CIS606 Mid-term exam, Spring 2013

Multiple Choice Questions (1pt each unless noted otherwise)

Notes

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

There are **thirty questions in total*

1. In the matrix factorization approach to collaborative filtering, the original larger matrix R is factorized into the product of two smaller matrices P and Q,

   i.e. $R=PQ$, where R ~ $m\times n$, P ~ $m\times k$ and Q ~ $k\times n$

   (*m* - number of users, *n* - number of items)

   The following statements correctly explain the workings of this process *except* for:

   a) Assumes that the true data lies on a subspace of the item (or user) space.

   b) Each of the *columns* of P corresponds to one feature, as does each of the *rows* of Q

   c) Typically, *k* has to be much smaller than *m* and *n* for this to make sense

   d) The total number of elements in P and Q combined should be smaller than the number of elements in R

   e) All are valid statements

2. Consider the following matrix factorization, where the matrix on the left is the user-item matrix as described above:

$$\begin{bmatrix} 6 & 3 & 9 & 3 & 0 \\ 0 & 1 & 2 & 0 & 3 \\ 0 & 2 & 4 & 0 & 6 \\ 2 & 1 & 3 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 3 \\ 1 & 0 \\ 2 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 2 & 0 & 3 \\ 2 & 1 & 3 & 1 & 0 \end{bmatrix}$$

   suppose we wish to add a new user into the mix:

$$U' = \begin{bmatrix} 2 & 2 & 5 & 1 & 3 \end{bmatrix}$$

Assuming that this does not change the item-feature mapping, what would the new “P” matrix be?

a) $\begin{bmatrix} 0 & 3 \\ 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 1 & 2 \end{bmatrix}$ b) $\begin{bmatrix} 2 & 1 \\ 0 & 3 \\ 1 & 0 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 2 \\ 0 & 3 \\ 1 & 0 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$ d) $\begin{bmatrix} 1 & 1 \\ 0 & 3 \\ 1 & 0 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$ e) $\begin{bmatrix} 0 & 3 \\ 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$

The next two questions are based on the following scenario

You are trying to design a "course recommendation" system so that future MI students will be able to choose electives which are of interest and use to them. To do this, you use a collaborative filtering approach.

The registrar provides you with the following information:

Student A: CIS503, CIS506, CIS511
Student B: CIS503, CIS505, CIS506, CIS511
Student C: CIS503, CIS505, CIS507
Student D: CIS505, CIS506, CIS507

3. A new user (Student E) joins the institute. She enrols in CIS503 based on the advise of her advisor. However she is not sure what other courses to take.

   To get her up and runnning ASAP, you decide to use an *item-based* recommendation approach. Why is this a good choice in her case?

   a) It is computationally more efficient
   b) It provides more accurate recommendations
   c) It can generate recommendations for users with little prior history
   d) It allows new items to be easily integrated into the system
   e) It provides a more accurate parameterization of the user-item space

4. For binary ratings, we can use the "Jaccard Index", defined as:

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}$$

   where the intersection and union operators denote the number of matches as well as the total coverage of the "1"s in the corresponding binary numbers.

   Using this similarity function, recommend one other course for student D.

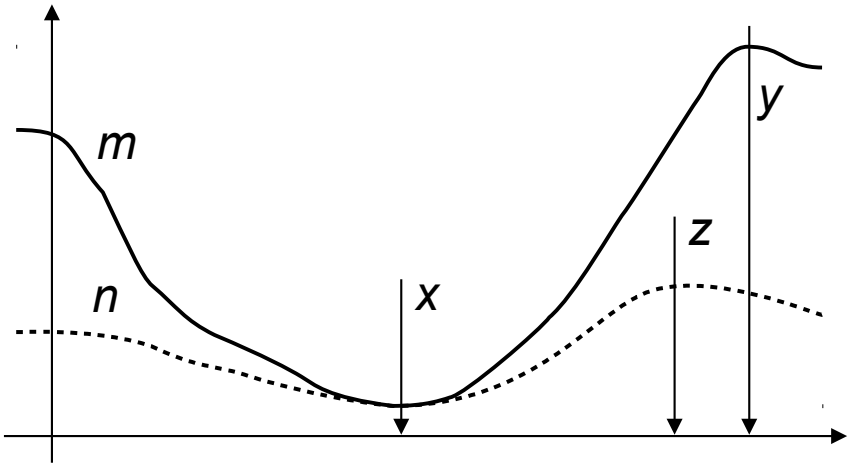
   a) CIS505
   b) CIS506
   c) CIS507
   d) CIS511
   e) Not possible to tell

