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

1. Use examples or draw diagrams to aid your explanations.
2. (3%) Briefly explain what is the Vector Space Model (VSM).
3. (3%) Explain how to use the k Nearest Neighbors (kNN) approach for document classification.
4. (3%) Briefly explain *tfij*, log and the formula w*ij = tfij* log(for term *tj* of the document *Di*).
5. (3%) Explain how to derive the category vector and use the Category vector for document classification (categorization).
6. There are two Classes, C1 and C2. The total number of documents in the training set is 100, and the number of documents belonging to C1 is 40. The following table shows the probability of P(Xi | Cj). Given a document D1 that contains some terms shown in the table.
7. (5%) Please use the Naive Bayesian document classification method to determine which Class does D1 belong to.
8. (3%) Give a simple example to explain how to calculate P(Xi|Cj)?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table** | | |  | **Document D1** | |
| Term(Xi) | P(Xi|C1) | P(Xi|C2) |  | Document term | Frequency |
| X1 | 1/16 | 3/32 |  | X1 | 0 |
| X2 | 1/4 | 1/16 |  | X2 | 2 |
| X3 | 1/8 | 1/16 |  | X3 | 2 |
| X4 | 1/16 | 1/8 |  | X4 | 0 |
| X5 | 1/8 | 1/4 |  | X5 | 0 |
| X6 | 3/16 | 1/32 |  | X6 | 0 |
| X7 | 1/16 | 1/4 |  | X7 | 1 |
| X8 | 1/8 | 1/8 |  | X8 | 1 |

1. (5%) Briefly explain the main idea of Probabilistic Model (Multinomial model) for document classification.
2. (4%) Briefly explain the differences between the unigram and bigram language models
3. (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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Gender** | **Income** | **Place** | **Buy Product** |
| Record 1 | Male | High | PcHome | No |
| Record 2 | Female | Middle | E-Bay | Yes |
| Record 3 | Female | High | Yahoo | No |
| Record 4 | Male | Low | PcHome | Maybe |
| Record 5 | Female | Middle | Yahoo | Yes |
| Record 6 | Male | High | E-Bay | Maybe |

1. (6%) 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. (3%) Explain why the SVM searches for the hyperplane with the largest margin?
3. (3%) Suppose that the two parallel hyperplanes for the decision boundary are:

w•x + b = 1

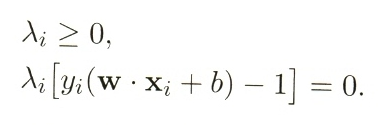
w•x + b = −1

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. (a) (4%) Explain the main ideas of Matrix factorization methods for recommendation. (explain the ideas of user latent factors, item latent factors and how the prediction score is derived)

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

(c) (3%) Briefly explain the main idea of alternative least square approach for solving the MF.

1. A database has seven transactions. Let min\_sup = 55% and min\_conf = 85%.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TID | DATE | ITEMS\_BOUGHT | | | | | | |  |
| T1 | 2015/10/10 | A | B | C |  | E | F |  | H |
| T2 | 2015/11/01 |  |  | C | D | E | F |  |  |
| T3 | 2016/01/01 | A |  | C |  | E | F | G |  |
| T4 | 2016/01/08 | A | B |  |  |  |  |  | H |
| T5 | 2016/01/20 | A | B | C |  | E | F | G | H |
| T6 | 2016/02/11 | A | B |  | D |  |  |  | H |
| T7 | 2016/03/01 |  |  | C | D | E |  | G |  |

(6%) 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.):

****

1. (a)(6%) Given the following large (frequent) **3-itemsets**:

|  |
| --- |
| < 1 3 5 >  < 1 3 7 >  < 1 4 5 >  < 1 4 7 >  < 1 5 7 >  < 2 4 5 >  < 2 4 7 >  < 3 5 7 >  < 4 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)(6%) Given the following large **3-sequences**:

|  |
| --- |
| < 1 2 3 >  < 1 4 6 >  < 1 4 7 >  < 1 6 7 >  < 3 7 9 >  < 4 6 7 >  < 5 3 7 >  < 5 3 9 >  < 5 7 9 > |

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

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

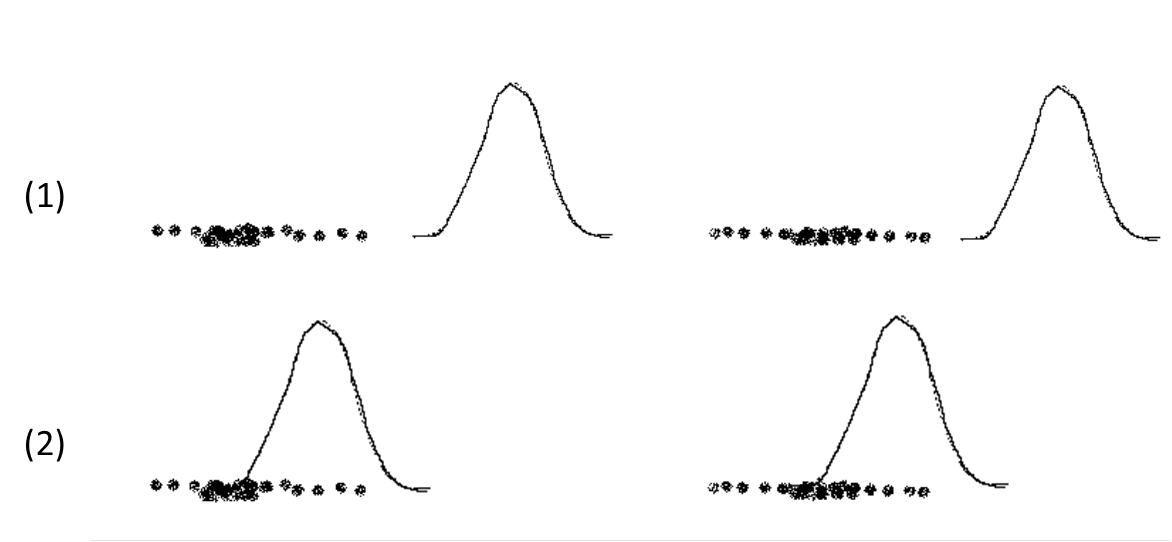
1. (10%) A database has six transactions. Let min\_sup=60% and min\_conf=85%.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TID | DATE | ITEMS\_BOUGHT | | | | | | |
| T1 | 01/01/15 | A |  | C |  |  | F | G |
| T2 | 01/02/15 |  | B | C | D | E |  |  |
| T3 | 01/03/15 | A | B | C |  | E |  | G |
| T4 | 01/04/15 | A | B | C | D |  | F | G |
| T5 | 01/05/15 | A | B | C |  | E |  | G |
| T6 | 01/06/15 | A | B | C |  |  | F |  |

Establish the FP-tree and find out “Conditional Pattern Base”, “Conditional FP-tree” and “Frequent Patterns Generated” for item G.

* 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 mc. 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,rc) 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, rc) and ( p(t) - mj(t) ).
2. (4%) Show the equation for updating the mc.
3. (a) (5%) Explain the following equation. You need to draw diagrams or use examples to assist your explanation.



(b) (5%) Suppose that *PR*(*p*1) is 1/6; *PR*(*p*2) is 1/3; *d*= 0.5; *N* = 6. What is the value of *PR* (*p*5)?

p1

p2

p3

p4

p5

p6