Machine Learning, Fall 2019

Qiang Liu

Name: EID:

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

$$\text{Prob}[X = k] = \alpha \exp(-k\lambda), \quad \forall k \in \{0, 1\},$$

where α and λ are two parameters to be decided.

(a) What constraints should we put on α and λ to ensure that $\text{Prob}[X = k]$ is a valid distribution?

Solution:
We must have $\text{Prob}[X = k] \geq 0$ *and* $\text{Prob}[X = 0] + \text{Prob}[X = 1] = 1$. *This reduces to (1)* $\alpha \geq 0$ *and (2)*

$$\alpha \exp(-0\lambda) + \alpha \exp(-\lambda) = 1.$$

Solving this gives $\alpha = \frac{1}{1+\exp(-\lambda)}$ *(the constraint of* $\alpha > 0$ *is automatically satisfied by this solution).*

(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 λ and propose a method to estimate λ (it is enough to frame an optimization problem without numerically solving it).

Solution:
Following above, we have

$$\text{Prob}[X=0] = \frac{1}{1+\exp(-\lambda)}, \qquad \text{Prob}[X=1] = \frac{\exp(-\lambda)}{1+\exp(-\lambda)}.$$

We have

$$\text{Prob}(D \mid \lambda) = \text{Prob}(X=0)^{n_0}\text{Prob}(X=1)^{n_1} = \left(\frac{1}{1+\exp(-\lambda)}\right)^{n_0}\left(\frac{\exp(-\lambda)}{1+\exp(-\lambda)}\right)^{n_1}$$

Taking the log, we have

$$\begin{aligned} L(\lambda) := \log \text{Prob}(D \mid \lambda) &= -n_0 \log(1+\exp(-\lambda)) + n_1(-\lambda - \log(1+\exp(-\lambda))) \\ &= -n_1\lambda - (n_0+n_1)\log(1+\exp(-\lambda)) \end{aligned}$$

The maximum likelihood estimator should maximize the log-likelihood:

$$\hat{\lambda} = \arg\max_{\lambda} \{L(\lambda) := -n_1\lambda - (n_0+n_1)\log(1+\exp(-\lambda))\}$$

Taking gradient:

$$\nabla L(\lambda) = -n_1 - (n_0+n_1)\frac{-\exp(-\lambda)}{1+\exp(-\lambda)}$$

$\hat{\lambda}$ should satisfy the zero gradient equation:

$$\frac{\exp(-\hat{\lambda})}{1+\exp(-\hat{\lambda})} = \frac{n_1}{n_0+n_1}.$$

Solving this gives $\exp(-\hat{\lambda}) = n_1/n_0$ and hence $\hat{\lambda} = \log(n_0/n_1)$.
Therefore, the maximum likelihood estimator is $\hat{\lambda} = \log(n_0/n_1)$.

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 positive and false negative of the device to be α and β, respectively, that is,

$$\begin{aligned} \alpha &= \text{Prob}(Y=0 \mid X=1) \\ \beta &= \text{Prob}(Y=1 \mid X=0). \end{aligned}$$

In addition, denote by γ the probability that this cancer happens in the population; this defines a prior distribution of Y, that is, $\gamma = \text{Prob}(Y=1)$.

(a) If the device claims cancer for a patient (that is, $Y=1$), the posterior probability that she actually has cancer is $\text{Prob}(X=1 \mid Y=1)$. Please calculate $\text{Prob}(X=1 \mid Y=1)$ in terms of α, β and γ.

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

$$\text{Prob}(X=1 \mid Y=1) \geq 90\%?$$

Solution:
From the prior, we have

$$\text{Prob}(X = 1) = \gamma.$$

$$\begin{aligned}\text{Prob}(Y = 1) &= \text{Prob}(Y = 1 \mid X = 1)\text{Prob}(X = 1) + \text{Prob}(Y = 1 \mid X = 0)\text{Prob}(X = 0) \\ &= (1 - \alpha)\gamma + \beta(1 - \gamma).\end{aligned}$$

$$\begin{aligned}\text{Prob}(X = 1 \mid Y = 1) &= \frac{\text{Prob}(Y = 1 \mid X = 1)\text{Prob}(X = 1)}{\text{Prob}(Y = 1)} \\ &= \frac{(1 - \alpha)\gamma}{(1 - \alpha)\gamma + \beta(1 - \gamma)}.\end{aligned}$$

When $\alpha = \gamma$, we have

$$\text{Prob}(X = 1 \mid Y = 1) = \frac{\gamma}{\gamma + \beta}.$$

We need $\beta \leq \gamma/9$ in order to achieve $\text{Prob}(X = 1 \mid Y = 1) \geq 90\%$.

