Exam #2

Instructions. This is a 150-minute test. You may use your notes. You may assume anything that we proved in class or in the homework is true.

Question	Score	Points
1		10
2		10
3		10
4		10
5		10
Out Of		50

Name: ______________________________

edX Username: ______________________________

1. **[Basic Probability]**

   Consider a joint distribution on X, Y, with $\text{Prob}(X = i, Y = j) = p_{ij}$, where $X \in \{1, 2\}$ and $Y \in \{1, 2, 3\}$. This is summarized in the following table:

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	p_{11}	p_{12}	p_{13}
$X = 2$	p_{21}	p_{22}	p_{23}

   Assume p_{ij} are all positive. In the following, write your answers in terms of elements of the joint distribution above.

   (a) Calculate the marginal distribution $\text{Prob}(X = 2)$ and $\text{Prob}(Y = 1)$.

   (b) Calculate $\text{Prob}(Y = 1 \mid X = 2)$.

   (c) Calculate the probability $\text{Prob}(X < Y)$, where $X < Y$ is the event that the value of X is smaller than the value of Y.

2. **[MLE and Bayesian Inference]**

   Every time when we go to Starbucks, we join a line with a number of people ahead of us. Let us build a probabilistic model to estimate the waiting time.

   From queueing theory, scientists have found that when there are k people ahead of us (k is a positive integer), the waiting time X follows a Gamma distribution, denoted by $\mathbf{Gamma}(k, \theta)$, whose density function is defined as follows:

$$p(x \mid \theta;\ k) = \frac{1}{\Gamma(k)} \times \theta^k x^{k-1} \exp\left(-\theta x\right), \qquad \forall x \in (0, \infty),$$

   where θ is a positive unknown parameter and $\Gamma(k)$ is the so called Gamma function, defined by an integration:

$$\Gamma(k) = \int_0^\infty z^{k-1} \exp(-z) dz.$$

   We want to estimate θ, because once we know θ, we would know the distribution of the waiting time when there are k people ahead. This would allow us to make prediction about the waiting time.

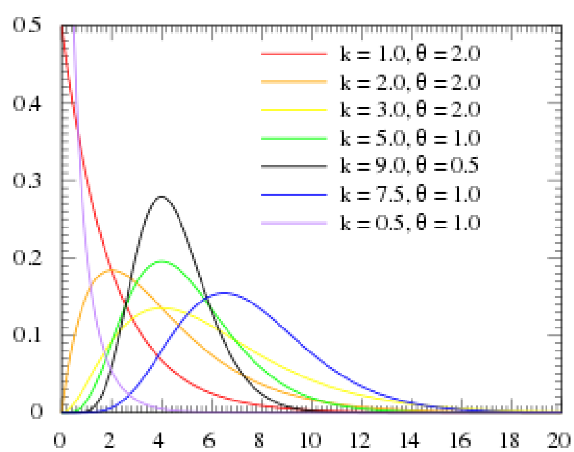

Figure 1: Examples of density functions of Gamma distributions with different parameters.

(a) Assume we went to the store for n times; at the i-th time, there were k_i people ahead and the waiting time was x_i. Assume $\{k_i, x_i\}_{i=1}^{n}$ are *i.i.d.* for different i. Please write down the likelihood function of θ based on those observations. Show your work.

(b) Please estimate θ with MLE based on $\{k_i, x_i\}_{i=1}^n$. So you just need to maximize the likelihood you get in (a). Please show your derivation and result.

(c) Let us consider the Bayesian approach now. Assume the prior of θ is $\mathbf{Gamma}(k_0, x_0)$, where k_0 and x_0 are fixed and known numbers. Please derive the posterior distribution $p(\theta \mid \{k_i, x_i\}_{i=1}^n)$. (*Hint: the posterior distribution is also a Gamma distribution, say* $\mathbf{Gamma}(k_{post}, x_{post})$*; please decide the value of* k_{post} *and* x_{post})

3. [**Multivariate Gaussian**]

Assume we have the following three dimensional normal random variable

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 3 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix} \right).$$

It will be useful to know that the inverse matrix of $\begin{bmatrix} 3 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix}$ is $\begin{bmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & 5/2 & 3/2 \\ 1/2 & 3/2 & 3/2 \end{bmatrix}$.

The inverse of $\begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 3/2 \end{bmatrix}$ is $\begin{bmatrix} 1 & -1 \\ -1 & 5/3 \end{bmatrix}$.

(a) Which two variables are independent with each other?