5. Which of the following statements about the EM algorithm is NOT true:

   a) It is a form of maximum likelihood estimation
   b) It is frequently used for optimizing the parameters of a model
   c) It is suited to problems where some of the data is hidden
   d) It is the only way by which mixture models can be trained
   e) They are all accurate

6. The user-based, item-based and matrix factorization approaches to collaborative filtering are:
    a) All memory based
    b) Memory based, memory based, model based
    c) Memory based, model based, model based
    d) Model based, model based, memory based
    e) All model based

The next three questions are based on the following figure:



7. The figure above is commonly used to portray the workings of the EM-algorithm. The two curves *m* and *n* depict:
    a) The log-likelihood of the observed data, and the expected incomplete log likelihood
    b) The expected incomplete log likehood, and the log-likelihood of the observed data
    c) The log-likelihood of the observed data, and the expected value of the complete data log-likelihood respectively.
    d) The expected value of the complete data log-likelihood, and the log-likelihood of the observed data respectively.
    e) The variational bound on the log likelihood, and the prior probability of the latent variables.
8. Consider the following three stages during the optimization process
    (i) Optimal solution (iii) Current iteration (ii) After the next M-step
    Match the points x, y and z to the three stages above:
    a) x,y,z
    b) y,z,x
    c) z,x,y
    d) x,z,y
    e) y,x,z

9. What spaces do the *y*- and *x*-axes represent:
    a) Temporal & spatial displacement, respectively
    b) Log-likelihood & parameter spaces, respectively
    c) Log-likelihood & spatial displacement, respectively
    d) Parameter space & spatial displacement, respectively
    e) Temporal and parameter spaces, respectively
10. In the acronym “EM”-algorithm, “E” refers to:
    a) The expected value of the complete data log likelihood
    b) The expected value of the observed data log likelihood
    c) The expectation that the algorithm will maximize the fit of the model to the data
    d) The expectation that each iteration is guaranteed to increase the log likelihood
    e) The expected value of the parameter update terms

The next three questions are based on the following scenario:

“A 1-D variable is distributed according to a Laplace Mixture Model (LMM). This is defined as follows:

$$p(x)=\sum_{z\in 1,2}\frac{1}{2s_z}\exp\left(\frac{-|x-\mu_z|}{s_z}\right)p(z)$$

Also, in the initial training set, you have the following:

Class 1: x=1,2,3

Class 2: x=6,7,8

11. What are the initial maximum likelihood estimates of μ_1, s_1, μ_2, s_2 ?
    a) 1,1/3,6
    b) 1,2/3,6
    c) 2,1/3,7
    d) 2,2/3,7
    e) 3,1/3,8
12. For subsequent use, the model will be updated in an incremental fashion using the EM-algorithm update rules. Suppose the following set of points are observed: x=2,4,5,7.

For these three values of x, what are the respective values of Q(z=1):

    a) 1,0.82,0.18,0
    b) 0,0.18,0.82,1
    c) 1,0.18,0.82,0
    d) 0,0.82,0.5,1
    e) 1,0.82,0.18,0.5

13. What is the updated value for μ_1 ?
    a) 1
    b) 2
    c) 3
    d) 5
    e) 6
14. The Binomial distribution defines the number of “successes” obtained from a series of Bernoulli trials (for e.g. the number of heads obtained after tossing a coin a number of times.

    It is written as follows: $$p(x=k)=\binom{n}{k}p^k(1-p)^{(n-k)}$$

    Where *k* is the number of successes, *n* is the number of trials and *p* is the probability of success for each trial.

