HOMEWORK 1

>>Yuanzhe Liu<<
>>9084055558<<

Instructions: This is a background self-test on the type of math we will encounter in class. If you find many questions intimidating, we suggest you drop 760 and take it again in the future when you are more prepared.
Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to Canvas. There is no need to submit the latex source or any code. Please check Piazza for updates about the homework.

1 Vectors and Matrices [6 pts]

Consider the matrix X and the vectors $\mathbf{y}$ and $\mathbf{z}$ below:

$$X = \begin{pmatrix} 3 & 2 \\ -7 & -5 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad \mathbf{z} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

1. Computer $\mathbf{y}^T X \mathbf{z}$

$$\begin{aligned} y^\top X z &= \left(\begin{array}{c} 2 \\ 1 \end{array} \right)^\top \left(\begin{array}{cc} 3 & 2 \\ -7 & -5 \end{array} \right) \left(\begin{array}{c} 1 \\ -1 \end{array} \right) \\ &= (-1, -1) \left(\begin{array}{c} 1 \\ -1 \end{array} \right) \\ &= -1 + 1 \\ &= 0 \end{aligned}$$

2. Is X invertible? If so, give the inverse, and if no, explain why not.
   By properity of invertible matrix: X is invertible if $\det(x) \neq 0$, that is:
   $\Rightarrow \det(X) = \det(\left| \begin{array}{cc} 3 & 2 \\ -7 & -5 \end{array} \right|) = 3 \cdot (-5) - 2 \cdot (-7) = -1 \neq 0$
   Thus, X is invertible.

2 Calculus [3 pts]

1. If $y = e^{-x} + \arctan(z) x^{6/z} - \ln \dfrac{x}{x+1}$, what is the partial derivative of y with respect to x?

$$\begin{aligned} \frac{\partial y}{\partial x} &= -e^{-x} + \frac{6}{z} \arctan(z) x^{\frac{6}{z}-1} - \frac{x+1}{x} \cdot \left(\frac{1}{(x+1)^2} \right) \\ &= -e^{-x} + \frac{6}{z} \arctan(z) x^{\frac{6}{z}-1} - \frac{1}{x(x+1)} \end{aligned}$$

3 Probability and Statistics [10 pts]

Consider a sequence of data $S = (1, 1, 1, 0, 1)$ created by flipping a coin x five times, where 0 denotes that the coin turned up heads and 1 denotes that it turned up tails.

1. (2.5 pts) What is the probability of observing this data, assuming it was generated by flipping a biased coin with $p(x = 1) = 0.6$?

$$f(x) = \begin{cases} p^x(1-p)^{n-x} \\ 0 \end{cases}$$
$$\Rightarrow f(x) = 0.6^4 \cdot 0.4^1 = 0.05184$$

2. (2.5 pts) Note that the probability of this data sample could be greater if the value of $p(x = 1)$ was not 0.6, but instead some other value. What is the value that maximizes the probability of S? Please justify your answer.

$$\begin{aligned} f(x) &= p^x(1-p)^{5-x} = p^4 \cdot (1-p) \\ \hat{p} &= \operatorname{argmax} f(x) \\ &= \operatorname{argmax} p^4(1-p) \\ &= \operatorname{argmax} \left(p^4 - p^5\right) \\ &= \operatorname{argmax} \log\left(p^4 - p^5\right) \\ &= \operatorname{argmax} \left(\log\left(p^4(1-p)\right)\right) \\ &= \operatorname{argmax} \left(\log p^4 + \log(1-p)\right) \\ &= \operatorname{argmax}(4\log p + \log(1-p)) \end{aligned}$$

Let $f(p) = 4\log p + \log(1-p) = 4\ln p + \ln(1-p)$

$$\begin{aligned} \frac{\partial f(p)}{\partial p} &= \frac{4}{p} - \frac{1}{1-p} = 0 \\ \Rightarrow \quad & 4(1-p) - p = 4 - 5p = 0 \\ p &= \frac{4}{5} \end{aligned}$$

Thus, when $p = \frac{4}{5} = 0.8$, we have the maximum of the probability of S.