(b) Define

$$Z = X_1 - aX_2 + bX_3, \tag{1}$$

where $a, b \in \mathbb{R}$ are two constants. Is it possible to set the values of a and b such that Z is independent with X_3 (that is, $Z \perp X_3$)? If so, give an example of such a and b.

(c) For $Z = X_1 - aX_2 + bX_3$, is it possible to set the value of a and b such that Z is independent with X_2 *conditional* on $X_1 = x_1$ (i.e., $Z \perp X_2 \mid X_1 = x_1$), for any fixed value $x_1 \in \mathbb{R}$? In other word, we hope to make X_2 and Z independent conditional on that X_1 equals to a fixed number x_1, regardless what the value of x_1 is. If this can be done, give an example of (a, b) that satisfy the condition.

4. **[Clustering, K-means]**

We want to cluster the following dataset into $K = 3$ clusters using the K-means algorithm:

$$\begin{aligned} x^{(1)} &= 4, \\ x^{(2)} &= 10, \\ x^{(3)} &= 16, \\ x^{(4)} &= 20, \\ x^{(5)} &= 26, \end{aligned}$$

where each $x^{(i)}$ is an one-dimensional data point.

(a) Please analyze how K-means updates the centroids of the clusters if we initialize the centroids of the $K = 3$ clusters by: $\mu_1 = 2$, $\mu_2 = 3$, and $\mu_3 = 4$. What values would the centroids (μ_1, μ_2, μ_3) converge to when K-means determines? Please show the centroid locations at each iteration of K-means. (*If no points are assigned to a cluster at a given iteration, do* **NOT** *update its centroid*).

(b) Is the solution unique regardless of the initialization? If not, show an example in which the final clustering is different from what the K-means algorithm estimated in (a).

5. **[General Knowledge]**

Please decide if the following statements are true. You can either provide a binary decision of 1 or 0, or, if you are uncertain, give a probabilisitic estimation in the interval $[0, 1]$. An answer of 0 corresponds to deciding the statement is false and, conversely, an answer of 1 corresponds to deciding the statement is true. Assume your estimation is q, then you will get $q \times 100\%$ credit if the statement is correct, and $(1 - q) \times 100\%$ if the statement is wrong. **Your answers should not be written as 'true' or 'false'**, but instead be in the form of a number in the interval $[0, 1]$. **Grading will be based on the number you provide**.

> **Example:** $1 + 1 = 2$ (Answer: 0.8)
> *[You will get 0.8 of the credit since the statement is true.]*
>
> **Example:** $1 + 1 = 4$ (Answer: 0.3)
> *[You will get $1 - 0.3 = 0.7$ of the credit since the statement is false.]*
>
> **Example:** $1 + 1 = 3$ (Answer: 0.8)
> *[You will get $1 - 0.8 = 0.2$ of the credit since the statement is false.]*

(a) Any random variables X_1 and X_2 are independent if they are uncorrelated.
(Answer:_______)

(b) The goal for Bayesian inference is to find a parameter that maximize the posterior.

(Answer:_______)

(c) Assume the prior distribution of a parameter is Gaussian, then its posterior distribution is always Gaussian.

(Answer:_______)

(d) EM algorithm is equivalent to coordinate ascend on a tight lower bound of the marginal likelihood function, so the objective will monotonically decrease and converge to global optimal.

(Answer:_______)

(e) K-means guarantees to monotonically improve the loss function, and will converge within a *finite* number of steps.

(Answer:_______)

(f) Assume $Q = [q_{ij}]_{ij=1}^{d}$ is the inverse covariance matrix (i.e. precision matrix) of a multivariate normal random variable $X = (X_1, \ldots, X_d)$. Then $X_i \perp X_j$ if and only if $q_{ij} = 0$.

(Answer:_______)

(g) Kernel regression yields a non-convex optimization if we pick Gaussian radial basis function(RBF) kernel.

(Answer:_______)

(h) In kernel regression, if we use a kernel $k(x, x') = x^\top x' + 1$, we would obtain a linear function (i.e., it is effectively doing a linear regression).

(Answer:_______)

(i) Consider a simple neural network with two ReLU neurons:

$$f(x; [w_1, w_2]) = \max(0,\ x - w_1) + \max(0,\ x - w_2).$$

Then $f(x; [w_1, w_2])$ is a convex function of both x and $[w_1, w_2]$, but estimating $[w_1, w_2]$ by minimizing the mean square error (MSE) loss would yield a non-convex optimization on $[w_1, w_2]$.

(Answer:_______)