# 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** Assume $X$ is a discrete random variable that takes values in $\{0, 1\}$. Assume we know in prior that

$$P(X = k) = \alpha \exp(-k\lambda), \quad \forall k \in \{0, 1\},$$

where $\alpha$ and $\lambda$ are two parameters to be decided.

   (a) What constraints should we put on $\alpha$ and $\lambda$ to ensure that we have a valid distribution?

   (b) Assume we observe a sequence $D = \{x_1, x_2, \ldots, x_n\}$ that is drawn independently from the distribution of $X$; we assume the observed number of $0, 1$ in $D$ are $n_0$, $n_1$, respectively. Please write down the log-likelihood function as a function of $\lambda$ and propose a method to estimate $\lambda$ (it is enough to frame an optimization problem without numerically solving it).

2. **Bayesian Inference** Assume we use a medical device to detect a type of rare cancer. Denote by $X \in \{0, 1\}$ if the cancer actually exists on a patient and $Y \in \{0, 1\}$ the output of the medical device. Denote by the false negative and false positive of the device to be $\alpha$ and $\beta$, respectively, that is,

$$\alpha = P(Y = 0 \mid X = 1)$$
$$\beta = P(Y = 1 \mid X = 0),$$

where $P$ denotes probability. In addition, denote by $\gamma$ the probability that this cancer happens in the population; this defines a prior distribution of $X$, that is, $\gamma = P(X = 1)$.

(a) If the device claims cancer for a patient (that is, $Y = 1$), the posterior probability that she actually has cancer is $P(X = 1 \mid Y = 1)$. Please calculate $P(X = 1 \mid Y = 1)$ in terms of $\alpha$, $\beta$ and $\gamma$.

(b) Assume we have $\alpha = \gamma$. What requirement on $\beta$ is needed in order to achieve

$$P(X = 1 \mid Y = 1) \geq 90\%?$$

3. **Multivariate Normal Distribution**

Assume $\boldsymbol{X} = [X_1, X_2]^\top$ is a two-dimensional standard normal random variable,

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

Let $\boldsymbol{Y} = [Y_1, Y_2]^\top$ is obtained by a linear transform of $\boldsymbol{X}$:

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

where $\rho$ is a real number constant.

1. What is the distribution of $\boldsymbol{Y}$? Decide its mean and covariance matrix.
2. Does there exist a value of $\rho$ such that $Y_1$ and $Y_2$ are independent with each other? Please explain the reason.

4. Which of the following statements are true? Please explain your answer for each one.

   (a) When training a neural network with 100 neurons using gradient descent or stochastic gradient descent, if we initialize the weights of all the neurons to be the same value, they will stay the same across the iterations (so effectively, we train a neural network with just a single neuron).

   (b) Learning neural networks is a non-convex optimization problem, and gradient descent algorithms are not guaranteed to find the global optima.

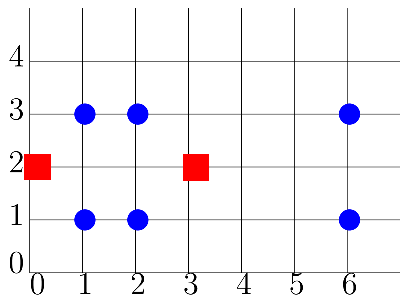
   (c) Kernel regression is guaranteed to outperform linear regression in practice because it allows us to fit more flexible nonlinear curves.

   (d) Estimating the coefficients of kernel regression yields a non-convex optimization problem, because it fits data with a non-linear curve.

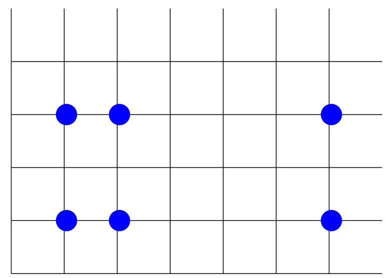
   (e) Expectation maximization (EM) is guaranteed to find the global optima of the log-likelihood of Gaussian mixture models, but $k$-means can only find local optima.

