<u>CIS506 Mid-term exam, Spring 2012</u>

Multiple Choice Questions (1pt each unless noted otherwise)

<u>*Notes*</u>

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

**The following three algorithms in your lecture notes are given in the Appendix for your reference.*

Perceptron Algorithm

Halving Algorithm

Weighted Majority Algorithm

There are **thirty questions* **in total**

1. Consider the following book ratings:

Name\Title	The Hobbit	The Bourne Identity	The Silmarillion
Wayne Rooney	2	4.5	1.5
Federer	3.5	3.2	5
Foo Kok Keong	2.5	5	?

Based on these ratings, what would be the predicted rating for Foo Kok Keong, for “The Silmarillion”, using first *1*-NN then using the weighted averaging technique?

(use the Cosine similarity function and the user-based approach)

a) 1.5,3.2

b) 1.5, 3.7

c) 5,3.2

d) 5, 3.7

e) None of the above

2. What if an item-based approach was taken?

a) 2.5,4

b) 2.5,3.6

c) 5,4

d) 5,3.6

e) None of the above

3. It is also possible to formulate the prediction as a matrix factorization problem. Where the problem is to decompose the ratings matrix *R* into the product of two smaller matrices:

$$R \approx PQ$$

(If R is $n \times m$, then P and Q are $n \times k$ and $k \times m$ respectively)

For a small data set, we can do this manually as follows: First, set k=2. Next, to reduce the degrees of freedom, set P to be the first two columns of R.

In this way, the problem then reduces to finding Q, which can be tackled as follows:

$$\begin{bmatrix} 2 & 4.5 & 1.5 \\ 3.5 & 3.2 & 5 \\ 2.5 & 5 & ? \end{bmatrix} = \begin{bmatrix} 2 & 4.5 \\ 3.5 & 3.2 \\ 2.5 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & q_1 \\ 0 & 1 & q_2 \end{bmatrix}$$

In this case, suggest appropriate values of q_1 and q_2 and subsequently, what is the predicted rating for “Foo Kok Keong” and “The Silmarillion”?

a) 0.5,1e.9,3.2

b) -0.5,1.9,2.2

c) 1.9,0.5,3.2

d) 1.9,0.5,2.2

e) 1.9,-0.5,2.2

4. In the context of the previous question, which of the following statements about k DO NOT apply?

   a) It controls the number of degrees of freedom of the subsequent optimization

   b) It is related with the intrinsic dimensionality of the column space of R

   c) It should ideally be a lot smaller than n or m.

   d) It's value depends on whether a user or item-based approach is adopted.

   e) By setting a smaller value of k, we can ensure that matrices P and Q are not sparse.

5. The following set of transactions were obtained from a supermarket's database:

   {*chips, salsa, popcorn*}

   {*salsa, popcorn, cheese*}

   {*chips, salsa, cheese*}

   {*cheese, pizza-base, baguettes*}

   {*salsa, cheese, pizza-base*}

   {*cheese, sausages, baguettes*}

   Assuming a confidence level of ≥80%, which of the following is a valid association rule:

   a) pizza-base → cheese

   b) cheese → chips

   c) chips → popcorn

   d) salsa → popcorn

   e) salsa → chips

6. Which of the following statements about association rule mining is NOT valid:
    a) Meeting the confidence threshold implies that the support threshold was met
    b) Meeting the confidence threshold helps to establish causality.
    c) Having a support threshold ensures only commonly encountered instances can form rules
    d) Having a support threshold helps to ensure that rules are generalizable
    e) Meeting the support threshold does not guarantee that an association is a valid rule.
7. The Expectation-Maximization algorithm is:
    a) A form of reinforcement learning
    b) A form of supervised learning
    c) A method for performing maximum likelihood estimation
    d) A method for performing maximum a-posteriori estimation
    e) None of the above
8. The "Expectation" in the EM-algorithm refers to the expected value of:
    a) The value of the parameters
    b) The value of the hidden data
    c) The complete data log likelihood
    d) The observed data log likelihood
    e) The probability of the latent variables

