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. 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}

(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 that of Y.

2. 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 **Gamma**(k, θ), 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.

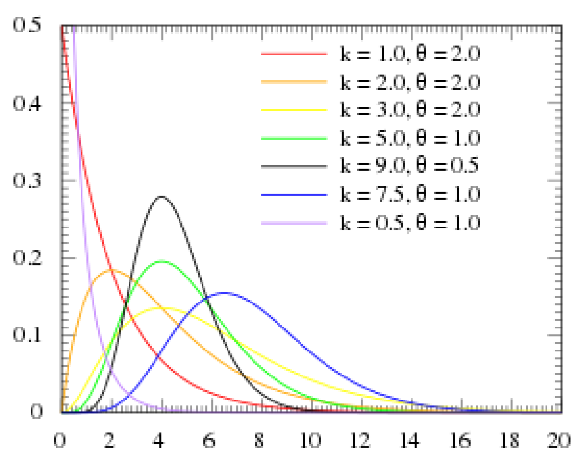

Figure 1: Examples of density functions of Gamma distributions with different parameters.

(a) Assume we went to a store once, and we found $k_1 = 5$ people ahead of us and the waiting time was x_1. Please estimate θ using maximum likelihood estimation (MLE) based on this information. Please show your derivation and result.

(b) 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 $\{k_i, x_i\}$ are independent for different i. Please estimate θ with MLE based on $\{k_i, x_i\}_{i=1}^{n}$. 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.*)

3. Assume we have the following three dimensional normal random variable

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 1 \\ -2 \end{bmatrix}, \quad \begin{bmatrix} 4 & 1 & -1 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \right).$$

It will be useful to know that the inverse matrix of $\begin{bmatrix} 4 & 1 & -1 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$ is $\begin{bmatrix} 1/2 & -1/2 & 1/2 \\ -1/2 & 3/2 & -1/2 \\ 1/2 & -1/2 & 3/2 \end{bmatrix}$.

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

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

(b) Define

$$Z = X_1 - \mathbf{p}^\top \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}, \tag{1}$$

where $\mathbf{p} \in \mathbb{R}^2$ is a deterministic 2×1 vector. Does there exist a $\mathbf{p}$ such that Z is independent with X_1? If so, give an example of $\mathbf{p}$. (Note that for two 2×1 vectors $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, $\mathbf{a}^\top \mathbf{b} = a_1 b_1 + a_2 b_2$.)

(c) Following the definition of Z in equation (1), does there exist a $\mathbf{p}$ such that Z is independent with X_1 *conditional* on $X_3 = x_3$ (i.e., $Z \perp X_1 \mid X_3 = x_3$) for any fixed value $x_3 \in \mathbb{R}$?

4. **[Clustering, K-means]** We want to cluster the following dataset into $K = 3$ clusters using the K-means algorithm:

$$x^{(1)} = 10,$$
$$x^{(2)} = 20,$$
$$x^{(3)} = 40,$$
$$x^{(4)} = 50,$$
$$x^{(5)} = 60,$$

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

(a) Initialize the centroids of the clusters as: $\mu_1 = 6$, $\mu_2 = 7$, and $\mu_3 = 8$. Where will the centroids (μ_1, μ_2, μ_3) converge to when K-means converges? 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 than what the K-means algorithm estimated in part (a).

5. Please decide if the following statements are true. You can either provide a binary decision of 1 (true) or 0 (false), or, if you are uncertain, give a probabilisitic estimation in interval $[0, 1]$. 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.

> **Example:** $1 + 1 = 2$ (Answer: 0.8)
> *[You will get 0.8 of the credit since the statement is true.]*
>
> **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 in 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 if we estimate $[w_1, w_2]$ by minimizing the mean square error (MSE) loss, we would have to solve a non-convex optimization on $[w_1, w_2]$.
(Answer:_______)

(j) Assume we train a neural network with one hidden layer consisting of 100 neurons. If we use *stochastic* gradient descent and initialize the weights of all the neurons to be the same value, then the weights of the different neurons will stay *the same* across the iterations (i.e., we effectively train a network with a single neuron).

(Answer:_______)