Exam #2

Instructions. This is a 120-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. [**Maximum Likelihood**] Given observations $\{x_i, y_i\}_{i=1}^n$, where both x_i and y_i are positive scalars, and each y_i is independently drawn from the following conditional density function:

$$p(y_i \mid x_i;\ \theta) = \frac{1}{\theta x_i} \exp\left(-\frac{y_i}{\theta x_i}\right),$$

where θ is an unknown positive parameter. You can think this as a *nonlinear regression* problem that seeks to fit a nonlinear relation between x_i and y_i.

   (a) Please write down the log-likelihood function as a function of θ.

(b) Please derive the maximum likelihood estimator of θ.

2. 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 $\text{Prob}(Y = 1)$.

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

(c) Decide if the following statements are correct or wrong

i. We must have $0 \leq p_{ij} \leq 1$ for all ij. **[Correct]** **[Wrong]**

ii. $\sum_{i=1}^{2} \text{Prob}(X = i) = 1$. **[Correct]** **[Wrong]**

iii. $\sum_{i=1}^{2} \text{Prob}(X = i \mid Y = 1) = 1$ **[Correct]** **[Wrong]**

iv. $\sum_{j=1}^{3} \text{Prob}(X = 1 \mid Y = j) = 1$ **[Correct]** **[Wrong]**

3. Assume $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is a two-dimensional standard normal random variable,

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

Let $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ be obtained by a linear transform of $\boldsymbol{X}$:

$$\begin{cases} Y_1 = 2X_1 + \rho X_2 \\ Y_2 = \ X_1 + \rho X_2, \end{cases}$$

where ρ is a real number constant.

(a) What is the distribution of $\boldsymbol{Y}$? Derive its mean and covariance matrix.

(b) Does there exist a value of ρ such that Y_1 and Y_2 are independent with each other? Please explain the reason.

4. Suppose that we are fitting a mixture of $K = 2$ Gaussians to the following dataset of scalar values:

$$x^{(1)} = 5,$$
$$x^{(2)} = 15,$$
$$x^{(3)} = 25,$$
$$x^{(4)} = 30,$$
$$x^{(5)} = 40.$$

We use the EM algorithm to find the maximum likelihood estimates for the model parameters; the parameters include the mixing proportions (α_1, α_2), the means (μ_1, μ_2), and the variances (σ_1^2, σ_2^2) of the two components.

(a) Suppose that at some point in the EM algorithm, the E-step found that the posterior probabilities of the two components for the five data items were as follows:

$x^{(i)}$	$\gamma_1^{(i)} := P(z^{(i)} = 1 \mid x^{(i)})$	$\gamma_2^{(i)} := P(z^{(i)} = 2 \mid x^{(i)})$
5	0.2	0.8
15	0.2	0.8
25	0.8	0.2
30	0.9	0.1
40	0.9	0.1

where we define $z^{(i)} \in \{1, 2\}$ to be the unobserved index that represents which Gaussian component $x^{(i)}$ is drawn from, and $\gamma_k^{(i)} = P(z^{(i)} := k \mid x^{(i)})$ the posterior probability that $x^{(i)}$ is drawn from the k-th component.

Please update the values of the parameters $\alpha_1, \alpha_2, \mu_1, \mu_2$ by performing one-step M-Step based on the posterior probabilities in the table. Show your calculation.

(b) Assume we initialize EM with the following parameters:

$$\mu_1 = \mu_2 = 0,$$
$$\sigma_1 = \sigma_2 = 1,$$
$$\alpha_1 = 0.2,\ \alpha_2 = 0.8.$$

Please describe the values of $[\mu_1, \mu_2, \sigma_1, \sigma_2, \alpha_1, \alpha_2]$ the EM algorithm converges to. Please also give the corresponding posterior probability $\gamma_1^{(i)} = p(z^{(i)} = 1 \mid x^{(i)})$ for $i = 1$ when EM converges. Please describe your reasoning concisely and clearly (*hint: the calculation is very easy once you understand the idea*).

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) K-means guarantees to monotonically improve the loss function, and find a global optima when it converges.

(Answer:_____)

(b) Both kernel regression and neural networks yield non-convex optimization.

(Answer:_____)

(c) Maximum likelihood estimation (MLE) yields convex optimization.

(Answer:_____)

(d) Bayesian inference is equivalent to maximum likelihood estimation (MLE) if the prior is a uniform distribution.

(Answer:_____)

(e) Assume the prior distribution of a parameter is Gaussian, then its posterior distribution is not always Gaussian.

(Answer:_____)

(f) Assume (X_1, X_2) is a multivariate normal random variable, then $X_1 \perp X_2$ if and only if $\text{cov}(X_1, X_2) = 0$.

(Answer:_____)

(g) Assume (X_1, X_2, X_3) is a multivariate normal random variable whose inverse covariance matrix (i.e. precision matrix) is

$$\begin{bmatrix} 3 & -1 & 0 \\ -1 & 5 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then its covariance graph and Markov graph happens to be the same.

(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) Neural network and kernel regression are guaranteed to be better than linear regression, because they can fit more flexible nonlinear curves.

(Answer:_____)

(j) Assume we train a neural network with 100 neurons. If we use deterministic 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). But if we apply stochastic gradient descent, the weights of different neurons will become different due to the randomness introduced by stochastic gradient descent, even if they were initialized to the same value.

(Answer:_____)