<u>The following scenario applies to Q9-Q12</u>

"There are two varieties of apples. In variant one, the weight is distributed according to a normal distribution, while in variant two, the weight is distributed according to a Binomial distribution → so, when the tree produces an apple, it literally pops into existence, and is immediately at weight w_1 or w_2, (yes, these are very strange apples, from the planet Qo'nos - pronounced "kronos")

The following probability distribution captures the situation above:

$$p(w)=\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(w-\mu)^2}{(2\sigma^2)}\right]\right]p(z=1)+\left[\delta(w-w_1)\,p+\delta(w-w_2)(1-p)\right]p(z=2)$$

w - weight of apple; μ,σ - mean and std of the gaussian; p - the probability that $w=w_1$ in the Bernoulli case; z - variant index.

Assume that $\mu=5$, $\sigma=2$, $w_1=3$ and $w_2=6$, $p=0.2$, and $p(z=1)=p(z=2)=0.5$."

9. What is the probability of w=6?
    a) 0.19
    b) 0.29

c) 0.35
d) 0.49
e) 0.55

10. What is the probability of variant 2 given the observed data mentioned in question 9?
    a) 0.52
    b) 0.62
    c) 0.69
    d) 0.75
    e) 0.82
11. Three apples are collected, and their weights are {3,5,6}. During the M-step, solving which of the following equations provides the update term for *p*?
    a) $\frac{p(z=2|w=3)}{p}+\frac{p(z=2|w=5)}{p}-\frac{p(z=2|w=6)}{1-p}=0$
    b) $\frac{p(z=2|w=3)}{p}-\frac{p(z=2|w=5)}{1-p}-\frac{p(z=2|w=6)}{1-p}=0$
    c) $\frac{p(z=2|w=3)}{p}-\frac{p(z=2|w=6)}{1-p}=0$
    d) $\frac{p(z=1|w=3)}{p}-\frac{p(z=2|w=6)}{1-p}=0$
    e) $\frac{p(z=1|w=3)}{p}-\frac{p(z=1|w=6)}{1-p}=0$
12. If instead, the weights were {7,8,9}, what would the next M-step updated value of μ be?
    a) 7
    b) 7.5
    c) 8
    d) 8.5
    e) 9

Q13-14

The following expression holds true for all forms of Q(z), the “variational distribution”:

$$\log\sum_{z} Q(z)\frac{p(x,z)}{Q(z)} \geqslant \sum_{z} Q(z)\log\frac{p(x,z)}{Q(z)} \cdots\cdots(1)$$

13. In the context of the EM-algorithm, this is an important property because:
    a) Q(z) captures the confidence we have in the likelihood

b) Q(z) helps to determine the uncertainty in the likelihood

c) The LHS of the equation is easily differentiable

d) The RHS of the equation is easily differentiable

e) None of the above

14. In theory, Q(z) could potentially be any valid distribution. In practice, however, Q(z) is always set to Q*(z) a particular distribution to ensure the fastest possible convergence of the EM-algorithm. The idea is to ensure that $\frac{p(x,z)}{Q(z)}=p(x)$. Why is this?

    a) We are only interested in *x*, the observed variable.

    b) EM is a maximum-likelihood estimation algorithm, maximizing *p*(*x*) (i.e. $p(x|\theta)$) achieves this.

    c) *p*(*x*) is constant w.r.t. z, which means that in (1) the RHS=LHS

    d) This can help to overcome local minima issues

    e) None of the above

15. Which of the following statements about *mixture models* is false?

    a) They allow complex probability distributions to be more intuitively modelled

    b) They involve the combination of 2 more distributions

    c) They can be formulated as a latent variable problem

    d) Directly optimizing the likelihood functions of mixture models is often very difficult

    e) Mixture models generally only occur in 1-D or low dimensional spaces

… 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 a set of given data points are linearly separable, there exists at least one (d–1)-dimensional decision hyperplane to separate those data points.

