**Data Mining Research & Practices –Final Exam**

1. (6 %)Design a multilayer feed-forward neural network for the given data in Table 1. Label the nodes in the input and output layers.

Table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Outlook | Wind | Place | Play basketball |
| Record 1 | Sunny | Middle | Indoor | Yes |
| Record 2 | Overcast | Strong | Outdoor | No |
| Record 3 | Rainy | Middle | Indoor | Yes |
| Record 4 | Sunny | Weak | Outdoor | Yes |
| Record 5 | Overcast | Middle | Indoor | No |
| Record 6 | Overcast | Weak | Outdoor | No |

1. (4%) Explain the concept of support vectors, maximum marginal hyperplane and linear separation between classes in SVM (Support Vector Machines). You should draw a diagram in a 2-D plane to aid your explanation.
2. (2%) Suppose that the two parallel hyperplanes for the decision boundary are:

w•x + b = 1

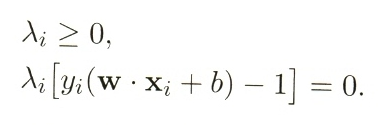
w•x + b = −1

Explain why maximizing the margin is equivalent to minimizing the following objective function: f(w) = ||w||2 / 2.

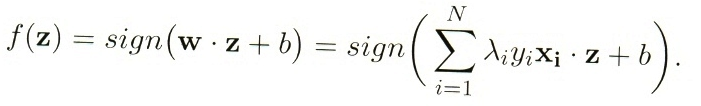
1. (4%) Explain the usage of Kernel function. Use the following example to assist your explanation.



1. (4%) Given the constraints on Lagrange multipliers ( λi )

Draw a diagram to clearly indicate and explain those training instances that have λi = 0 and those training instances that have λi > 0. You also need to indicate the support vectors.

1. (4%) Explain the following equation. You need to explain z, *xi, yi* , *λi* , z, *xi* ***⋅*** z, and the usage of *f* (z). Are all the training instances used in computing *f* (z)?



1. (a) (3%) Explain why we want to minimize the following equation.

(b) (4%) Briefly explain the main idea of gradient descent approach for solving the MF.

(c) (5%) Explain how the ALS (Alternating-Least-Squares) approach handles the implicit feedback (presence or absence of an event). You need to use an example to explain the values of the rating matrix and the usage of cost function (confidence) in ALS.

1. A database has five transactions. Let min\_sup=50% and min\_conf=80%.

Table 1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TID | DATE | ITEMS\_BOUGHT | | | | | | |
| T1 | 14/01/11 | A | B, | C, | D, | E |  | G |
| T2 | 14/02/11 |  |  | C, | D, | E | F |  |
| T3 | 14/03/11 |  | B, | C, |  | E, | F | G |
| T4 | 14/04/11 |  |  | C, |  | E | F, |  |
| T5 | 14/05/11 | A, | B, |  |  | E |  | G |

1. (5%) List all of the *strong* association rules (with support *s* and confidence *c*) matching the following metarule, where X is a variable representing customers and itemi denotes variables representing items (e.g., “A”, “B”, etc.): 
2. (6%) Establish the ***FP-tree*** and find out “*Conditional Pattern Base”,* *“Conditional FP-tree”* and “*Frequent Patterns Generated”* for **item F.**
3. (a)(4%) Given the following large (frequent) **3-itemsets**:

|  |
| --- |
| < 1 2 4>  < 1 2 6 >  < 1 4 6 >  <1 5 7 >  < 2 4 6 >  < 2 4 7 >  < 2 6 7 >  < 3 5 7 > |

(a-1) Find the candidate 4-itemsets according to the **Apriori-generate** algorithm.

(a-2) Find the candidate 4-itemsets after pruning.

(b)(4%) Given the following large **3-sequences**:

