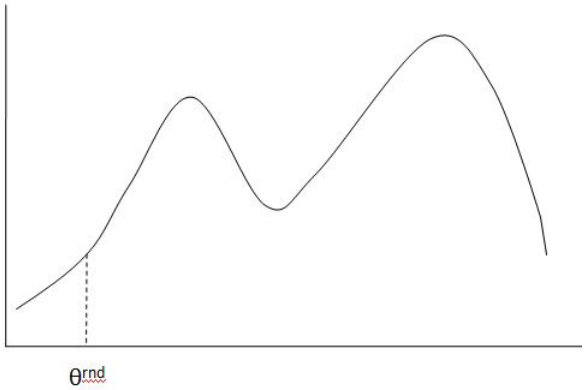
Assignment 11 (Sol.)

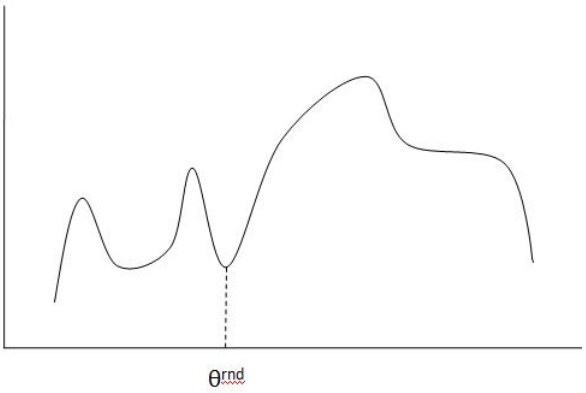
Introduction to Machine Learning

Prof. B. Ravindran

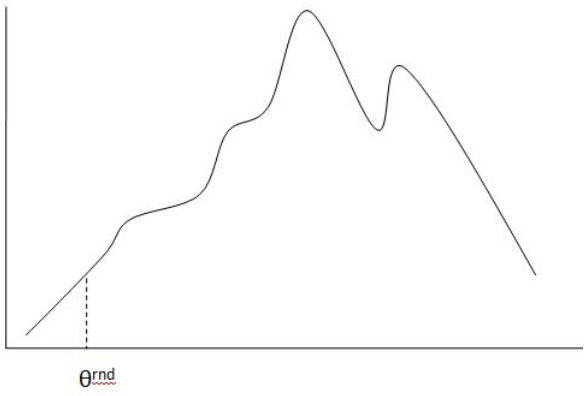
1. In each of the options the diagram contains a log likelihood function for a particular problem as well as an initial value of the parameter used in the execution of the EM algorithm. In which of the given scenarios will the EM algorithm be able to achieve the global maximum?



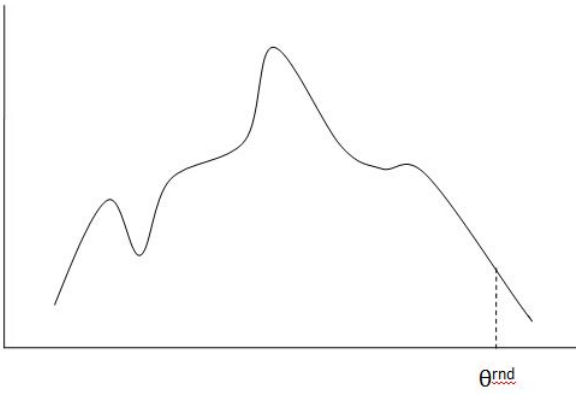
Q1 (a)



Q1 (b)



Q1 (c)



Q1 (d)

Sol. (c)
In each scenario other than (c), the EM algorithm will get stuck in a local minima.

2. Suppose we are given n p-dimensional data points and the corresponding class labels (k different classes). We want to build a decision tree classifier to classify the data. However, we find that there are missing values in the data set. Is it possible to use the EM algorithm to fill the missing data given the above information and making no further assumptions?

   (a) Yes

   (b) No

   Sol. (b)
   Estimating values for missing data is one of the uses of the EM algorithm. However, this cannot be done without specifying the corresponding distribution. Without any knowledge of the distribution, as in this example, we cannot apply the EM algorithm.

3. Suppose you have a PAC learning algorithm A for a concept class C such that with probability at least 0.5, the algorithm will output an approximately correct hypothesis. Suppose that for deployment purposes, you need an algorithm which can output the approximately correct hypothesis with probability at least 0.998. Is it possible to make use of algorithm A for this purpose? If so, how?

