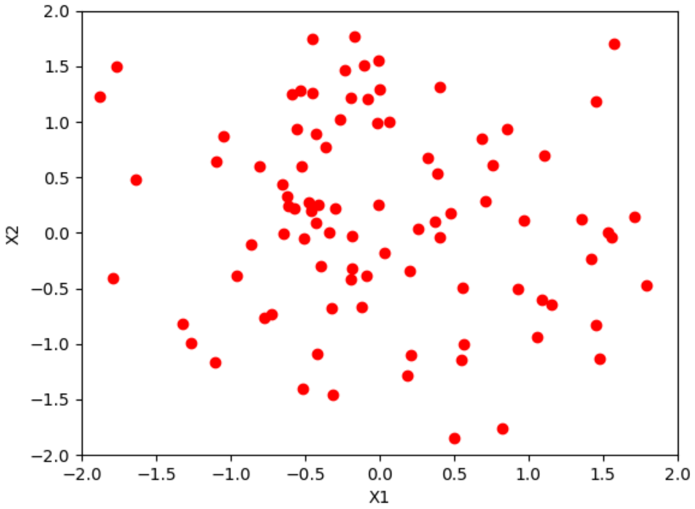

Figure 1: n = 100, 2D scatter plot with 100 samples of (X1,X2)

c

Using python, the probability that an observation of (X1,X2) falls within a circle of radius 0.5 centered at (0,0) is about 0.16. Notice that every experiment leads to different answers, the range is mainly between 0.12 and 0.17.

d

Using python, the empirical probability that an observation falls within a sphere of radius 0.5 centered at (0,0,0) is about 0.05. Notice that every experiment leads to different answers, the range is mainly between 0.02 and 0.07.

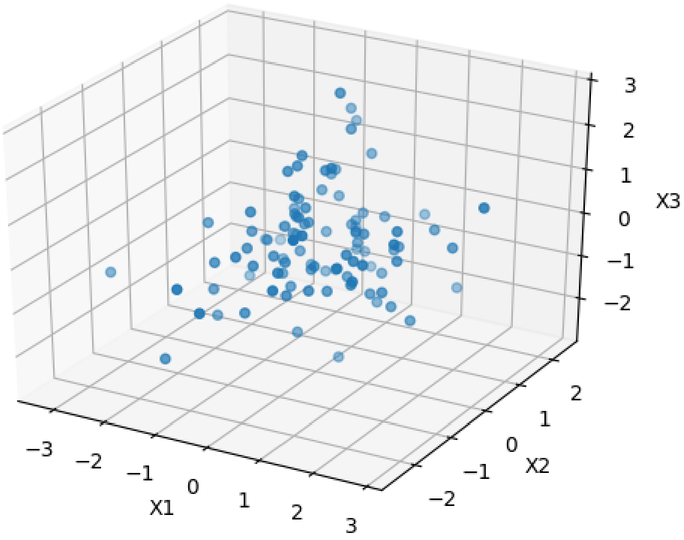

Figure 2: n = 100, 3D scatter plot with 100 observations from (X1,X2,X3)

e

Repeat the above exercise for n = 1000: Using python, the probability that an observation of (X1,X2) falls within a circle of radius 0.5 centered at (0,0) is about 0.125, while the empirical probability that an observation falls within a sphere of radius 0.5 centered at (0,0,0) is about 0.039. Notice that every experiment leads to different answers, the range for 2D experiment is mainly between 0.110 and 0.160, and the range for 3D experiment is mainly between 0.020 and 0.040.

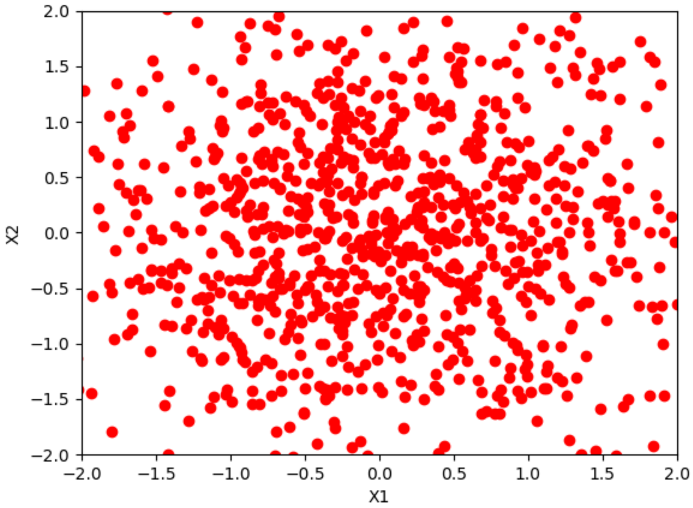

Figure 3: n = 1000, 2D scatter plot with 100 samples of (X1,X2)

