CIS506 Mid-term exam, Spring 2011

Section (A) – Multiple Choice Questions (20 Points)

*Answer all questions. Unless stated otherwise, select a single **best** answer to each question.*

1. Which of the following statements concerning *association rules* is **false**:
    a. Allows for relationships between groups of items to be extracted
    b. Is considered a form of supervised learning
    c. Can be used with different data from a wide variety of sources
    d. Allows the user to pre-set the required confidence levels.
    e. Can be efficiently found using the *apriori* algorithm

2. The following are elements in the apriori algorithm **except** for:
    a. Generation of "candidate" itemsets of size n+1 by doing a self join of the sets of frequent itemsets of size n
    b. Pruning of candidate itemsets which contain non-frequent subsets
    c. Validation of the induced rules on a separate test data set.
    d. Checking for causality based on the confidence of an association rule.
    e. These are all important steps.

3. Collaborative filtering is an approach for matching items to other similar items based on the inputs of a large group of users. There are two popular approaches to this challenge – user and item based algorithms. What is the main difference between the two:
    a. Item-based approaches solve the issue of the sparsity of the ratings matrix
    b. User-based approaches can take the cultural differences of the users into account.
    c. Item-based approaches are not affected by the sparsity of the ratings matrix
    d. Item-based approaches allow new users to quickly receive reasonable recommendations
    e. User-based approaches can not be applied to cases with binary (Like/Dislike) ratings matrices

4. How does the matrix factorization (MF) approach of collaborative filtering provide ratings for previously unrated user-item pairings?
    a. It cannot.
    b. By projecting the requested item rating into a lower dimensional "topic space", MF can provide approximate ratings based on the ratings of the overall topic.
    c. Before use, the factorized matrices need to be trained with existing data – this provides the matrix coefficients to reflect the user-structure
    d. By finding the most similar user ratings.
    e. None of the above

5. In the probabilistic framework, the prediction of a user rating can be written as p($h|a$) (i.e. the probability that a user a would consume item h). Using the matrix factorization approach, this is calculated using:

$$p(h \mid a) = \sum_z p(h \mid z) p(z \mid a)$$

which of the following statements best describes the operation on the right hand side of the equation above:

   a. The expression is wrong.
   b. The probability p($h|a$) is approximated by the probability of h being chosen given topic z, multiplied by the probability of user a choosing z, summed over all topics.
   c. The probability p($h|a$) is approximated by multiplying the probability of h being chosen given topic z, with the probability of user a chosing topic z, which is the topic with the highest probability.
   d. The probability p($h|a$) is approximated by the probability of h being chosen, multiplied by the probability of the most likely topic z given user a.
   e. None of the above

6. In the case of user-based collaborative filtering, what is the main difference between the Cosine similarity and the Pearson correlation?
   a. The Cosine similarity provides a measure of *similarity* while the cosine distance provides a measure of *distance* between users
   b. Calculating the Pearson correlation is less computationally demanding
   c. Calculating the Cosine distance is less computationally demanding
   d. The Pearson correlation treats the user's rating pattern as a zero-mean random variable
   e. None of the above

7. The Expectation-Maximization (EM) algorithm is a form of:
   a. Stochastic parameter optimization
   b. Maximum a-posteriori (MAP) parameter optimization
   c. Maximum likelihood parameter optimization
   d. Classification
   e. Density estimation

8. Jensen's inequality allows us to write the following relationship:

$$\log\left[\sum_z Q(z) p(z \mid x) p(x)\right] \geq \sum_z Q(z) \log p(z \mid x) p(x)$$

Why is this an important step?

a. The right hand side term is a tight bound on the left hand side term
b. It allows us to find the value of Q(z) for all values of x
c. Finding the derivative of the term on the right hand side is difficult
d. Finding the derivative of the term of the left hand side is difficult
e. The right hand side term can be factorized easily

9. The following are reasons why we frequently work with the *logarithm* of a likelihood (instead of directly with the likelihood term itself) **except**:
    a. When dealing with the product of multiple likelihoods terms, using the logarithm means that we only have to work with a sum of terms.
    b. Some interesting distributions are expressed in terms of the exponentials of the data. Using the logarithm makes these easier to deal with
    c. Likelihood terms can be extremely small – using the logarithm prevents numerical problems
    d. Logarithms allow for mixture distributions to be combined into more tractable forms
    e. None of the above

10. In certain situations, the Gaussian Mixture Model (GMM) closely approximates the *k*-means algorithm. However, the GMM is a much richer model and addresses many of the limitations of standard k-means. Which of the following is one of these:
    a. In the *k*-means algorithm, determining the value of *k* can be a difficult challenge
    b. In the *k*-means algorithm, each cluster MUST have the same number of objects
    c. In the *k*-means algorithm, each cluster is assumed to have the same class variance
    d. The *k*-means algorithm is computationally very expensive
    e. None of the above

11. Which of the following statements is **false**?
    a. A data vector with *d* dimensions can be represented as a data point in *d*-dimensional space.
    b. In a *d*-dimensional space, if the given points are linearly separable, the decision hyperplane of $d - 1$ dimension.
    c. A (d–1)-dimensional decision hyperplane can be defined by a *d*-dimensional parameter vector perpendicular to it.
    d. Two parameter vectors of the same direction but different magnitudes (norms) correspond to two distinct decision hyperplanes.
    e. The magnitude (norm) of a vector must be always non-negative.