    $\binom{n}{k}$ is the number of ways for choosing *k* items from a set of n, and is defined:

    $$\binom{n}{k}=\frac{\prod_{i=0}^{k-1}(n-i)}{\prod_{j=1}^{k}j}$$

    A biased coin is tossed 5 times and “heads” appears 2 times. By using maximum likelihood estimation or any other method, what is *p* the probability of getting “heads”?
    a) 0.1
    b) 0.2
    c) 0.3
    d) 0.4
    e) 0.5
15. You have two coins, one with p=0.2 and one with p=0.6 (p is the probability of flipping a head).

    Let's label the first coin z=1, and the second coin z=2. One of these coins is picked randomly and flipped twice, obtaining first a head then a tail. What is $p(z=1|X)$, where X → {head,tail}?
    a) 0.2
    b) 0.3
    c) 0.4
    d) 0.5
    e) 0.6

… Continued on following page …

16. 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, there exists a decision hyperplane is of $d - 1$ dimensions.
c. A (d–1)-dimensional decision hyperplane can be defined by a d-dimensional parameter vector perpendicular to it and a scalar offset value from the origin.
d. If a set of data points cannot be linearly separable in d-dimensional space, there is a chance that they can be linearly separable in l-dimensional space,where $l < d$.
e. None of the above.

17. What is the best name for the following feature selection method?

1. S starts with the full feature set
2. Find the best feature to *remove* (by checking which removal from S gives best performance on a test set).
3. If overall performance has improved, return to step 2; else stop

a. Complete feature selection method.
b. Forward feature selection method.
c. Backward feature selection method.
d. Relief feature selection method.
e. None of the above.

18. If we have the following 4 training data points with their respective class labels, which of the following method **CANNOT** provide us a classifier with 0% training error (supposed no slacks are allowed)?

$$\underline{x}_1 = [1.0, 1.0], \quad y_1 = -1$$
$$\underline{x}_2 = [1.0, 2.0], \quad y_2 = +1$$
$$\underline{x}_3 = [2.0, 1.0], \quad y_3 = +1$$
$$\underline{x}_4 = [2.0, 2.0], \quad y_4 = -1$$

a.Perceptron algorithm with RBF kernel.
b. SVM with linear kernel.
c. SVM with polynomial kernel.
d. SVM with RBF kernel.
e. None of the above.

19. Which of the following directly affects the number of mistakes (updates) that the perceptron algorithm makes if all the data points are linearly separable?

a. The number of data points.
b. The dimensionality of data points.
c. The input order of data points.
d. The geometric margin (the distance between the decision boundary and its closest data point)
e. None of the above.

20. 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 smaller the constant 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. 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).
e. None of the above.

21. For the Weighted Majority algorithm given below, assume that there are 4 experts and the penalty parameter β = 0.5. If $w^{(t)}$ = {1.0, 1.0, 0.5, 0.25}, $q_+^{(t)}$ = 2.0, $q_-^{(t)}$ = 0.75, and the actual label $y^{(t)}$ = +1, what will be $w^{(t+1)}$?

The Weighted Majority Algorithm

- Parameter: $0 < \beta < 1$
- Initialization: set $w_j = 1$ for $j = 1 \dots d$.
- For $t = 1 \dots T$
  1. I receive some input $\underline{x}^{(t)}$
  2. Define
  $$q_+^{(t)} = \sum_{j:x_j^{(t)}=+1} w_j; \quad q_-^{(t)} = \sum_{j:x_j^{(t)}=-1} w_j$$
  If $q_+^{(t)} > q_-^{(t)}$ predict $\hat{y}^{(t)} = +1$, else $\hat{y}^{(t)} = -1$
  3. I receive the correct label $y^{(t)} \in \{-1, +1\}$. If $\hat{y}^{(t)} \neq y^{(t)}$ I have made an error.
  4. Update: for all j such that $x_j^{(t)} \neq y^{(t)}$, set $w_j = w_j \times \beta$

a. {1.0, 1.0, 0.5, 0.25}
b. {1.0, 1.0, 0.25, 0.125}
c. {0.5, 0.5, 0.5, 0.25}
d. {0.5, 0.5, 0.25, 0.125}
e. None of the above.