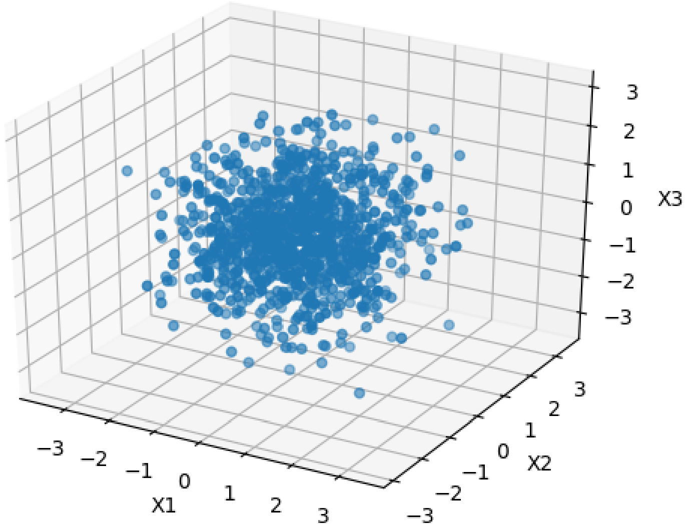

Figure 4: n = 1000, 3D scatter plot with 100 observations from (X1,X2,X3)

f

If $X_1, X_2...X_k$ are independent, standard normal random variables, then the sum of their squares,$Q = \sum_{i=1}^{k} {X_i}^2$ is distributed according to the chi-squared distribution with k degrees of freedom, i.e.

$$Q \sim \chi^2(k)$$

Suppose the demension of hypersphere is n. Consider event A as an observation falls into hypersphere of radius 0.5 centered at the origin (0, 0, . . .), so the probability:

$$P(A) = P(\sum_{i=1}^{n} {X_i}^2 \leq 0.25) = \frac{\gamma(n/2, x/2)}{\Gamma(n/2)} \mid_{x=0.25} = \frac{\gamma(n/2, 0.125)}{\Gamma(n/2)}$$

After checking the table for big degrees of freedom, it is said that When $n \to \infty$

$$\lim_{n\to\infty} P(A) = 0$$

Code for Q1

```
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
import math

X1 = np.random.randn(1000)
X2 = np.random.randn(1000)
plt.plot(X1, X2, 'ro')
plt.axis([-2, 2, -2, 2])
plt.xlabel('X1')
plt.ylabel('X2')
plt.savefig("Q1(b)_n1000.png")

R = X1**2 + X2**2
r = [i for i in R if i <= 0.5**2]
p_b = len(r)/1000
print(p_b)

X3 = np.random.randn(1000)
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(X1, X2, X3)
ax.set_xlabel('X1')
ax.set_ylabel('X2')
ax.set_zlabel('X3')
plt.savefig("Q1(c)_n1000.png")

R3 = R + X3**2
r3 = [i for i in R3 if i <= 0.5**2]
p_c = len(r3)/1000
print(p_c)

plt.show()
```

Q2

a

Prove the conditional version of Bayes rule:

$$P(B \mid A, C) = \frac{P(A \mid B, C)\, P(B \mid C)}{P(A \mid C)}$$

Based on Bayes rule, we have:

$$P(B \mid A, C) = \frac{P(A, B, C)}{P(A, C)}$$

$$P(B \mid A, C) = \frac{P(A \mid B, C)\, P(B, C)}{P(A, C)}$$

$$P(B \mid A, C) = \frac{P(A \mid B, C)\, P(B \mid C)\, P(C)}{P(A \mid C)\, P(C)}$$

Finally, we get the result by cancellation:

$$P(B \mid A, C) = \frac{P(A \mid B, C)\, P(B \mid C)}{P(A \mid C)}$$

b

For

$$X_i \sim Normal(\mu, \sigma^2)$$

which has probability density function

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

the corresponding probability density function for a sample of n independent identically distributed normal random variables (the likelihood) is:

$$f(x_1, x_2..., x_n \mid \mu, \sigma) = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2}$$

$$f(x_1, x_2..., x_n \mid \mu, \sigma) = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-(\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2)/2\sigma^2}$$

we maximize the likelihood,

$$L(\mu, \sigma) = f(x_1, x_2..., x_n \mid \mu, \sigma)$$

$$log(L(\mu, \sigma)) = -(n/2)log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Let the derivative equals to 0:

$$\frac{\partial log(L(\mu, \sigma))}{\partial \mu} = 0$$

We get the estimated parameter:

$$\hat{\mu} = \bar{x}$$

This is the maximum of the function since it is the only turning point in μ and the second derivative is strictly less than zero. Its expectation value is equal to the parameter μ of the given distribution.