12. Which of the following **directly** affects the number of mistakes (updates) that the perceptron algorithm makes?
    a. The number of data points

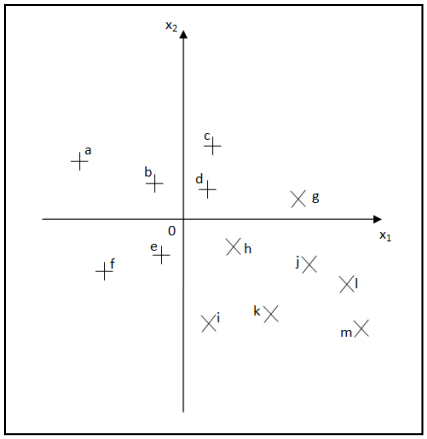
b. The dimensionality of data points
c. The input order of data points
d. The sparseness of data points
e. The radius of the bounding sphere of data points

13. Suppose we have a data vector $\underline{x}$ = {2.0, 3.0, –4.0} and its label y = –1. After executing the perceptron algorithm, we have the optimal parameter vector $\underline{\theta}^*$ = {2.0, 4.0, 4.0}. What is the distance between the vector $\underline{x}$ and the decision boundary?
    a. 5.0
    b. 0.333
    c. –5.0
    d. –0.333
    e. None of above

14. After executing the perceptron algorithm on linearly separable training data points, we have the optimal parameter vector $\underline{\theta}^*$ = {10.0, 10.0, 10.0}. If we have a particular data vector $\underline{x}$ with its label y = –1 in the training data, what is a possible vector for $\underline{x}$?
    a. {–0.9, 0.5, 0.5}
    b. {0.9, –0.9, 0.9}
    c. {0.5 , 0.5, –0.5}
    d. {–0.5, –0.5, 0.5}
    e. None of above

15. Which of the following statement regarding the perceptron algorithm is **false**?
    a. The algorithm will converge if the training data points are linearly separable.
    b. The margin is the distance between the decision hyperplane and its nearest data point.
    c. The wider the margin, the smaller the upper bound of the number of updates required.
    d. The number of updates cannot exceed the number of training data points.
    e. Even if the training data points are linearly separable, the margin found by the perceptron algorithm is not guaranteed to be optimal. (That is, we may still be able to find a wider margin for the same training data points using some other algorithms.)

16. Suppose we the trying to build an SVM classifier based on the positive samples (denoted as +) and the negative samples (denoted as x) in the following figure. Which points are most likely to be the support vectors?



a. d, e, h
b. d, g, h
c. c, d, e, f
d. e, f, i
e. d, e, h , i

17. In a linear SVM (with no slacks allowed), if the number of training data points is 50, the dimensionality of each data point is 5, and the number of support vectors is 4, what is the upper bound of the error rate if we conduct a leave-one-out cross validation test on it?
    a. 4 / 50
    b. (4 * 5) / 50
    c. 4 / 49
    d. (4 * 5) / 49
    e. None of above

18. Regarding a support vector machine, which of the following is **false**?
    a. The objective is to minimize one half the square of the norm of the parameter vector ($\underline{\theta}$).
    b. The purpose of introducing an offset θ_0 is to obtain a wider margin if possible.
    c. If the training data points in *d* dimensional space are not linearly separable, we can try to map those data points into *k* dimensional feature space (where $k > d$) in which the points in the new feature space may become linearly separable.
    d. The quadratic programming problem in SVM is one with quadratic objective and quadratic constraints.
    e. The disadvantage of SVM is that it is sensitive to labeling errors.

19. Regarding a support vector machine, which of the following is **false**?
    a. The purpose of introducing slacks in SVM is to avoid the problem of overfitting.
    b. In general, the bigger the contact value *C* to control the effect of slack variables, the wider the margin between the support vectors and the decision hyperplane.
    c. For a non-linear SVM, a linear decision hyperplane in a higher dimensional feature space can be viewed as a non-linear decision hypercurve in the original data space.

d. For a non-linear SVM, it is not really necessary to explicitly convert the input vectors $\underline{x}_1$ and $\underline{x}_2$ into their respective feature vectors $\underline{\phi}(\underline{x}_1)$ and $\underline{\phi}(\underline{x}_2)$ in order to compute the output from the kernel function $K(\underline{x}_1, \underline{x}_2)$.
e. The advantage of SVM is good generalization because the solution is sparse (i.e. the number of support vectors are much smaller than the number of training samples).