3. (5 pts) Consider the following joint probability table where both A and B are binary random variables:

A	B	$P(A, B)$
0	0	0.3
0	1	0.1
1	0	0.1
1	1	0.5

(a) What is $P(A = 0|B = 1)$?

$$\begin{aligned} P(A = 0 \mid B = 1) &= \frac{P(A = 0, B = 1)}{P(B = 1)} \\ &= \frac{0.1}{0.6} \\ &= \frac{1}{6} \end{aligned}$$

(b) What is $P(A = 1 \vee B = 1)$?

$$\begin{aligned} P(A = 1 \cup B = 1) &= P(A = 1, B = 0) + P(A = 1, B = 1) + P(A = 0, B = 1) \\ &= 0.1 + 0.5 + 0.1 \\ &= 0.7 \end{aligned}$$

4 Big-O Notation [6 pts]

For each pair (f, g) of functions below, list which of the following are true: $f(n) = O(g(n))$, $g(n) = O(f(n))$, both, or neither. Briefly justify your answers.

1. $f(n) = \ln(n)$, $g(n) = \log_2(n)$.
   If $\lim_{n\to+\infty} \frac{f(n)}{g(n)} = c$, then $f(n) = O(g(n))$

$$\lim_{n\to+\infty} \frac{f(n)}{g(n)} = \ln(2) \tag{1}$$

   If $\lim_{n\to+\infty} \frac{g(n)}{f(n)} = c$, then $g(n) = O(f(n))$

$$\lim_{n\to+\infty} \frac{g(n)}{f(n)} = \ln(2) \tag{2}$$

   Thus, we have $f(n) = O(g(n)), g(n) = O(f(n))$.

2. $f(n) = \log_2 \log_2(n)$, $g(n) = \log_2(n)$.
   If $\lim_{n\to+\infty} \frac{f(n)}{g(n)} = c$, then $f(n) = O(g(n))$

$$\lim_{n\to+\infty} \frac{f(n)}{g(n)} = \lim_{n\to+\infty} \frac{1}{\log_2(n)\ln(2)} = 0 \tag{3}$$

   If $\lim_{n\to+\infty} \frac{g(n)}{f(n)} = c$, then $g(n) = O(f(n))$

$$\lim_{n\to+\infty} \frac{g(n)}{f(n)} = \lim_{n\to+\infty} \log_2(n)\ln(2) = +\infty \tag{4}$$

   Thus, we have $f(n) = O(g(n)), g(n) \neq O(f(n))$.

3. $f(n) = n!$, $g(n) = 2^n$.
   Firstly, we will prove $g(n) = O(f(n))$ By mathematical induction, we will prove that $2^n \leq n!$ if $n \geqslant 4$
   - Base case: $2^4 = 16 < 4! : 24$, it is satisfied when $n = 4$
   - Induction Step: we assume that $2^k \leq k!$ if $n = k$, then We will prove that it is also satisfied when $n = k+1$

$$2^{k+1} = 2 \cdot \left(2^k\right) \leqslant 2 \cdot k! < (k+1)k! = (k+1)! \tag{5}$$

   Thus, it is satisfied when $n = k + 1$. Also, it is satisfied when $n = 1, 2, 3$. Hence, $g(n) = O(f(n))$

   Secondly, we will prove $f(n) = O(g(n))$

$$\begin{aligned}
\frac{f(n)}{g(n)} &= \frac{n!}{2^n} \\
&= \frac{1 \cdot 2 \cdots n}{2 \cdot 2 \cdots 2} \\
&= \frac{1}{2} \cdot \frac{2}{2} \cdot \frac{3}{2} \cdots \frac{n}{2} \\
&> \frac{1}{2} \cdot 1 \cdot 1 \cdots \frac{n}{2} \\
&= \frac{n}{4}
\end{aligned}$$

   that is,

$$\begin{aligned}
\lim_{n\to\infty} \frac{f(n)}{g(n)} &= \lim_{n\to\infty} \frac{n!}{2^n} \\
&> \lim_{n\to\infty} \frac{n}{4} \\
&= +\infty
\end{aligned}$$

   Thus, $\lim_{n\to\infty} \frac{f(n)}{g(n)} = +\infty, f(n) \neq O(g(n))$

5 Probability and Random Variables

5.1 Probability [12.5 pts]

State true or false. Here Ω denotes the sample space and A^c denotes the complement of the event A.

1. For any $A, B \subseteq \Omega$, $P(A|B)P(A) = P(B|A)P(B)$.
   False, should be $P(B \mid A)P(A) = P(A \mid B)P(B)$

2. For any $A, B \subseteq \Omega$, $P(A \cup B) = P(A) + P(B) - P(B \cap A)$.
   True, by properties : $P(A \cup B) = P(A) + P(B) - P(B \cap A)$

3. For any $A, B, C \subseteq \Omega$ such that $P(B \cup C) > 0$, $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq P(A|B \cup C)P(B)$.
   True, we know that $P(A \cup B \cup C) \geqslant P(B \cup C)$, that is, $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geqslant 1$. Then, $P(A \mid B \cup C) \leqslant 1$ and $P(B) \leqslant 1$, that is, $P(A \mid B \cup C)P(B) \leqslant 1$.
   Thus, we have $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geqslant P(A \mid B \cup C) \cdot P(B)$ for any $A, B, C \subseteq \Omega$ such that $P(B \cup C) > 0$.

