Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering

CS60050 : Machine Learning
Spring 2021 | Long Test 2 (ALL Q&A) | Marks : 50
Date : 05-Apr-2021 (Monday) | Time : 8:30pm-9:45pm (75 min)

Question-1: [Hierarchical Clustering] [Marks: 4 + 4 = 8]

Suppose, six points (P1, P2, P3, P4, P5 and P6) are provided in a 2-D plane. The Euclidean distance between a pair of these points are provided in the table below.

	P1	P2	P3	P4	P5	P6
P1	0.00	0.24	0.22	0.37	0.34	0.23
P2	0.24	0.00	0.15	0.20	0.14	0.25
P3	0.22	0.15	0.00	0.15	0.28	0.11
P4	0.37	0.20	0.15	0.00	0.29	0.22
P5	0.34	0.14	0.28	0.29	0.00	0.39
P6	0.23	0.25	0.11	0.22	0.39	0.00

Answer the following questions:

(i) If you use Single Linkage Hierarchical Agglomerative Clustering technique to form the single-link dendrogram, initially each point will form separate clusters, {P1}, {P2}, {P3}, {P4}, {P5} and {P6}. Then, at the first (bottom-most grouping) phase, the algorithm selects {P3} and {P6} clusters to merge and form new cluster {P36}, as the distace considered for grouping here was, dist(P3, P6) = 0.11.
Now, you need to complete the rest of the bottom-up phases (Phase-2 to Phase-5) by mentioning the next new cluster formed and the distance considered that time. [4 x (0.5 x 2) = 4 marks]

Note / Notation:
-- To denote the clusters, please write {P123} to indicate new clusters of {P1, P2, P3}; {P14} to indicate new clusters of {P1, P4}; etc. -- that is, within brackets ({}), P followed by participating point-indices in sorted ascending order.

-- For REAL valued answers (distance), give the approximated results upto THREE places after the decimal point.

Complete the following table:

Phases -->	Phase-1	Phase-2	Phase-3	Phase-4	Phase-5
New Cluster	{P36}				
Distance	0.11				

(ii) If you use Complete Linkage Hierarchical Agglomerative Clustering technique to form the complete-link dendrogram, initially each point will form separate clusters, {P1}, {P2}, {P3}, {P4}, {P5} and {P6}. Then, at the first (bottom-most grouping) phase, the algorithm selects {P3} and {P6} clusters to merge and form new cluster {P36}, as the distace considered for grouping here was, dist(P3, P6) = 0.11.
Now, you need to complete the rest of the bottom-up phases (Phase-2 to Phase-5) by mentioning the next new cluster formed and the distance considered that time.
[4 x (0.5 x 2) = 4 marks]

Note / Notation:
-- To denote the clusters, please write {P123} to indicate new clusters of {P1, P2, P3}; {P14} to indicate new clusters of {P1, P4}; etc. -- that is, within brackets ({}), P followed by participating point-indices in sorted ascending order.

-- For REAL valued answers (distance), give the approximated results upto THREE places after the decimal point.

Complete the following table:

Phases -->	Phase-1	Phase-2	Phase-3	Phase-4	Phase-5
New Cluster	{P36}				
Distance	0.11				

Answers:
(i)

Phases -->	Phase-1	Phase-2	Phase-3	Phase-4	Phase-5
New Cluster	{P36}	{P25}	{P2356} or P{346}	{P23456}	{P123456}
Distance	0.11	0.14	0.15	0.15	0.22

(ii)

Phases -->	Phase-1	Phase-2	Phase-3	Phase-4	Phase-5
New Cluster	{P36}	{P25}	{P346}	{P125}	{P123456}
Distance	0.11	0.14	0.22	0.34	0.39