5. **K-Means**

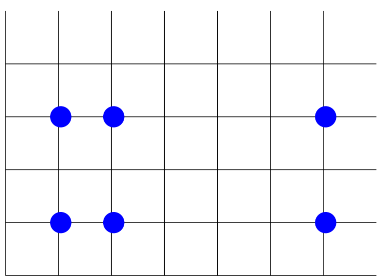
Let us practice $k$-means in this problem. Consider Figure (a) below where we have six data points (blue circles), and we have chosen two initial centroid locations (red squares). Please run $k$-means on this data set and plot the location of centroids at each iteration in Figure (b)-(d) (if the algorithm converges within the first or second iteration, there is no need to fill in the remaining figures).



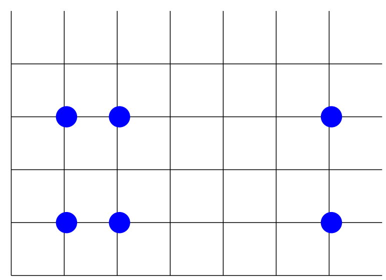

(a) Initialization



(b) Iteration 1

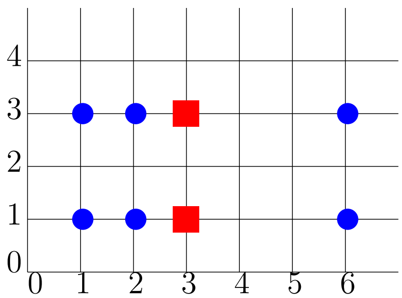


(c) Iteration 2

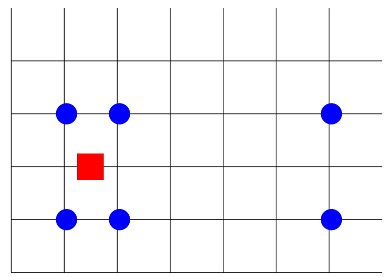


(d) Iteration 3

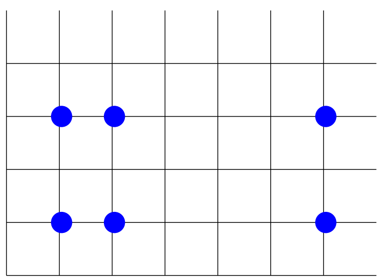
The result of $k$-means is not unique. Different initializations may yield different final results. For example, Figure (a) below shows another possible clustering of the same dataset. In Figure (b), we have initialized one of the centroids. Please initialize the other centroid properly, so that $k$-means converges to the clustering result in Figure (a). Show your initialization and the location of the centroids at each iteration of K-means in Figure (b)-(f). Again, if your algorithm converges in less than 4 iterations, you do not need to fill in the remaining figures.



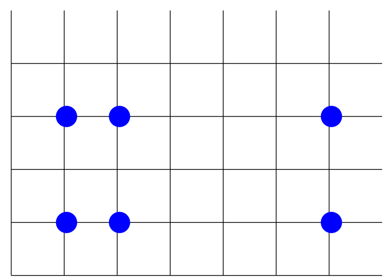

(a) Desirable Result



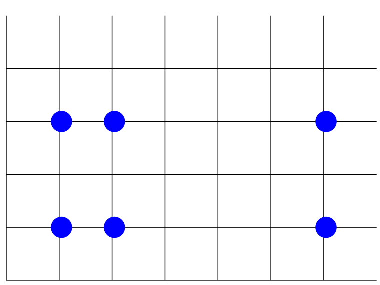
(b) Initialization



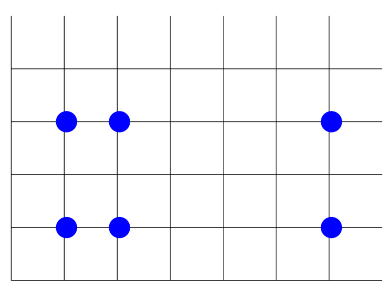
(c) Iteration 1



(d) Iteration 2



(e) Iteration 3



(f) Iteration 4