$$E[\hat{\mu}] = \mu$$

which means that the estimator $\hat{\mu}$ is unbiased.

$$\frac{\partial log(L(\mu, \sigma))}{\partial \sigma} = 0$$

We get the estimated parameter:

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

Inserting the estimate $\mu = \hat{\mu}$ we obtain:

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

the estimated σ :

$$E[\hat{\sigma}] = \frac{n-1}{n}\sigma^2$$

which means that the estimator $\hat{\sigma}$ is biased.

c

Estimate parameters by maximum likelihood using gradient ascent:

```
# 15 lines
import numpy as np
import math
[n,i] = [10000, 0] # n is dimension of X and i is iteration number
x = np.random.randn(n)
[tol,error] = [0.0001,1]# tolerance of error, and error initialization
gama = 10e-6  # step size multiplier
[mu,sigma] = [5,5] # initial value
df_mu = lambda mu, sigma: n*(np.mean(x)-mu)/(sigma**2) # derivative by mu
df_sigma = lambda mu, sigma: -n/sigma+(n*np.var(x))/(sigma**3) # derivative by sigma
while error > tol and i < 10e6:
        [mu,sigma] = map(lambda a,b:a+b,[mu,sigma],\
                [gama*df_mu(mu, sigma), gama*df_sigma(mu, sigma)])
        error = np.linalg.norm([df_mu(mu, sigma), df_sigma(mu, sigma)])
        i = i+1
print([mu,sigma])
```

Finally, we get $\mu = -0.0061, \sigma = 1.0023$ when error tolerance is 10e-6, step size is 10e-6, max iteration is 10e6, and n equals to 10e4. This is very close to real values $\mu = 0, \sigma = 1$ After doing many experiments, it is shown that the larger the n and iteration number, the smaller the step size and error tolerance, the precise the estimated mean and variance will be.

d

Prior distribution of μ is $N(\mu_0, \sigma_0)$ the posterior distribution $P(\mu \mid x_1, x_2..., x_n, \sigma)$ is:

$$P(\mu \mid x_1, x_2..., x_n, \sigma) = \frac{\frac{1}{\sigma_0\sqrt{2\pi}} e^{-(\mu-\mu_0)^2/2\sigma_0^2} (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2}}{\int_{-\infty}^{+\infty} \frac{1}{\sigma_0\sqrt{2\pi}} e^{-(\mu-\mu_0)^2/2\sigma_0^2} (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2} d\mu}$$

After cancellation:

$$P(\mu \mid x_1, x_2..., x_n, \sigma) = \frac{A}{\int_{-\infty}^{+\infty} A d\mu}$$

where

$$A = e^{-(\frac{(\mu-\mu_0)^2}{2\sigma_0^2} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2})}$$

$$A = e^{-(\frac{(\mu-\mu_0)^2}{2\sigma_0^2} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2})}$$

$$A = e^{-(\frac{(\mu-\mu_0)^2}{2\sigma_0^2} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2})}$$

$$A = e^{-\frac{1}{2}((\frac{(\mu^2-2\mu\mu_0+\mu_0^2}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 + n\mu^2 - 2n\mu\bar{x}}{\sigma^2}))}$$

$$A = e^{-\frac{1}{2}B}$$

where B is:

$$B = \mu^2(\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}) - 2\mu(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}) + (\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x} + \sum_{i=1}^{n} x_i^2}{\sigma^2})$$

$$B = (\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2})[\mu^2 - 2\mu\frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}} + C]$$

where C is a constant. Finally, we get:

$$B = (\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2})\{[\mu - \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}}]^2 + C_1\}$$

where C_1 is a constant and can be cancelled cause it exists both in numerator and denominator.

Notice that the core part is a normal distribution:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-0.5\{(\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2})[\mu - \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}}]^2\}} d\mu = 1$$

So:

$$P(\mu \mid x_1, x_2..., x_n, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-0.5\{(\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2})[\mu - \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}}]^2\}}$$

We can easily see that it is a normal distribution with mean as:

$$\frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}}$$

variance as

$$\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}$$

In conclusion, the posterior distribution $P(\mu \mid x_1, x_2..., x_n, \sigma)$ is proved to be the same normal distribution with the same mean and variance as that in the question.