Explanations:
(i) Phase-2: The minimum among the minimum distances between a pair of existing clusters, [{P1},{P2},{P36},{P4},{P5}], is = dist({P2},{P5}) = 0.14.
Phase-3: The minimum among the minimum distances between a pair of existing clusters, [{P1},{P25},{P36},{P4}], is = MIN[dist({P25},{P36})] = MIN[dist({P2},{P3}), dist({P2},{P6}), dist({P5},{P3}), dist({P5},{P6})] = 0.15.
Phase-4: The minimum among the minimum distances between a pair of existing clusters, [{P1},{P2356},{P4}], is = MIN[dist({P2356},{P4})] = MIN[dist({P2},{P4}), dist({P3},{P4}), dist({P5},{P4}), dist({P6},{P4})] = 0.15.
Phase-5: The minimum among the minimum distances between a pair of existing clusters, [{P1},{P23456}], is = MIN[dist({P23456},{P1})] = MIN[dist({P2},{P1}), dist({P3},{P1}), dist({P4},{P1}), dist({P5},{P1}), dist({P6},{P1})] = 0.22.

(ii) Phase-2: The minimum among the maximum distances between a pair of existing clusters, [{P1},{P2},{P36},{P4},{P5}], is = dist({P2},{P5}) = 0.14.

Phase-3: The minimum among the maximum distances between a pair of existing clusters, [{P1},{P25},{P36},{P4}], is = MAX[dist({P36},{P4})] = MAX[dist({P3},{P4}), dist({P6},{P4})] = 0.22.

Phase-4: The minimum among the maximum distances between a pair of existing clusters, [{P1},{P25},{P346}], is = MAX[dist({P1},{P25})] = MAX[dist({P1},{P2}), dist({P1},{P5})] = 0.34.

Phase-5: The minimum among the maximum distances between a pair of existing clusters, [{P125},{P346}], is = MAX[dist({P125},{P346})] = MAX[dist({P1},{P3}), dist({P1},{P4}), dist({P1},{P6}), dist({P2},{P3}), dist({P2},{P4}), dist({P2},{P6}), dist({P5},{P3}), dist({P5},{P4}), dist({P5},{P6})] = 0.39.

Question-2: [Clustering Evaluation] [Marks: 4 + 1 = 5]

Suppose, four points (P1, P2, P3 and P4) are grouped into two separate clusters, where Cluster-1 contains {P1,P2} and Cluster-2 contains {P3,P4}. The distance (or dissimilarity) between these points are provided in the table below.

	P1	P2	P3	P4
P1	0.00	0.10	0.65	0.55
P2	0.10	0.00	0.70	0.60
P3	0.65	0.70	0.00	0.30
P4	0.55	0.60	0.30	0.00

Answer the following questions:

(i) Compute the silhouette coefficient (score) for each of the four points. [4 x 1 = 4 marks]

(ii) Compute the average silhouette coefficient of the overall clustering (considering the two cluster). [1 mark]

Note: For REAL valued answers, give the approximated results upto THREE places after the decimal point.

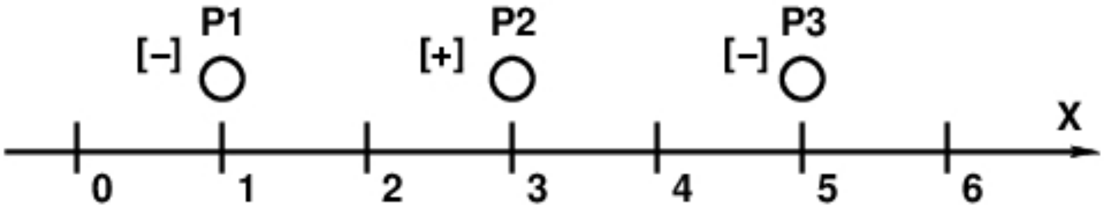
Answers:
(i) SC(P1) = 0.833 , SC(P2) = 0.846 , SC(P3) = 0.556 , SC(P4) = 0.478
(ii) Overall Average SC = 0.67825

