Homework Assignment # 1

Due: Thursday, September 26, 2019, 11:59 p.m.
Total marks: 100

Question 1. [10 MARKS]

Let X be a random variable with outcome space $\Omega = \{a, b, c\}$ and $p(a) = 0.1, p(b) = 0.2$, and $p(c) = 0.7$. Let

$$f(x) = \begin{cases} 10 & \text{if } x = a \\ 5 & \text{if } x = b \\ 10/7 & \text{if } x = c \end{cases}$$

(a) [3 MARKS] What is $E[f(X)]$?

$$E[f(x)] = 10 * 0.1 + 5 * 0.2 + 10/7 * 7/10 = 3$$

(b) [3 MARKS] What is $E[1/p(X)]$?

$$E[1/p(x)] = 0.1 * 1/0.1 + 0.2 * 1/0.2 + 0.7 * 1/0.7 = 3$$

(c) [4 MARKS] For an arbitrary pmf p, what is $E[1/p(X)]$?

$$E[1/p(X)] = \sum_{x \in X} p(x) * 1/p(x) = 3$$

Question 2. [15 MARKS]

Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be independent multivariate Gaussian random variables, with $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, with $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d \times d}$ for dimension $d \in \mathbb{N}$. Define $\mathbf{X} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_m\mathbf{X}_m$ as a convex combination, $a_i \geq 0$ and $\sum_{i=1}^m a_i = 1$.

(a) [5 MARKS] Write the expected value $E[\mathbf{X}]$ in terms of the givens $a_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$. Show all you steps. What is the dimension of $E[\mathbf{X}]$?

$$\begin{aligned} E[\mathbf{X}] &= E[a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \cdots + a_m\mathbf{X}_m] \\ &= a_1E[\mathbf{X}_1] + a_2E[\mathbf{X}_2] + \cdots + a_mE[\mathbf{X}_m] \\ &= \sum_{i=1}^m a_i * \boldsymbol{\mu}_i \in \mathbb{R}^d \end{aligned}$$

(b) [10 MARKS] Write the covariance $\text{Cov}[\mathbf{X}]$ in terms of the givens $a_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$. Show all you steps. What is the dimension of $\text{Cov}[\mathbf{X}]$? Briefly explain how the result for $\text{Cov}[\mathbf{X}]$ would be different if the variables $\mathbf{X}_1$ and $\mathbf{X}_2$ are not independent and have covariance $\text{Cov}[\mathbf{X}_1, \mathbf{X}_2] = \boldsymbol{\Lambda}$ for $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$.

By definition, we have:

$$\text{Cov}[X] = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T]$$

$$= E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}]^T$$

As the result is a $d \times d$ matrix, we use i, j to denote the element in the ith row and the jth column. We use $\mathbf{X}_{i,j}$ to denote the jth element in $\mathbf{X}_i$ and $\mathbf{X}_i$ to denote ith element in $\mathbf{X}$:

$$\begin{aligned}
\text{Cov}[\mathbf{X}]_{i,j} &= E[\mathbf{X}\mathbf{X}^T]_{i,j} - E[\mathbf{X}]E[\mathbf{X}]^T{}_{i,j} \\
&= E[\mathbf{X}_i\mathbf{X}_j] - E[\mathbf{X}_i]E[\mathbf{X}_j] \\
&= E[\sum_{t=1}^m a_t\mathbf{X}_{t,i}\sum_{t=1}^m a_t\mathbf{X}_{t,j}] - E[\sum_{t=1}^m a_t\mathbf{X}_{t,i}]E[\sum_{t=1}^m a_t\mathbf{X}_{t,j}]
\end{aligned}$$

It's obvious that, for any $t \neq t'$, $E[a_t\mathbf{X}_{t,i}a_{t'}\mathbf{X}_{t',i}] = E[a_t\mathbf{X}_{t,i}]E[a_{t'}\mathbf{X}_{t',i}]$ due to the independence. In this case, the above formula can be simplified as:

$$\begin{aligned}
\text{Cov}[\mathbf{X}]_{i,j} &= \sum_{t=1}^m a_t^2 E[\mathbf{X}_{t,i}\mathbf{X}_{t,j}] - \sum_{t=1}^m a_t^2 E[\mathbf{X}_{t,i}]E[\mathbf{X}_{t,j}] \\
&= \sum_{t=1}^m a_t^2 \left[E[\mathbf{X}_{t,i}\mathbf{X}_{t,j}] - E[\mathbf{X}_{t,i}]E[\mathbf{X}_{t,j}]\right] \\
&= \sum_{t=1}^m a_t^2 \mathbf{\Sigma}_{t,i,j}
\end{aligned}$$