   (a) No, we cannot make use of A

   (b) Yes, repeat A three times and choose the best hypothesis

   (c) Yes, repeat A five times and choose the best hypothesis

   (d) Yes, repeat A nine times and choose the best hypothesis

   Sol. (d)
   Given the probability of 0.5 of obtaining an approximately correct hypothesis, if we execute algorithm A nine times, the probability that one of the hypotheses obtained in the nine executions is approximately correct is equal to the probability of tossing a fair coin nine times and observing heads at least once. This can be calculated as

$$\Pr(\text{observing at least one head in nine tosses}) = \frac{2^9 - 1}{2^9} = 0.9980.$$

4. To say that the VC-dimension of a class is at least k, is it necessary for the class to be able to shatter any configuration of k points?

   (a) Yes

   (b) No

   Sol. (b)
   No. According to the definition, it must be able to shatter at least one configuration of k points. For example, we saw in the lecture that the VC-dimension of the straight lines hypothesis class was three even though it is not possible to shatter three co-linear points with straight lines.

5. What is the VC-dimension of the class of axis-parallel rectangles?

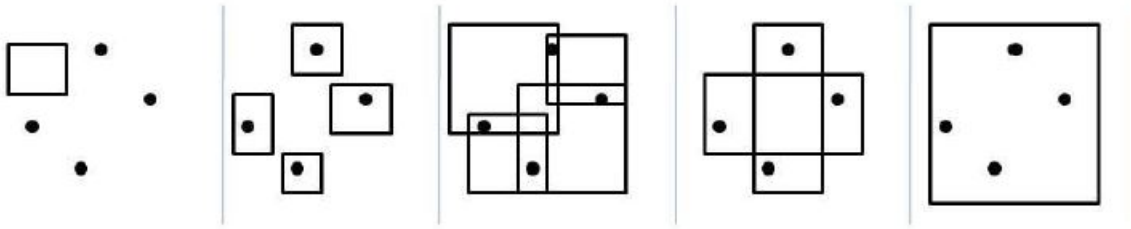
   (a) 3

   (b) 4

   (c) 5

   (d) 6

   Sol. (b)
   The VC-dimension for axis-parallel rectangles is 4. To prove this, we first have to show that the VC-dimension is at least 4. This is shown in the image below by the shattering of 4 points by using axis-parallel rectangles.



Next we need to show that the VC-dimension of axis parallel rectangles is no more than 4. To do so, we consider two cases. In the first case, consider five non-collinear points such that the tightest axis-parallel rectangle that can be drawn around them has at least one point in the convex hull of the rectangle. In this case, the points excluding the point in the interior cannot be enclosed by the axis-parallel rectangle. Now consider five non-collinear points such that the tightest axis-parallel rectangle that can be drawn around them has all the points on its edges. By the pigeon hole principle, at least one of the edges must contain two points whereas each of the remaining three edges contain a point each. Without loss of generality, assume that the edge containing two points is the left edge of the rectangle. Then it is clear that the two points which are the lower point on the left edge and the point on the top edge cannot be enclosed by an axis-parallel rectangle without also enclosing the upper point on the left edge. In contrast, if more than two points lie on a single edge, then the two points on such an edge with another point in between cannot be enclosed. Thus, five points cannot be shattered by axis parallel rectangles.

6. Which of the following statements are true about similarity graph based representations which are used for spectral clustering? (Note that more than one statements may be correct)

   (a) One can give a tighter upper bound than $O(n)$ (where n is the number of data points) on the maximum degree of the vertex corresponding to a point in its kNN based similarity graph representation

   (b) One can give a tighter upper bound than $O(n)$ (where n is the number of data points) on the maximum degree of the vertex corresponding to a point in its epsilon neighborhood based similarity graph representation

   (c) If a is in the k nearest neighbors of b, then b is in the k nearest neighbors of a

   (d) If a is in the epsilon neighborhood of b, then b is in the epsilon neighborhood of a

   Sol. (a) & (d)
   In the kNN similarity graph based representation, every vertex has a degree of k. Since k can be considered O(1), it is a much tighter upper bound than $O(n)$. Hence, (a) is True. On the other hand, an epsilon neighborhood graph could even have a vertex with degree n if all the points lie within a distance epsilon of each other. Hence, (b) is False. We can easily see why (c) is false, point x may be the closest point to y, but z may be the closest point to x. Hence, the 1 nearest neighbor of y will be x, but the 1 nearest neighbor of x will be z. (d) is true, since by definition being in the epsilon neighborhood is a symmetric relation. If point x is within a distance epsilon of point y, then point y is within epsilon of point x.