4. For any $A, B \subseteq \Omega$ such that $P(B) > 0, P(A^c) > 0$, $P(B|A^C) + P(B|A) = 1$.
   False, should be $P\left(A^c \mid B\right) + P(A \mid B) = 1$

5. If A and B are independent events, then A^c and B^c are independent.
   True, we know $P(A \cap B) = P(A)P(B)$ if A and B are independent events
   $$\begin{aligned} P\left(A^c\right) \cdot P\left(B^c\right) &= (1 - P(A))(1 - P(B)) \\ &= 1 - (P(A) + P(B)) + P(A)P(B) \\ &= 1 + P(A \cap B) - (P(A) + P(B)) \\ &= 1 - P(A \cup B) \\ &= P\left(A^c \cap B^c\right) \end{aligned}$$
   Thus, A^c and B^c are independent if A, B are independent

5.2 Discrete and Continuous Distributions [12.5 pts]

Match the distribution name to its probability density / mass function. Below, $|\boldsymbol{x}| = k$.

(a) Gamma (j)
(b) Multinomial (i)
(c) Laplace (h)
(d) Poisson (l)
(e) Dirichlet (k)

(f) $f(\boldsymbol{x}; \boldsymbol{\Sigma}, \boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$

(g) $f(x; n, \alpha) = \binom{n}{x} \alpha^x (1 - \alpha)^{n-x}$ for $x \in \{0, \dots, n\}$; 0 otherwise

(h) $f(x; b, \mu) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$

(i) $f(\boldsymbol{x}; n, \boldsymbol{\alpha}) = \frac{n!}{\Pi_{i=1}^k x_i!} \Pi_{i=1}^k \alpha_i^{x_i}$ for $x_i \in \{0, \dots, n\}$ and $\sum_{i=1}^k x_i = n$; 0 otherwise

(j) $f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x \in (0, +\infty)$; 0 otherwise

(k) $f(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$ for $x_i \in (0, 1)$ and $\sum_{i=1}^k x_i = 1$; 0 otherwise

(l) $f(x; \lambda) = \lambda^x \frac{e^{-\lambda}}{x!}$ for all $x \in Z^+$; 0 otherwise

5.3 Mean and Variance [10 pts]

1. Consider a random variable which follows a Binomial distribution: $X \sim \text{Binomial}(n, p)$.

   (a) What is the mean of the random variable?

$$\begin{aligned}
\mu &= E(x) \\
&= \sum x \cdot p(x) \\
&= \sum x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum x \cdot \frac{n!}{x(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum \frac{n(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \cdot p \\
&= np \sum \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\
&= np \sum \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{x-1} (1-p)^{(n-1)-(x-1)} \\
&= np \sum \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} \\
&= np \cdot (p+1-p)^{n-1} \\
&= np
\end{aligned}$$

(b) What is the variance of the random variable?

$$\begin{aligned}
\operatorname{Var}(x) &= E\left(x^2\right) - E(x)^2 \\
&= n(n-1)p^2 + np - n^2p^2 \\
&= np(1-p)
\end{aligned}$$

2. Let X be a random variable and $\mathbb{E}[X] = 1$, $\operatorname{Var}(X) = 1$. Compute the following values:

   (a) $\mathbb{E}[5X]$
   $E(5X) = 5E(X) = 5$

   (b) $\operatorname{Var}(5X)$
   $\operatorname{Var}(5X) = 5^2 \operatorname{Var}(X) = 25$

   (c) $\operatorname{Var}(X+5)$
   $\operatorname{Var}(X+5) = 1^2 \operatorname{Var}(X) = 1$

5.4 Mutual and Conditional Independence [12 pts]

1. (3 pts) If X and Y are independent random variables, show that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

$$\begin{aligned}
E(XY) &= \sum p(xy)xy \\
&= \sum p(x)p(y)xy \\
&= \sum xp(x) \cdot \sum yp(y) \\
&= E(X) \cdot E(Y)
\end{aligned}$$

2. (3 pts) If X and Y are independent random variables, show that $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.
Hint: $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + 2\operatorname{Cov}(X,Y) + \operatorname{Var}(Y)$

$$\begin{aligned}
\operatorname{Var}(X+Y) &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y) \\
&= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2 \cdot 0 \\
&= \operatorname{Var}(X) + \operatorname{Var}(Y)
\end{aligned}$$

3. (6 pts) If we roll two dice that behave independently of each other, will the result of the first die tell us something about the result of the second die?