The full matrix can be written as:

$$\text{Cov}(\mathbf{X}) = \begin{bmatrix}
\sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,1,1} & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,1,2} & \dots & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,1,d} \\
\sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,2,1} & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,2,2} & \dots & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,2,d} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,d,1} & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,d,2} & \dots & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,d,d}
\end{bmatrix} \in \mathbb{R}^{d\times d}$$

In this case, the above simplication step would be different

$$\begin{aligned}
\text{Cov}[\mathbf{X}]_{i,j} &= \sum_{t=1}^m a_t^2 E[\mathbf{X}_{t,i}\mathbf{X}_{t,j}] - \sum_{t=1}^m a_t^2 E[\mathbf{X}_{t,i}]E[\mathbf{X}_{t,j}] \\
&+ a_1a_2E[\mathbf{X}_{1,i}\mathbf{X}_{2,j}] - a_1a_2E[\mathbf{X}_{1,i}]E[\mathbf{X}_{2,j}] + a_2a_1E[\mathbf{X}_{2,i}\mathbf{X}_{1,j}] - a_2a_1E[\mathbf{X}_{2,i}]E[\mathbf{X}_{1,j}] \\
&= \sum_{t=1}^m a_t^2\left[E[\mathbf{X}_{t,i}\mathbf{X}_{t,j}] - E[\mathbf{X}_{t,i}]E[\mathbf{X}_{t,j}]\right] + 2a_1a_2\left[E[\mathbf{X}_{1,i}\mathbf{X}_{2,j}] - E[\mathbf{X}_{1,i}]E[\mathbf{X}_{2,j}]\right] \\
&= \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,i,j} + 2a_1a_2\mathbf{\Lambda}_{i,j}
\end{aligned}$$

Thus, the full matrix would be:

$$\text{Cov}(\mathbf{X}) = \begin{bmatrix}
\sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,1,1} + 2a_1a_2\mathbf{\Lambda}_{1,1} & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,1,2} + 2a_1a_2\mathbf{\Lambda}_{1,2} & \dots & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,1,d} + 2a_1a_2\mathbf{\Lambda}_{1,d} \\
\sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,2,1} + 2a_1a_2\mathbf{\Lambda}_{2,1} & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,2,2} + 2a_1a_2\mathbf{\Lambda}_{2,2} & \dots & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,2,d} + 2a_1a_2\mathbf{\Lambda}_{2,d} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,d,1} + 2a_1a_2\mathbf{\Lambda}_{d,1} & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,d,2} + 2a_1a_2\mathbf{\Lambda}_{d,2} & \dots & \sum_{t=1}^m a_t^2\mathbf{\Sigma}_{t,d,d} + 2a_1a_2\mathbf{\Lambda}_{d,d}
\end{bmatrix}$$

Question 3. [15 MARKS]

This question involves some simple simulations, to better visualize random variables and get some intuition for sampling, which is a central theme in machine learning. Use the attached code called `simulate.py`. This code is a simple script for sampling and plotting with python; play with some of the parameters to see what it is doing. Calling `simulate.py` runs with default parameters; `simulate.py 1 100` simulates 100 samples from a 1d Gaussian. The generated plot is generically for 3 dimensions. If you call the function with 1d, then it simply plots the points on a line, but on a 3-dimensional plot. The maximum dimension that can be given to the script is 3.

Note that if you do not have matplotlib installed, you will have to install it.

(a) [5 MARKS] Run the code for 10, 100 and 1000 samples with dim=1 and $\sigma = 1.0$. Next run the code for 10, 100 and 1000 samples with dim=1 and $\sigma = 10.0$. What do you notice about the sample mean?

	σ=1.0	σ=10.0
samples=10	0.272	-6.109
samples=100	0.114	1.484
samples=1000	0.027	0.771

We can notice that as the sample number gets larger, the sample mean gets closer to the true mean. As the variance gets larger, the sample mean deviates more from the true mean. Such result is supported by theoretic result that the sample mean is an unbiased estimator of the true mean and the variance of sample mean is σ^2/n where n is the number of samples.

(b) [5 MARKS] The current covariance for dim=3 is

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

What does that mean about the multivariate Gaussian (i.e., the vector random variable composed of random variables X, Y and Z)?

Each of the random variable has variance 1, but their covariance is 0, which means any two of the random variables are uncorrelated.