20. What is the time complexity of the following counting-based kernel perceptron algorithm for *n* training samples each mapped into a *k*-dimensional feature vector (denoted as $\phi(\underline{x})$)?

Initialize: $\alpha_i = 0,\ i = 1, \ldots, n$

Repeat for $t = 1, \ldots, n$

if $y_t\left(\sum_{i=1}^{n} \alpha_i y_i [\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x}_t)]\right) \leq 0$ (mistake)

$\alpha_t \leftarrow \alpha_t + 1$

a. $O(n \,.\, k)$
b. $O(n \,.\, k^2)$
c. $O(n^2 \,.\, k)$
d. $O(n^2 \,.\, k^2)$
e. None of above

Section (B) – Structured questions (45 Points)

Answer all questions. Please also include all the workings and derivations used to obtain your answers (working paper will be provided).

1. The Laplace distribution is a member of the exponential distribution of functions, and is defined as:

$$p(x \mid \mu, s) = \frac{1}{2b} \exp\left[-\frac{|x - \mu|}{s}\right]$$

where x is an instance of the random variable **X**, and the distribution is parameterized by μ and s.

1	5	8	12	15

a) Sketch a graph that depicts this probability.
b) Given a sample of points $\{x_1, x_2 \ldots . x_N\}$,:
   i. Given that this data is drawn from a Laplace distribution, write down its log-likelihood, and find its derivatives w.r.t. the parameters μ and *s*.

ii. Sketch the graph of the derivative w.r.t. μ vs μ, using the points given in the table above as the data sample.

iii. Show that the Maximum Likelihood Estimate (MLE) of the quantity μ is given by the sample median, and the estimate of s by:

$$\hat{s} = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

c) You would like to create a "Laplace Mixture Model" (LMM) – similar to the Gaussian Mixture Model but using Laplace instead of Gaussian distributions as the mixture components:

$$p(x;\theta) = \sum_z p(x \mid \mu_z, s_z) p(z)$$

(z – variable which indexes the "active" mixture component)

To derive an algorithm for learning the model parameters, you decide to use the EM-algorithm. In this context:

i. Write down the full expression (in terms of the Laplace distribution) for the Variational Distribution *Q(z)* which will provide a tight lower bound on the Log Likelihood.

ii. Explain, using appropriate mathematical expressions if necessary, how the expression for the MLE of μ is affected in the context of the LMM (compared to the case in b.iii).

iii. Suppose we replace the expression of Q(z) with the following approximation:

$$Q(z) = \begin{cases} 1 & z = \arg\max_z p(z \mid x;\theta) \\ 0 & Otherwise \end{cases}$$

What algorithm does the EM-iterations reduce to?

(*25 Points*)

2. Suppose we have the following six 2-dimensional training data points and their respective labels.

$\underline{x}_1 = \{4, 1\}, y_1 = +1$

$\underline{x}_2 = \{1, 1\}, y_2 = -1$

$\underline{x}_3 = \{2, 3\}, y_3 = -1$

$\underline{x}_4 = \{3, 1\}, y_4 = +1$

$\underline{x}_5 = \{5, 2\}, y_5 = +1$

$\underline{x}_6 = \{2, 3.5\}, y_6 = -1$

Suppose you run the following perceptron algorithm on the training data.

Initialize: $\underline{\theta} = 0$

Repeat until convergence:

  for $t = 1, \ldots, n$

    if $y_t(\underline{\theta} \cdot \underline{x}_t) \leq 0$ (mistake)

      $\underline{\theta} \leftarrow \underline{\theta} + y_t \underline{x}_t$

Show the change of $\underline{\theta}$ after each update. What is the final $\underline{\theta}$? How many updates are required?

(Hint: you are required to pass through the whole data set not more than twice to get the final $\underline{\theta}$.)

(20 Points)

Section (C) – Open questions (15 Points)

This section is designed to test your broader perspective of Machine Learning. Consequently your solutions do not have to be extremely technical but should involve some careful reflection on your part.

1. When computer scientists discuss algorithms and models, we frequently talk about approaches that are "Elegant". While this is a somewhat subjective concept, there is a general sense that this term should refer to methods that have some or all of the following properties:
   i. **Parsimonious** – i.e. the technique should be "simple" in the sense that it adequately describes the process or phenomenon being modeled without introducing additional complexity.
   ii. **Efficient** – Should be computationally efficient/scales well with the size of the data set.
   iii. **"Cute"** – Employs some neat trick or technique to solve an otherwise difficult or intractable challenge
   iv. **Principled** – based on a sound mathematical or scientific foundation
   v. **Broadly applicable** – the algorithm or model should be easily customized to a variety of problems or applications, rather than being a very specialized technique.

   Select one of the algorithms that you have learnt about (either from the Machine Learning or Data Mining courses, or perhaps through your own research interests) which you feel is "elegant", and explain how it possesses as many of these properties as is possible.