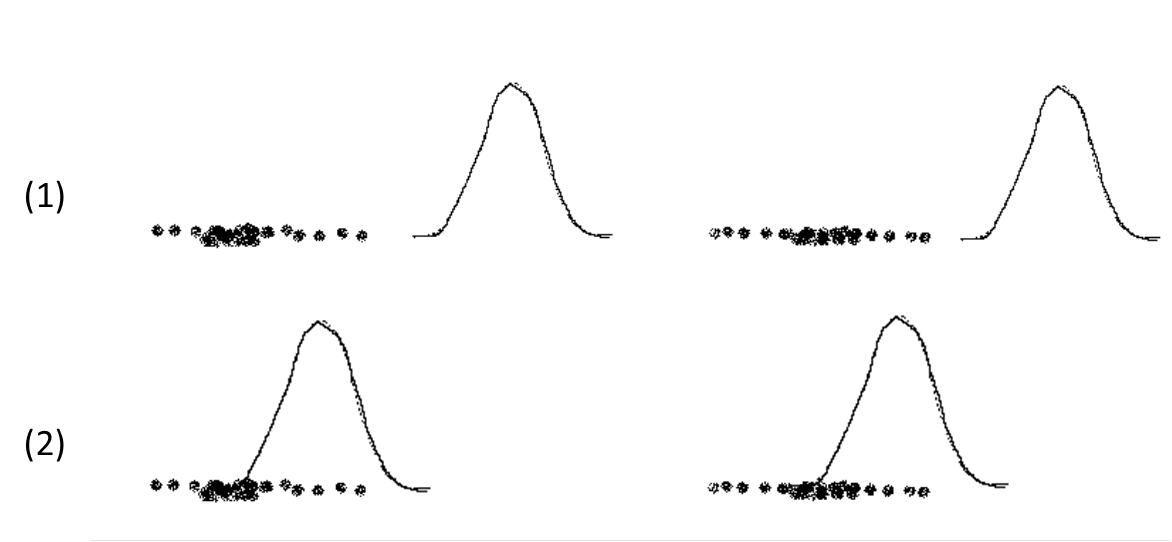
|  |
| --- |
| < 1 2 3>  < 1 2 4 >  < 1 3 4 >  < 1 4 3 >  < 2 3 4 >  < 2 4 3 >  < 3 5 4 >  < 3 5 6 >  < 3 6 4 >  <5 6 4> |

(b-1) Find the candidate 4-itemsets according to the **Apriori-generate** algorithm.

(b-2) Find the candidate 4-itemsets after pruning.

* 1. (8%)Explain the basic concept of EM (Expectation-Maximization) clustering. What are the differences between K-means and EM clustering in terms of **the assignment of data points to clusters** and **the computation of centroids / model parameters**?
  2. (4%) Given the following two mixture models (1) and (2). Which one has higher expected likelihood? Why?





1. In SOM algorithm, we update centroids after selecting the closest centroid. Let m1,….,mj be the centroids. For time step t, let p(t) be the current object (point) and assume that the closest centroid to p(t) is mk. Then, for time t+1, the j­’th centroid is updated by using the following equation.

mj(t+1) = mj(t) + hj(t) (p(t) - mj(t)) , hj­(t) could be Gaussian function as follow.

hj(t) = α(t), α(t) is learning rate, which decreases monotonically with time. rj = (xj,yj) is the two-dimensional point that gives the grid coordinates of the jth centroid. dist(rj,rk) is the Euclidean distance between the grid location of the two centroids.

1. (6%) Please explain the main function hj(t) played. **Use diagram to illustrate the effect** of updating the centroids in terms of the dist(rj, rk) and ( p(t) - mj(t) ).
2. (4%) Show the equation for updating the mk.
3. Assume that there are two latent topics A and B, and the LDA (Latent Dirichlet Allocation) outputs the following assignments of the words to the topics.

Doc1: Broccoli: A, Banana: A, Spinach: B, Hamster: B, Cute: A

Doc2: Banana: A, Spinach: A, Kitten: B

Doc3: Hamster: B, Cute: B, Broccoli: A, Kitten: B

1. (4%) Derive the topic distributions (proportion) for each document.
2. (4%) Derive the probability distributions of words for each topic.
3. (5%) Explain the idea of updating the topic assignment of current word *w* in document *d* based on the topic distributions of all the documents (*p*(topic *k* | document *d*)) and word distributions of all the topics (*p*(word *w* | topic *k*) ).
4. (a) (8%) Explain how k-means clustering is executed in Map-Reduce by using three Map tasks and three Reduce tasks to cluster data into Six clusters. You should use examples (partial data) and draw a diagram to aid your explanations. Clearly indicate the key values that are shuffled to the Reduce tasks.
5. (4%) Explain the map method and reduce method for k-means clustering. You also need to clearly indicate the input and output of the two methods.
6. (a) (6%) Briefly explain the Hadoop Distributed File System (HDFS). You should draw a diagram to aid your explanations. You need to explain the functions of namenode and datanode.

(b) (5%) Assume that the input file is split into six partitions F1, F2, F3, F4, F5 and F6. F3 and F6 are stored in datanode D1. F2 and F4 are stored in datanode D2. F1 and F5 are stored in datanode D3. Six map tasks M1, M2, M3, M4, M5 and M6 are started to process the input file partitions F1, F2, F3, F4, F5 and F6 respectively. There are two reduce tasks R1 and R2. Five workers W1 ~ W5 are started to run the map tasks and reduce tasks. Draw a diagram to show a cluster of computer nodes (machines) and indicate the following – the datanodes with input file partitions, the map tasks, reduce tasks, the workers, the job tracker and the namenode. Clearly indicate the tasks that are assigned to the workers. Assume that the job tracker decides on where to run each map task based on the concept of locality.

(c) (3%) Explain the functions of the Job tracker.