(c) [5 MARKS] Change the covariance to

$$\Sigma = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

What happens?

In this case, random variable X, Z are correlated. More specifically, they are now linearly associated to some degree.

Question 4. [30 MARKS]

Suppose that the number of accidents occurring daily in a certain plant has a Poisson distribution with an unknown mean λ. Based on previous experience in similar industrial plants, suppose that our initial feelings about the possible value of λ can be expressed by an exponential distribution with parameter $\theta = \frac{1}{2}$. That is, the prior density is

$$p(\lambda) = \theta \mathrm{e}^{-\theta\lambda}$$

where $\lambda \in [0, \infty)$.

(a) [5 MARKS] Before observing any data (any reported accidents), what is the most likely value for λ?

$$\begin{aligned}
\lambda &= \arg\max_{\lambda} p(\lambda) \\
\lambda &= \arg\max_{\lambda} \frac{1}{2} e^{-\lambda/2} \\
\lambda &= 0
\end{aligned}$$

(b) [5 MARKS] Now imagine there are 79 accidents over 9 days. Determine the maximum likelihood estimate of λ.

Assume the data obtained is i.i.d and denote the dataset as $\mathcal{D}$ and each data point as x_i:

$$p(\mathcal{D}|\lambda) = \prod_{i=1}^{m} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

The log-likelihood of p can be written as:

$$\begin{aligned}
\ln p(\mathcal{D}|\lambda) &= \sum_{i=1}^{m} \ln \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\
&= \sum_{i=1}^{m} x_i \ln \lambda - \lambda - \ln x_i!
\end{aligned}$$

$\lambda_{MLE} = \arg\max_\lambda \ln p(\mathcal{D}|\lambda)$ and the maximum of $\ln p(\mathcal{D}|\lambda)$ can be computed by taking derivatives w.r.t λ:

$$\frac{d}{d\lambda} \ln p(\mathcal{D}|\lambda) = \frac{\sum_{i=1}^{m} x_i}{\lambda} - \sum_{i=1}^{m} 1$$

As we already know $\sum_{i=1}^{m} x_i = 79$ and $m = 9$, λ can be computed by setting the derivative as 0:

$$\begin{aligned}
\frac{\sum_{i=1}^{m} x_i}{\lambda} - \sum_{i=1}^{m} 1 = 0 \\
\rightarrow \lambda_{MLE} = 79/9
\end{aligned}$$

(c) [5 MARKS] Again imagine there are 79 accidents over 9 days. Determine the maximum a posteriori (MAP) estimate of λ.

Similar to the procedures in (b) with the same notation and assumption, the log-likelihood of p can be written as:

$$\begin{aligned}\ln p(\mathcal{D}|\lambda)p(\lambda) &= \ln p(\mathcal{D}|\lambda) + \ln p(\lambda) \\ &= \sum_{i=1}^{m} x_i \ln \lambda - \lambda - \ln x_i! + \ln \theta - \theta\lambda\end{aligned}$$

$\lambda_{MAP} = \arg\max_{\lambda} \ln p(\mathcal{D}|\lambda)p(\lambda)$ and the maximum can be computed by taking partial derivatives w.r.t λ:

$$\frac{\partial}{\partial \lambda} \ln p(\mathcal{D}|\lambda)p(\lambda) = \frac{\sum_{i=1}^{m} x_i}{\lambda} - \sum_{i=1}^{m} 1 - \theta$$

As we already know $\sum_{i=1}^{m} x_i = 79$, $m = 9$ and $\theta = 1/2$, λ can be computed by setting the derivative as 0:

$$\begin{aligned}\frac{\sum_{i=1}^{m} x_i}{\lambda} - \sum_{i=1}^{m} 1 - \theta = 0 \\ \rightarrow \lambda_{MAP} = 158/19\end{aligned}$$

(d) [5 MARKS] Imagine you now want to predict the number of accidents for tomorrow. How can you use the maximum likelihood estimate computed above? What about the MAP estimate? What would they predict?

I would use the mode of the distribution, which is the most likely value to predict. For Poisson distribution, the mode is the largest integer less than or equal to λ. Thus, the result predicted by MLE would be 8 and the result predicted by MAP would be 8 too.

(e) [5 MARKS] For the MAP estimate, what is the purpose of the prior once we observe this data?

The purpose is to correct the variance and bias caused by limited number of observations and partial observability. For example, if we flip a coin and observe a head, from MLE, we will obtain the result that this coin is 100% biased to head. However, we know this estimate is unlikely to be correct because a single data point has high variance. For MAP, with prior that the coin is totally unbiased, observing the previous event would only shift the estimate slightly toward that the coin is biased to head which is more reasonable.