   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. The magnitude (norm) of a vector must be always non-negative.

   e. None of the above.

17. In the perceptron algorithm (without offset), after certain number of updates, the parameter vector $\underline{\theta}$ is now [1.0, 2.0]. Then, the parameter vector $\underline{\theta}$ is updated for a data point $\underline{x}_i$ = [2.0, 1.0] whose label $y_i = -1$. What is the new value of $\underline{\theta}$?

   a. [3.0, 3.0]

   b. [–1.0, 1.0]

   c. [–1.0, –2.0]

   d. [–2.0, –1.0]

   e. None of the above.

18. Which of the following about the perceptron algorithm is **false**?

   a. The algorithm will always converge if the given data points are linearly separable.

   b. The number of updates to the parameter vector $\underline{\theta}$ cannot exceed R^2 / γ_g^2 where R is the radius of the minimum bounding sphere of all data points and γ_g is geometric margin (i.e., the distance between the decision boundary and the nearest data point).

   c. The algorithm can always guarantee to obtain the maximum geometric margin γ_g.

   d. The number of updates to the parameter vector $\underline{\theta}$ can sometimes be more than the number of given data points.

   e. None of the above.

19. What is the radius of the minimum bounding sphere (centered at the origin) for the data set composed of the following 4 data points?

$$\underline{x}_1 = [2.0, 1.0]$$
$$\underline{x}_2 = [-2.0, -2.0]$$
$$\underline{x}_3 = [-2.0, 1.0]$$
$$\underline{x}_4 = [2.0, -1.0]$$

   a. 5.0

   b. 8.0

   c. 2.828

   d. 2.236

   e. None of the above.

20. Suppose you have a data point $\underline{x}$ = [−1.0, −2.0] with its label $y = -1$. After running the perceptron algorithm (without offset) on a set of given data points including $\underline{x}$, you have the final parameter vector $\underline{\theta}^*$ = [3.0, 4.0]. What is the shortest distance between $\underline{x}$ and the decision boundary defined by $\underline{\theta}^*$?

   a. 2.2

   b. −2.2

   c. 0.4545

   d. −0.4545

   e. None of the above.

21. 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. $2K_1(\underline{a}, \underline{b})$

   b. $K_1(\underline{a}, \underline{b}) + (K_2(\underline{a}, \underline{b}))^2$

   c. $\underline{a}^2 K_1(\underline{a}, \underline{b})\, \underline{b}^2$

   d. $\underline{a}\, K_2(\underline{a}, \underline{b})\, \underline{b}$

   e. None of the above.

22. Suppose the number of vectors (data points) in a data set is 10 and the dimensionality of each vector is 2. Suppose we deploy an instance of SVM algorithm $\mathscr{A}$ on that data set and found that the number of support vectors is 4. What is the maximum possible error rate if we conduct a "ten-fold cross validation" on that data set using the same algorithm $\mathscr{A}$?

   a. 0.8

   b. 0.889

   c. 0.444

   d. 0.4

   e. None of the above.

23. If we construct a linear SVM (with an offset but without slacks) on the following 6 data points (vectors), what are the most likely ones to become the "support vectors"?

$$\underline{x}_1 = [2, 1], \quad y_1 = +1$$
$$\underline{x}_2 = [1, 2], \quad y_2 = +1$$
$$\underline{x}_3 = [3, 2], \quad y_3 = +1$$
$$\underline{x}_4 = [2, 3], \quad y_4 = -1$$
$$\underline{x}_5 = [1, 4], \quad y_5 = -1$$
$$\underline{x}_6 = [3, 4], \quad y_6 = -1$$

a. $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$, $\underline{x}_4$, $\underline{x}_5$, $\underline{x}_6$

b. $\underline{x}_2$, $\underline{x}_3$, $\underline{x}_4$

c. $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$

d. $\underline{x}_4$, $\underline{x}_5$, $\underline{x}_6$

e. None of the above.