22. In which of the following algorithm, is "random sampling with replacement" used?

a. Kernel perceptron algorithm.
b. AdaBoost algorithm.
c. Bagging algorithm.
d. Dual SVM algorithm.
e. None of the above.

23. Suppose we build an ensemble classifier with equally-weighted majority voting by 3 base classifiers. If the accuracy of each base classifier is 0.8, what is the expected accuracy of the ensemble classifier?

a. 0.896
b. 0.8
c. 0.999
d. 0.854
e. None of the above.

24. Which of the following statement about a kernel function is FALSE?

a. The purpose of a kernel function $K(\underline{x}_1, \underline{x}_2)$ is to find the dot product of the feature vectors $\underline{\phi}(\underline{x}_1)$ and $\underline{\phi}(\underline{x}_2)$, where $\underline{x}_1$ and $\underline{x}_2$ are the input vectors.
b. For a non-linear kernel function, 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 a higher dimensional space in order to compute the output from a kernel function $K(\underline{x}_1, \underline{x}_2)$.
c. Any distinct set of training points, regardless of their labels, are separable using the Radial Basis kernel function.
d. A kernel function is valid if it is symmetric and for all the training data points, the Gram matrix is positive semi-definite.
e. None of the above.

25. If $\underline{a}$ and $\underline{b}$ are two data points of dimension ($d > 1$) and $K_1(\underline{a}, \underline{b})$ and $K_2(\underline{a}, \underline{b})$ are two valid kernel functions, which of the following is NOT a valid kernel function?

a. $(K_2(\underline{a}, \underline{b}))^3$
b. $(K_1(\underline{a}, \underline{b}) \cdot K_2(\underline{a}, \underline{b})) + (K_2(\underline{a}, \underline{b}))^2$
c. $(K_1(\underline{a}, \underline{b}) + 1) \cdot K_2(a, b)$
d. $\underline{a}^2 K_1(\underline{a}, \underline{b}) \underline{b}^2$
e. None of the above.

26. Suppose you have a data vector $\underline{x} = [-2.0, 2.0]$ with its label $y = -1$. After running the perceptron algorithm, you have the optimal parameter vector $\underline{\theta}^* = [4.0, 3.0]$. What is the distance between the vector $\underline{x}$and the decision boundary?

a. 2.5
b. 5
c. 0.2
d. 0.4
e. None of the above

27. In the counting-based perceptron algorithm on the following 4 data points, if the mistake counts on those 4 data point are: $\underline{\alpha} = \{3, 0, 2, 0\}$, what is the final parameter vector $\underline{\theta}^*$?

$$\underline{x}_1 = [1.0, 2.0], \ y_1 = -1$$
$$\underline{x}_2 = [1.0, 3.0], \ y_2 = -1$$
$$\underline{x}_3 = [3.0, 1.0], \ y_3 = +1$$
$$\underline{x}_4 = [3.0, 2.0], \ y_4 = +1$$

a. [3.0, 4.0]
b. [3.0, -4.0]
c. [-3.0, 4.0]
d. [-3.0, -4.0]
e. None of the above.

28. We **CANNOT** build an ensemble learner using the following setup.

a. Different learning algorithms.
b. Algorithms with different choices for parameters.
c. Datasets with different features.
d. Different subsetsfrom the same pool of dataset.
e. None of the above.

29. In AdaBoost algorithm, which of the following is the parameter that a user can manually tune?

a. Number of base learners (hypotheses) in the ensemble.
b. Weight of each data points.
c. Weighted error of each hypotheses.
d. Weight of each hypotheses.
e. None of the above.

30. If you have a dataset with 10 features (attributes), if you want to select 5 features from it, what it the total number of distinct feature sets consisting of5 features?

a. 30240
b. 252
c. 3628800
d. 120
e. None of the above.