(f) [5 MARKS] Imagine that now new safety measures have been put in place and you believe that the number of accidents per day should sharply decrease. How might you change θ to better reflect this new belief about the number of accidents? *Hint:* Look at the plots of some exponential distributions to better understand the prior chosen on λ.

I would increase the value of θ because this would make small numbers even more likely for the prior distribution of λ.

Question 5. [30 MARKS]

Imagine that you would like to predict if your favorite table will be free at your favorite restaurant. The only additional piece of information you can collect, however, is if it is sunny or not sunny. You collect paired samples from visit of the form (is sunny, is table free), where it is either sunny (1) or not sunny (0) and the table is either free (1) or not free(0).

(a) [10 MARKS] How can this be formulated as a maximum likelihood problem?

In this problem, we have two random variables, they are X={sunny, not sunny} and Y={free, not free}. Both the random variables are binary and discrete. The number of outcomes of the two random variable is 4, which can be well modeled by a multinomial distribution f. If we denote the parameter of a multinomial distribution as p, p would have 4 components $p = \{p_1, p_2, p_3, p_4 | \sum_{i=1}^{4} p_i = 1\}$ and each represents the probability of data in each class.

If we denote our dataset as $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, by the definition of maximum likelihood estimation, we can write:

$$p_{MLE} = \arg\max_p f(\mathcal{D}|p)$$

With the assumption that the data obtained is i.i.d, we can obtain:

$$p_{MLE} = \arg\max_p \prod_{i=1}^{n} f((x_i, y_i)|p)$$

By applying the log-likelihood trick, we can obtain:

$$p_{MLE} = \arg\max_p \sum_{i=1}^{n} \ln f((x_i, y_i)|p)$$

p_{MLE} here can be solved by taking derivatives to find the maximum.

(b) [10 MARKS] Assume you have collected data for the last 10 days and computed the maximum likelihood solution to the problem formulated in (a). If it is sunny today, how would you predict if your table will be free?

In this case, we would have a joint distribution w.r.t the learned parameter p_{MLE}:

$$P(X, Y|p_{MLE})$$

where X={sunny, not sunny} and Y={free, not free}.

To predict whether the table is free or not today, we need to compute the conditional distribution of Y given X:

$$P(Y|X; p_{MLE})$$

By plugging in that X=sunny, we will know the probability of the table is free or not and we can pick the one with higher probability for prediction:

$$P(Y|X\text{=sunny}; p_{MLE})$$

(c) [10 MARKS] Imagine now that you could further gather information about if it is morning, afternoon, or evening. How does this change the maximum likelihood problem?

We denote Z = {morning, afternoon, evening}. First, this will change our data in $\mathcal{D}$ from 2-dimensional to 3-dimensional. Second, the number of all outcomes of X, Y, Z will be 12 and we need to model the distribution with a 12-dimensional multinomial distribution. Last, the joint distribution learned will be $P(X, Y, Z|p_{MLE})$ rather than $P(X, Y|p_{MLE})$

Bonus. [20 MARKS]

(d) [10 MARKS] Using a computer, generate 1000 samples from a d-dimensional multivariate Gaussian with mean **0** and identity covariance matrix. Compute the average ℓ_2 distance of each

Dimension	1	2	4	8	16	32	64	128	256
Distance	0.79	1.26	1.88	2.75	3.93	5.61	7.97	11.29	15.98

sample to the origin. Repeat this experiment for $d \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$. What happens to the average distance as d increases? For k-means clustering, what is one implication of this outcome?

From the table, we can conclude that the average distance increases as the dimension increases. This implies that for k-means clustering, l_2 distance may not be a good matric for high-dimensional data.

(e) [10 MARKS] Consider two d-dimensional hypercubes centered at the origin. The first has a side length of 1 and the second has a side length of 1 - ϵ ($0 < \epsilon < 1$). Give an expression for the ratio of the volume of the second hypercube to the volume of the first (in terms of d and ϵ). What happens as d gets large? How does this help explain the result about average distances from the previous question?

The ratio of the two hypercubes is: $(1 - \epsilon)^d$. This implies even though ϵ can be close to 0, as the dimension d approaches infinity, the ratio approaches 0. This implies for 2 data points, no matter how close they are, in high-dimensional space, they might be grouped into different clusters by k-means clustering. This again, explains the k-means clustering may not be suitable for high-dimensional data.