Explanations:
(i) Let a indicate the average distance of a point to other points in its cluster, and b indicate the minimum of the average distance of a point to points in another cluster.

```
    Point P1: SC = 1- a/b = 1 - 0.1/((0.65+0.55)/2) = 0.833
    Point P2: SC = 1- a/b = 1 - 0.1/((0.70+0.60)/2) = 0.846
    Point P3: SC = 1- a/b = 1 - 0.3/((0.65+0.70)/2) = 0.556
    Point P4: SC = 1- a/b = 1 - 0.3/((0.55+0.60)/2) = 0.478

(ii) Cluster-1: Average SC = (0.833+0.846)/2  = 0.8395
     Cluster-2: Average SC = (0.556+0.478)/2  = 0.517
     Overall:   Average SC = (0.8395+0.517)/2 = 0.67825
```

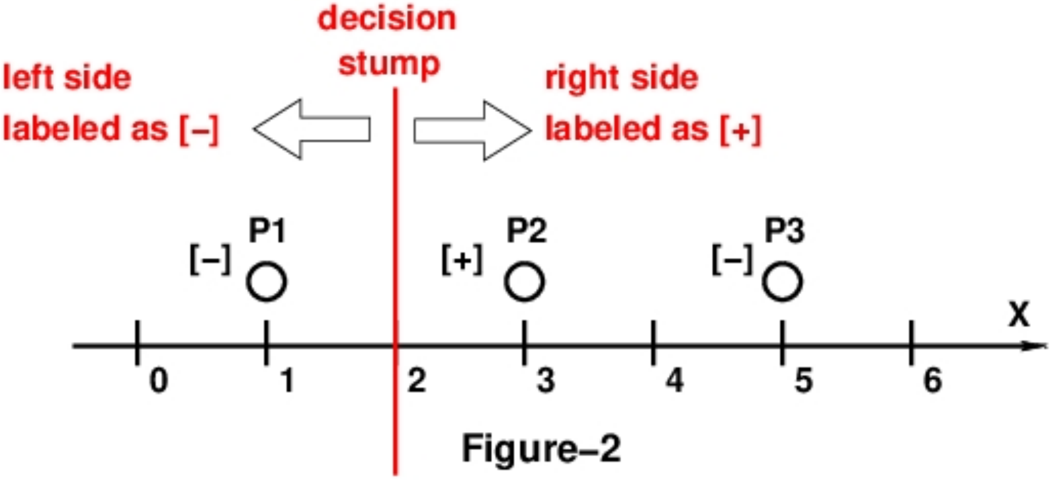
Question-3: [Boosting]   [Marks: 1 + 1 + 1 + 3 + 2 = 8]

In this problem, we study how boosting algorithm performs on a very simple classification problem shown in Figure 1.



Figure–1

In the above figure, three training points (indicated using 'o'), P1, P2 and P3, are given in a 1-D line with their corresponding 2-class (+/-) labels as [-], [+] and [-], respectively.



Figure–2

We shall use decision stumps as our weak learner / hypothesis. Decision stump classifier chooses a constant value c and classifies all points where x > c as one class and other points where x ≤ c as the other class. One such decision stump (where x > 2 region is classified as [+] zone and x ≤ 2 region is classified as [-] zone) is mentioned above in Figure-2.

Answer the following questions:

Note: For REAL valued answers, give the approximated results upto THREE places after the decimal point.

(i) What is the initial weight assigned to each data point? [1 mark]

(ii) How many different decision stumps are possible for the data points given in Figure-1? [1 mark]

(iii) Which point(s) will have weights increased after the boosting process as per the decision stump considered in Figure-2? [1 mark]
Choose all the correct options:

(a) P1
(b) P2
(c) P3
(d) NO points
(e) ALL points

(iv) What will be weights of the data points after boosting is done? [3 x 1 = 3 marks]

Weight(P1) = ?
Weight(P2) = ?
Weight(P3) = ?

(v) Indicate whether the following statements are TRUE / FALSE? [2 x 1 = 2 marks]

[A] Boosting algorithm cannot perfectly classify all the training examples given in Figure-1 (above).

[B] The training error of boosting classifier (combination of all the weak classifier)
monotonically decreases as the number of iterations in the boosting algorithm increases.