   No, since they are two independent trials, the result of the first die will not affect the second die.

   If, however, the first die's result is a 1, and someone tells you about a third event — that the sum of the two results is even — then given this information is the result of the second die independent of the first die?

   No, since the result of first dice will depend the distribution of the result of second dice, they are conditional.

5.5 Central Limit Theorem [3 pts]

Prove the following result.

1. Let $X_i \sim \mathcal{N}(0,1)$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then the distribution of $\bar{X}$ satisfies

$$\sqrt{n}\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

By the Central Limit Theorem, we suppose $\{X_1, \dots, X_n, \dots\}$ is a sequence of i.i.d. random variables, then we have,

$$\mathbb{E}\left[X_i\right] = \mu \tag{6}$$

$$\operatorname{Var}\left[X_i\right] = \sigma^2 < +\infty \tag{7}$$

the random variables $\sqrt{n}\left(\bar{X}_n - \mu\right)$ converge in distribution to a normal $\mathcal{N}\left(0, \sigma^2\right)$ as $n \to +\infty$. In this case, $\mu = 0$ and $\sigma^2 = 1$, then we have $\sqrt{n}\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$ as $n \to +\infty$.

6 Linear algebra

6.1 Norms [5 pts]

Draw the regions corresponding to vectors $\mathbf{x} \in \mathbb{R}^2$ with the following norms:

1. $||\mathbf{x}||_1 \leq 1$ (Recall that $||\mathbf{x}||_1 = \sum_i |x_i|$)

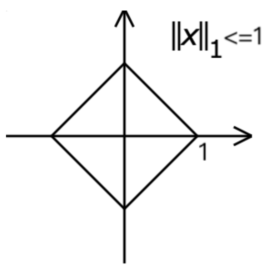

Figure 1

2. $||\mathbf{x}||_2 \leq 1$ (Recall that $||\mathbf{x}||_2 = \sqrt{\sum_i x_i^2}$)

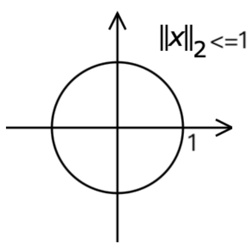

Figure 2

3. $||\mathbf{x}||_\infty \leq 1$ (Recall that $||\mathbf{x}||_\infty = \max_i |x_i|$)

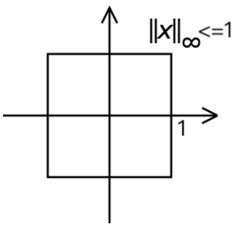

Figure 3

For $M = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 3 \end{pmatrix}$, Calculate the following norms.

4. $||M||_2$ (L2 norm)
   $\|M\|_2 =$ largest eigenvalue of $M^\top M^{\frac{1}{2}} = 7$

5. $||M||_F$ (Frobenius norm)
   $\|M\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} M_{ij}} = \sqrt{5^2 + 7^2 + 3^2} = \sqrt{83}$

6.2 Geometry [10 pts]

Prove the following. Provide all steps.

1. The smallest Euclidean distance from the origin to some point $\mathbf{x}$ in the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is $\frac{|b|}{||\mathbf{w}||_2}$. You may assume $\mathbf{w} \neq 0$.

$$
\begin{aligned}
d &= \|proj_w\,(x_0 - x)\|_2 \\
&= \left\|(x_0 - x) \cdot \frac{w^\top w}{\|w\|_2^2}\right\|_2 \\
&= \left\|(x_0 - x) \cdot w^\top\right\|_2 \cdot \frac{\|w\|_2}{\|w\|_2^2} \\
&= \left\|(x_0 - x) \cdot w^\top\right\|_2 \frac{1}{\|w\|_2} \\
&= \left\|w^\top x_0 - w^\top x\right\|_2 \frac{1}{\|w\|_2} \\
&= \|0 - (-b)\|_2 \frac{1}{\|w\|_2} \\
&= \frac{b}{\|w\|_2}
\end{aligned}
$$

2. The Euclidean distance between two parallel hyperplane $\mathbf{w}^T\mathbf{x} + b_1 = 0$ and $\mathbf{w}^T\mathbf{x} + b_2 = 0$ is $\frac{|b_1 - b_2|}{||\mathbf{w}||_2}$ (Hint: you can use the result from the last question to help you prove this one).