24. Regarding a support vector machine, which of the following is **false**?

a. The objective of the primal SVM 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. The quadratic programming problem in SVM is one with quadratic objective and quadratic constraints.

d. If the training data points in d dimensional space are not linearly separable, we can try to map those data points into d' dimensional feature space (where $d' > d$) in which the points in the new feature space may become linearly separable.

e. None of the above.

25. Regarding a support vector machine, which of the following is **false**?

a. 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).

b. We can readily use kernel functions in the primal formulation of SVM.

c. The dual SVM formulation is a quadratic programming problem with simple box constraints.

d. After solving the dual SVM formulation, only those data points (vectors) whose resultant Lagrange multipliers (α_i*) are greater than 0 are regarded as the support vectors.

e. None of the above.

26. 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 cross product of the feature vectors $\phi(\underline{x}_1)$ and $\phi(\underline{x}_2)$, 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 $\phi(\underline{x}_1)$ and $\phi(\underline{x}_2)$ 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. It is also possible to define valid kernel functions to measure pairwise similarities of complex objects like strings, trees, and graphs.

   e. None of the above.

27. For a linear classifier in two dimensional space (where dimension #1 is denoted as X_1, and dimension #2 is denoted as X_2), the equation of the decision boundary is $X_1 - X_2 = 0$. If the coordinates of the nearest point to the decision boundary is $[X_1=2.0, X_2=1.0]$, what is the geometric margin (γ_g) of the classifier?

   a. 1

   b. $\sqrt{2}$

   c. $\sqrt{2}$

   d. 2

   e. None of the above.

28. For the Halving Algorithm, if $Q^{(t)}$ = {A, H, I, K, L, M, Z}, $Q_+^{(t)}$ = {A, I, L, Z}, $Q_-^{(t)}$ = {H, K, M}, and the actual label $y^{(t)} = -1$, what will be $Q^{(t+1)}$?

   a. {A, H, I, K, L, M, Z}

   b. {A, I, L, Z}

   c. {H, K, M}

   d. { }

   e. None of the above.

29. For the Halving Algorithm, if there are 20 experts and at least one of them is a "super expert" who always makes correct predictions, what is the maximum number of prediction error that the algorithm can commit?

   a. 4

   b. 5

   c. 2

   d. 20

   e. None of the above.

30. For the Weighted Majority Algorithm, 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 $w^{(t+1)}$?

    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.

Appendix

Perceptron algorithm

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

Repeat until convergence:

for $t = 1, \ldots, n$

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

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

The Halving Algorithm

- Initialization: $\mathcal{Q}^{(1)} = \{1, 2, 3, \ldots d\}$
- For $t = 1 \ldots T$
  1. I receive some input $\underline{x}^{(t)}$
  2. Define
  $$\mathcal{Q}^{(t)}_{+} = \{j \in \mathcal{Q}^{(t)} : x^{(t)}_j = +1\}$$
  $$\mathcal{Q}^{(t)}_{-} = \{j \in \mathcal{Q}^{(t)} : x^{(t)}_j = -1\}$$
  If $|\mathcal{Q}^{(t)}_{+}| > |\mathcal{Q}^{(t)}_{-}|$ 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: if $y^{(t)} = +1$ then $\mathcal{Q}^{(t+1)} = \mathcal{Q}^{(t)}_{+}$, else $\mathcal{Q}^{(t+1)} = \mathcal{Q}^{(t)}_{-}$.

The Weighted Majority Algorithm

- Parameter: $0 < \beta < 1$
- Initialization: set $w_j = 1$ for $j = 1 \ldots d$.
- For $t = 1 \ldots T$
  1. I receive some input $\underline{x}^{(t)}$
  2. Define
  $$q^{(t)}_{+} = \sum_{j : x^{(t)}_j = +1} w_j; \quad q^{(t)}_{-} = \sum_{j : x^{(t)}_j = -1} w_j$$
  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 $x^{(t)}_j \neq y^{(t)}$, set $w_j = w_j \times \beta$