Answers:
(i) Weight = 1/3 = 0.333
(ii) 6
(iii) Option-(c) P3
(iv) Weight(P1) = 0.25 , Weight(P2) = 0.25 , Weight(P3) = 0.5
(v) (a) TRUE , (b) FALSE

Explanations:

(i) As, Weight(P1) = Weight(P2) = Weight(P3) and
Weight(P1) + Weight(P2) + Weight(P3) = 1,
So, Weight(P1) = Weight(P2) = Weight(P3) = 1/3 = 0.333

(ii) (3 separators/stumps) x (2 different class organizations for each) = 6

(iii) The given decision stump in Figure-2 will missclassify P3 as [+].

(iv) $\epsilon_t = 1/3$ and $\alpha_t = (1/2).\ln[(1-1/3)/(1/3)] = (1/2)\ln(2) = 0.3465$.
Also, normalization factor, $Z = 2.\sqrt{[\epsilon_t.(1-\epsilon_t)]} = 2.\sqrt{[(1/3).(2/3)]}$
$= 2\sqrt{2}/3 = 0.94281$
So, for correctly classified data-points (P1 and P2),
Weight(P1) = Weight(P2) = (1/Z) . [(1/3).exp(-0.3465)] = 0.25
Weight(P3) = (1/Z) . [(1/3).exp(0.3465)] = 0.5

(v) (a) TRUE, since the data is not linearly separable.

(b) FALSE, since boosting minimizes loss function: $\sum_{i=1}^{m} \exp(-y_i . f(x_i))$
which does not necessary mean that training error monotonically decrease.

Question-4: [Kernels] [Marks: 4 + 3 + 2 + 2 = 11]

Answer the following questions:

(i) Consider two finite-dimensional feature transform Φ_1 and Φ_2 and their corresponding Kernels K_1 and K_2. [2 x 2 = 4 marks]

[A] Define $\Phi(x) = [\Phi_1(x),\Phi_2(x)]$. Express the corresponding Kernel of Φ in terms of K_1 and K_2.
Choose the correct option:
(a) $K_1 + K_2$
(b) K_1K_2
(c) $\sqrt{2}K_1K_2$

(d) $\sqrt{2}K_1 + \sqrt{2}K_2$

[B] Consider the matrix $\phi_1(x).\phi_2(x)^T$ and let $\phi(x)$ be the vector representation of the matrix (say, by concatenating all the rows). Express the corresponding Kernel ϕ in terms of K_1 and K_2.
Choose the correct option:
(a) $K_1 + K_2$
(b) K_1K_2
(c) $\sqrt{2}K_1K_2$
(d) $\sqrt{2}K_1 + \sqrt{2}K_2$

(ii) A kernel function $K(x,z)$ measures the similarity between two instances x and z in a transformed space. For a feature transform $x \rightarrow \phi(x)$ the kernel function is $K(x,z) = \phi(x).\phi(z)$. Consider the two dimensional input vectors $x = (x_1,x_2)$. For each of the kernel function below what is the corresponding feature transform? [3 x 1 = 3 marks]

[A] $K(x,z) = 1 + x.z$
Choose the correct option:
(a) $\phi(x) = (x_1,x_2)$
(b) $\phi(x) = (1,x_1,x_2)$
(c) $\phi(x) = (x_1^2,x_2^2)$
(d) $\phi(x) = (1,x_1^2,x_2^2)$

[B] $K(x,z) = (x.z)^2$
Choose the correct option:
(a) $\phi(x) = (x_1^2,x_2^2)$
(b) $\phi(x) = (1,x_1^2,x_2^2)$
(c) $\phi(x) = (x_1^2,x_2^2,\sqrt{2}x_1x_2)$
(d) $\phi(x) = (1,x_1^2,x_2^2,\sqrt{2}x_1x_2)$

[C] $K(x,z) = (1 + x.z)^2$
Choose the correct option:
(a) $\phi(x) = (1,x_1^2,x_2^2)$
(b) $\phi(x) = (1,x_1^2,x_2^2,\sqrt{2}x_1x_2)$
(c) $\phi(x) = (1,x_1^2,x_2^2,\sqrt{2}x_1x_2,\sqrt{2}x_1,\sqrt{2}x_2)$
(d) $\phi(x) = (1,x_1^2,x_2^2,\sqrt{2}x_1x_2,x_1,x_2)$