$$\begin{aligned}\text{Prob}(X = 0 \mid Y = 0) &= \frac{\text{Prob}(Y = 0 \mid X = 0)\text{Prob}(X = 0)}{\text{Prob}(Y = 0)} \\ &= \frac{\text{Prob}(Y = 0 \mid X = 0)\text{Prob}(X = 0)}{\text{Prob}(Y = 0 \mid X = 1)\text{Prob}(X = 1) + \text{Prob}(Y = 0 \mid X = 0)\text{Prob}(X = 0)} \\ &= \frac{(1 - 10^{-4}) \times (1 - 10^{-4})}{(10^{-4}) \times 10^{-4} + (1 - 10^{-4}) \times (1 - 10^{-4})} \\ &= \frac{(10^4 - 1) \times (10^4 - 1)}{1 + (10^4 - 1) \times (10^4 - 1)} \\ &\approx 1.\end{aligned}$$

3. **Multivariate Normal Distribution**

Assume $\vec{x} = [x_1, x_2]^\top$ is a two-dimensional normal random variable,

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

Let $\vec{y} = [y_1, y_2]^\top$ is obtained by a linear transform of $\vec{x}$:

$$\begin{cases} y_1 = 2x_1 + \rho x_2 \\ y_2 = x_1 + \rho x_2, \end{cases}$$

where ρ is a real number constant.

1. What is the distribution of $\vec{y}$? Decide its mean and covariance matrix.

2. Does there exist a value of ρ such that y_1 and y_2 are independent with each other? Please explain the reason.

Solution:

(a) Define $A = \begin{bmatrix} 2 & \rho \\ 1 & \rho \end{bmatrix}$, *we have* $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, *we have*

$$E[y] = E[Ax] = AE[x] = [0, 0]^\top$$
$$cov[y] = cov[Ax] = A\ I\ A^\top = \begin{bmatrix} 4+\rho^2 & 2+\rho^2 \\ 2+\rho^2 & 1+\rho^2 \end{bmatrix}$$

(b) Impossible, $2 + \rho^2 > 0$.

4. Which of the following statements are true?

   (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 keep the same across the iterations (so effectively we only train a neural network with a single neuron).

   (b) Learning neurons networks yields a non-convex optimization, and it is not guaranteed to find the global optima using gradient descent algorithms.

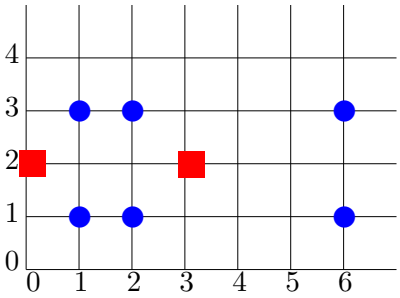
   (c) Kenrel regression is guaranteed to outperform linear regression in practice because it allows us to fit nonlinear curves.

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

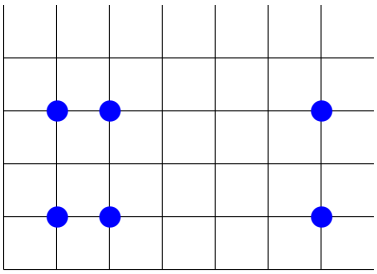
   (e) Expectation maximization (EM) is guaranteed to find the global optima of the log-likelihood of Gaussian mixture models, but K-means can only find local optima.

5. **K-Means**

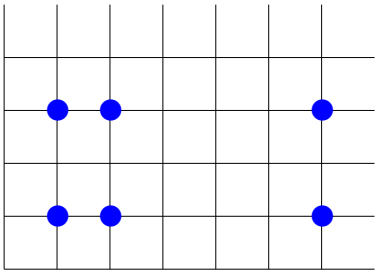
Let us practice K-means in this problem. Consider Figure (a) blow where we have six data points (blue circles), and we have chosen two initial centroid locations (red squares). Please run K-Means on this data set and plot the location of centroids at each iteration in Figure (b)-(d) (if the algorithm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



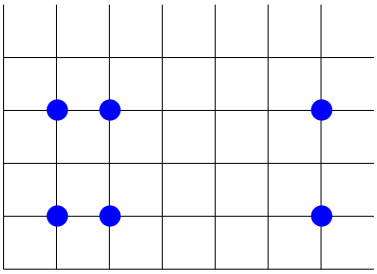
(a) Initialization



(b) Iteration 1

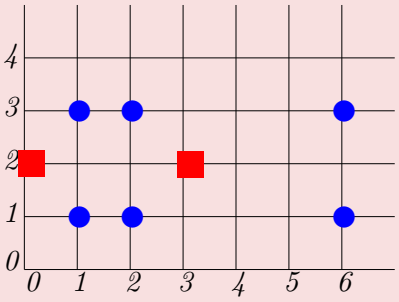


(c) Iteration 2

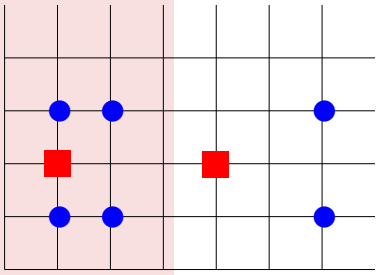


(d) Iteration 3

Solution:

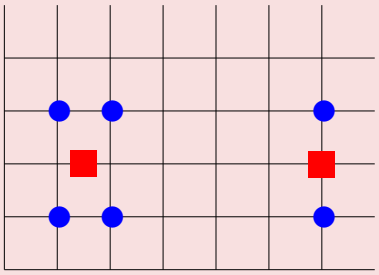


(a) Initialization

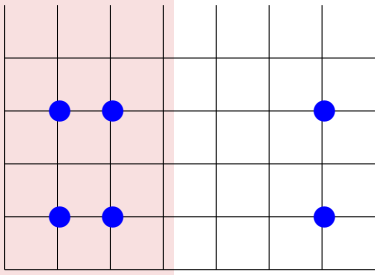


(b) Iteration 1

***Solution**:*

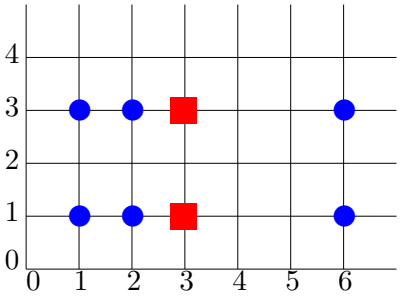


(c) Iteration 2

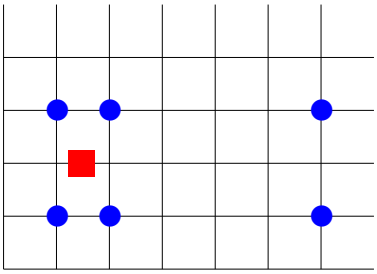


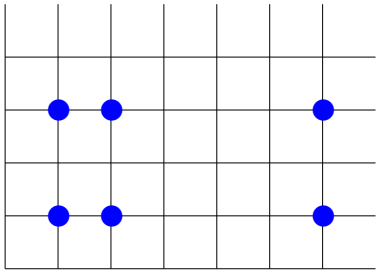
(d) Iteration 3

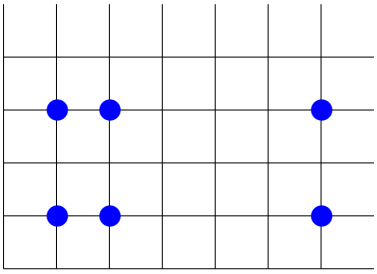
The result of K-means is not unique. Different initialization may yield different final results. For example, Figure (a) below shows another possible clustering of the same dataset. In Figure (b), we have initialized one of the centroid. Please initialize the other centroid properly, so that K-means converges to the clustering result in Figure (a). Show your initialization and the location of centroid 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 the remaining figures.

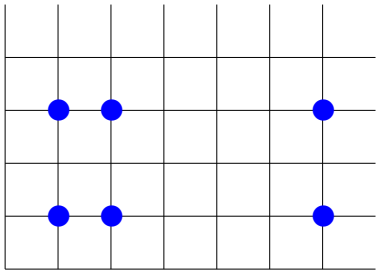


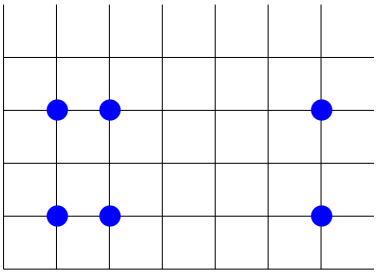
(a) Desirable Result


(b) Initialization

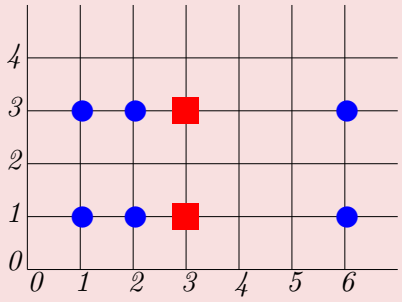

(c) Iteration 1


(d) Iteration 2

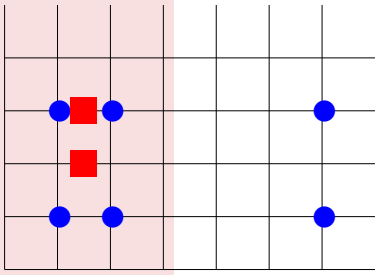

(e) Iteration 3


(f) Iteration 4

Solution:

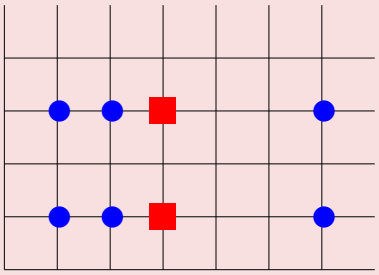


(a) Desirable Result

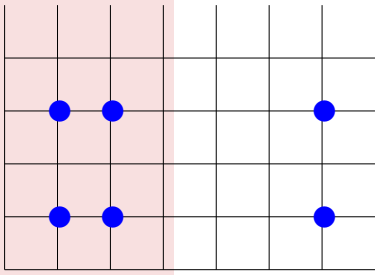


(b) Initialization

***Solution**:*



(c) Iteration 2



(d) Iteration 3