   Let x_0 be the point on $\omega^\top x + b_1 = 0$,

$$w^\top x_0 + b_1 = 0$$

$$
\begin{aligned}
d &= \|proj_w\,(x_0 - x)\|_2 \\
&= \left\|(x_0 - x)\, w^\top\right\|_2 \frac{1}{\|w\|_2} \\
&= \left\|w^\top x_0 - w^\top x\right\|_2 \frac{1}{\|w\|_2} \\
&= \|-b_1 + b_2\|_2 \frac{1}{\|w\|_2} \\
&= \frac{|b_1 - b_2|}{\|w\|_2}
\end{aligned}
$$

7 Programming Skills [10 pts]

Sampling from a distribution. For each question, submit a scatter plot (you will have 2 plots in total). Make sure the axes for all plots have the same ranges.

1. Make a scatter plot by drawing 100 items from a two dimensional Gaussian $N((1,-1)^T, 2I)$, where I is an identity matrix in $\mathbb{R}^{2\times 2}$.

```
import numpy as np
import matplotlib.pyplot as plt
```

```
# <1>
mean_x = np.array([1, -1]).T
cov_x = np.array([[2, 0], [0, 2]]) #y=1
num_x = 100
```

```
rand_x = np.random.multivariate_normal(mean_x, cov_x, num_x)
```

```
plt.scatter(x=rand_x[:,0], y=rand_x[:,1], s=2) # label
plt.legend()
```

```
WARNING:matplotlib.legend:No handles with labels found to put in legend.
<matplotlib.legend.Legend at 0x7fb498d80040>
```

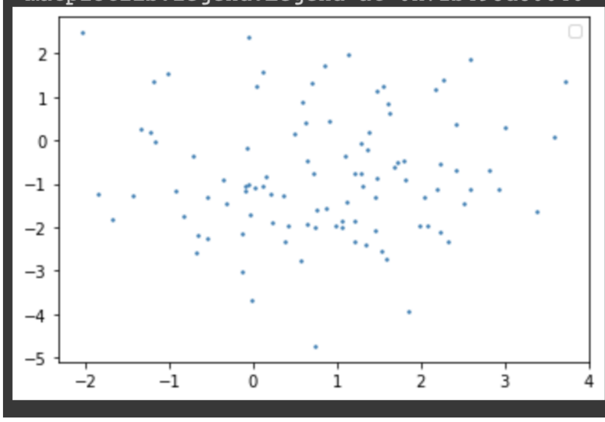

Figure 4

2. Make a scatter plot by drawing 100 items from a mixture distribution $0.3N\left((5,0)^T, \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}\right) + 0.7N\left((-5,0)^T, \begin{pmatrix} 1 & -0.25 \\ -0.25 & 1 \end{pmatrix}\right)$.

```python
# Y = 0
mean0 = np.array([5, 0]).T
cov0 = np.array([[1, 0.25], [0.25, 1]]) #y=1
mean1 = np.array([-5, 0]).T
cov1 = np.array([[1, -0.25], [-0.25, 1]]) # realization
```

```python
rand0 = np.random.multivariate_normal(mean0, cov0, 30)
rand1 = np.random.multivariate_normal(mean1, cov1, 70)
```

```python
#Y=0
plt.scatter(x=rand0[:,0], y=rand0[:,1], s=2, label="Y = 0") #Y=1
plt.scatter(x=rand1[:,0], y=rand1[:,1], s=2, label="Y = 1") # label
plt.legend()
```

```
<matplotlib.legend.Legend at 0x7fb499ec55e0>
```

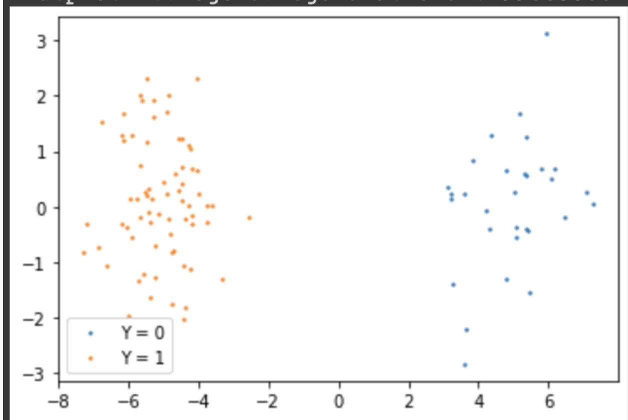

Figure 5