CS5691: Pattern recognition and machine learning
Quiz - 4 Solutions
Course Instructor : Prashanth L. A.
Date : Apr-12, 2019 **Duration** : 40 minutes

I. Short answer questions

1. Consider the dataset $\{(1,1),(2,2),(3,3),(4,4)\}$. Use the K-means clustering algorithm, initialized with the first two data points as the cluster centers, and with $K = 2$. Provide the mapping of the data points to the clusters, and the cluster centers. (1 mark)

   Solution: First cluster: $\{(1,1),(2,2)\}$ with center $(1.5,1.5)$, and second cluster: $\{(3,3),(4,4)\}$ with center $(3.5,3.5)$.
   Alternate solution:
   First cluster: $\{(1,1)\}$, and second cluster: $\{(2,2),(3,3),(4,4)\}$.

2. Recall that, given a dataset $\{(\mathbf{x}_i, y_i), i = 1,\dots,n\}$, the support vector regression method would solve the following optimization problem: Letting $\boldsymbol{\xi} = (\xi_1,\dots,\xi_n)$, and $\boldsymbol{\xi'} = (\xi'_1,\dots,\xi'_n)$,

$$\min_{w,b,\boldsymbol{\xi},\boldsymbol{\xi'}} w^\mathsf{T} w + C\left(\sum_{i=1}^{n} \xi_i + \xi'_i\right) \tag{1}$$

subject to

$$y_i - w^\mathsf{T}\mathbf{x}_i - b \le \epsilon + \xi_i, i = 1,\dots,n \tag{2}$$

$$w^\mathsf{T}\mathbf{x}_i + b - y_i \le \epsilon + \xi'_i, i = 1,\dots,n \tag{3}$$

$$\xi_i \ge 0, \xi'_i \ge 0, i = 1,\dots,n. \tag{4}$$

   Let α_i and α'_i be the Lagrange multipliers associated with the constraints (2) and (3), respectively. From the solution to the dual optimization problem, for each of the following training examples, identify whether the Lagrange multipliers α_i, α'_i both vanish or if one of them vanishes or if none vanish: (1 mark for each table entry)

   Solution:

example	α_i, α'_i
$\lvert y_i - w^\mathsf{T}\mathbf{x}_i - b\rvert \ge \epsilon$	$\alpha_i\alpha'_i = 0$
$\lvert y_i - w^\mathsf{T}\mathbf{x}_i - b\rvert < \epsilon$	$\alpha_i = \alpha'_i = 0$

3. Consider a specific three layer feedforward network with sigmoidal activation functions for all the hidden nodes. Can we construct another three layer feedforward network (with same architecture), where the hidden nodes use the hyperbolic tangent as the activation function such that the two networks compute the same function? ($\frac{1}{2}$ mark)

   Answer: Yes.

4. Consider a feedforward neural network with a linear activation function. Can a polynomial of degree two be represented by such a network? (1 mark)

   Answer: No.

5. For the case of a mixture of Bernoulli distributions, suppose that EM algorithm is initialized such that the mean vector for each component distribution is the same. Does the EM algorithm converge? If yes, provide an estimate of the number of iterations it takes to converge. (1 mark)

   Answer: Yes, and one iteration.

II. A problem that requires a detailed solution

1. Consider a dataset $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, with $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,d})^\intercal$ and $x_{i,j} \in \{0, 1\}$. Consider a mixture of two Bernoulli distributions:

$$f(\mathbf{x}_i) = \lambda_1 f_1(\mathbf{x}_i) + \lambda_2 f_2(\mathbf{x}_i),$$

   where f_j is the mass function of a d-vector of Bernoulli r.v.s with parameter θ_j, $j = 1, 2$. To elaborate, each $\mathbf{x}_i$ is drawn, w.p. λ_j, from a vector of Bernoulli r.v.s with mean vector $\theta_j = (\mu_{j,1}, \dots, \mu_{j,d})^\intercal$.

   Answer the following:

   (a) Write down the expression for the log-likelihood $\log f(\mathcal{X} \mid \theta, \lambda)$ of the data, where $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. (1 mark)

   (b) Introduce the hidden variables $z_{i,j}$ that indicate whether data point $\mathbf{x}_i$ was drawn from component f_j or not, for $i = 1, \dots, n$ and $j = 1, 2$. Let $\mathbf{z}_i = (z_{i,1}, z_{i,2})$, $i = 1, \dots, n$. Write the conditional distribution of $\mathbf{z}_i$ given $\boldsymbol{\lambda}$, and the conditional distribution of $\mathbf{x}_i$, given $\mathbf{z}_i, \boldsymbol{\lambda}, \boldsymbol{\theta}$. Use these quantities to derive the expression for the likelihood of the data, and hidden variables, given λ and θ. (1 mark)

   (c) Apply the EM algorithm to the mixture of Bernoulli distributions, specified above. In particular, show the E and M steps in detail. (2.5 marks)

Solution: See Section 9.3.3 of Bishop's book.