Q3

a

i

$+ \mid Y$: The test correctly identifies the presence of WNV in 95% of cases $+ \mid N$: The test gives false positives in 1/10000 cases

$$P(+ \mid Y) = 0.95$$

$$P(+ \mid N) = 0.0001$$

Y: The West Nile virus infected approximately 2000 people with 300 million people susceptible N: The West Nile virus did not infect approximately 2000 people with 300 million people susceptible

$$P(Y) = \frac{2000}{300,000,000} = \frac{1}{150,000}$$

$$P(N) = 1 - \frac{2000}{300,000,000} = \frac{149,999}{150,000}$$

By Bayer rules:

$$P(+) = P(+ \mid N)P(N) + P(+ \mid Y)P(Y) = 0.0001 \cdot \frac{149,999}{150,000} + 0.95 \cdot \frac{1}{150,000} = 0.0001063$$

The probability that Sheldon has WNV(event A) is:

$$P(A) = P(Sheldon\, has\, WNV) = P(Y \mid +) = \frac{P(+ \mid Y)P(Y)}{P(+)} = \frac{0.95 \cdot \frac{1}{150,000}}{0.0001063} = 0.05956$$

ii

Assume dying of a fatality rate of any cause of 0.2% to be event B:

$$P(B) = 0.002$$

The probability that Sheldon will die due to WNV(event C):

$$P(C) = 0.04 \cdot 0.05956 = 0.002382$$

The probability that Sheldon will die is:

$$P(Sheldon\, will\, die) = P(B) + P(C) - P(B, C)$$

$$P(Sheldon\, will\, die) = 0.002 + 0.002382 - 0.002382 \cdot 0.002 = 0.004377 \approx 0.004$$

b

The probability for one person to win at one rolling is:

$$P(Alice\, wins\, at\, one\, time) = \frac{15}{36} = \frac{5}{12}$$

The probability for one person to win at the end is:

$$P(Alice\, wins\, at\, one\, time) = \frac{5}{12}(1 + \frac{1}{6} + \frac{1}{6^2} + ...) = \frac{1}{2}$$

It is interesting that the final result is intuitive, which is half-half.

The probability for Alice rolled 3 and win at the end is:

$$P(Alice\, wins\, at\, one\, time) = \frac{2}{36}(1 + \frac{1}{6} + \frac{1}{6^2} + ...) = \frac{1}{15}$$

So the probability for Alice rolled 3 given that she won is:

$$P(Alice\, rolled\, 3 \mid she\, won) = \frac{P(Alice\, rolled\, 3\, and\, won)}{P(Alice\, won)} = \frac{\frac{1}{15}}{\frac{1}{2}} = \frac{2}{15}$$

Q4

```
# Part a, 14 lines #
import numpy as np
import csv

def season2code(season):
        if "winter" in season: code = -1
        elif 'summer' in season: code = -1
        elif 'spring' or 'fall' in season: code = 0
        else: code = 999
        return code

A = 999*np.ones((101,101))
with open('matrix.csv', 'r') as csvfile:
        reader = csv.reader(csvfile, delimiter=',')
        for row in reader:
                (i,j) = (int(row[0]),int(row[1]))
                A[i,j]= season2code(row[2])

# Part b #
print(A[1:3,9:11])

# Part c #
(u,v) = (np.ones((1,10)),np.ones((1,10)))
print(np.inner(u, v))
```

In part a, I define a function to change text to number, which is not perfect as using a dictionary. Inspired by Meng Liu, I modified the code to make it simpler. Plus by this way I can define the shape of matrix A after reading all the data. The modified code is shown below.

```
# Part a, 11 lines #
import numpy as np
import csv

seasons = {'winter': -1, 'summer':1, 'fall':0, 'spring':0}
List = []; # A flat list to store data
with open('matrix.csv', 'r') as csvfile:
        reader = csv.reader(csvfile, delimiter=',')
        for row in reader:
                List.append([int(row[0].strip()),int(row[1].strip()),\
                seasons[row[2].strip()]])
        A = 999*np.ones((max(List,key = lambda x:x[0])[0]+1,\
        max(List,key = lambda x:x[1])[1]+1))
        for row in reader:
                A[row[0],row[1]]=row[2] # convert List to matrix A
```

Q5

By Singular Value Decomposition, we have:

$$A = U\Sigma V^T = u_1\sigma_1 v_1^T + u_2\sigma_2 v_2^T + ... + u_r\sigma_r v_r^T \tag{3}$$

Where u from 1 to r is an orthonormal basis for the column space, v from 1 to r is an orthonormal basis for the row space, and σ are singular values.

$$B = A^T A = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T \Sigma V^T \tag{4}$$

On the right you see the eigenvector matrix V for the symmetric positive (semi) definite matrix ATA since $\sigma^2 \geq 0$. And $\sigma^T \sigma$ must be the eigenvalue matrix of $A^T A$: Each σ^2 is $\lambda A^T A$ The v's are eigenvectors of $A^T A$.

Similarly,

$$C = AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma^T V^T V\Sigma U^T = U\Sigma^T \Sigma U^T \tag{5}$$

The u's are eigenvectors of AA^T.