(iii) Multiple kernels can be combined to produce new kernels. For example, $K(x,z) = K_1(x,z) + K_2(x,z)$ is a valid combination. Suppose kernel K_1 has the associated feature transformation ϕ_1 and K_2 has the associated feature transformation ϕ_2. What is the feature transform associated with the combinations given below?
[2 x 1 = 2 marks]

[A] $K(x,z) = \alpha.K_1(x,z)$
Choose the correct option:
(a) $\phi(x) = \phi_1(x)$
(b) $\phi(x) = \alpha^2.\phi_1(x)$
(c) $\phi(x) = \alpha.\phi_1(x)$
(d) $\phi(x) = \sqrt{\alpha}.\phi_1(x)$

[B] $K(x,z) = \alpha.K_1(x,z) + \beta.K_2(x,z)$
Choose the correct option:
(a) $\phi(x) = \alpha.\phi_1(x) + \beta.\phi_2(x)$
(b) $\phi(x) = \sqrt{\alpha}.\phi_1(x) + \sqrt{\beta}.\phi_2(x)$
(c) $\phi(x) = [\ \alpha.\phi_1(x)\ ,\ \beta.\phi_2(x)\ ]$
(d) $\phi(x) = [\ \sqrt{\alpha}.\phi_1(x)\ ,\ \sqrt{\beta}.\phi_2(x)\ ]$

(iv) One of the most commonly used kernels in SVM is the Gaussian RBF kernel:
$k(a,b) = \exp(- ||a-b||^2 / 2\sigma)$. Suppose we have three points, z_1, z_2, and x. z_1 i

s geometrically very close to x, and z_2 is geometrically far away from x. What is the value of $k(z_1,x)$ and $k(z_2,x)$? [2 marks]

Choose the correct option:

(a) $k(z_1,x)$ will be close to 1 and $k(z_2,x)$ will be close to 0.

(b) $k(z_1,x)$ will be close to 0 and $k(z_2,x)$ will be close to 1.

(c) $k(z_1,x)$ will be close to c_1, $c_1 >> 1$ and
$k(z_2,x)$ will be close to c_2, $c_2 << 0$ (where $c_1,c_2 \in \mathbb{R}$)

(d) $k(z_1,x)$ will be close to c_1, $c_1 << 0$ and
$k(z_2,x)$ will be close to c_2, $c_2 >> 1$ (where $c_1,c_2 \in \mathbb{R}$)

Answers:

(i) [A] Option-(a) $K_1 + K_2$
[B] Option-(b) K_1K_2

(ii) [A] Option-(b) $\phi(x) = (1,x_1,x_2)$
[B] Option-(c) $\phi(x) = (x_1^2,x_2^2,\sqrt{2}x_1x_2)$
[C] Option-(c) $\phi(x) = (1,x_1^2,x_2^2,\sqrt{2}x_1x_2,\sqrt{2}x_1,\sqrt{2}x_2)$

(iii) [A] Option-(d) $\phi(x) = \sqrt{\alpha}.\phi_1(x)$
[B] Option-(d) $\phi(x) = [\ \sqrt{\alpha}.\phi_1(x)\ ,\ \sqrt{\beta}.\phi_2(x)\ ]$

(iv) Option-(a) $k(z_1,x)$ will be close to 1 and $k(z_2,x)$ will be close to 0.

Explanations:

(i) [A] We have, $K_1 = \phi_1^T.\phi_1$ and $K_2 = \phi_2^T.\phi_2$
Now, $\phi = \phi^T.\phi = [\phi_1^T \quad \phi_2^T].[\phi_1 \quad \phi_2]^T = \phi_1^T.\phi_1 + \phi_2^T.\phi_2 = K_1 + K_2$

[B] Let $\phi_1 = [a_1 \ ... \ a_n]^T$ and $\phi_2 = [b_1 \ ... \ b_m]^T$.
So, we have $K_1 = \phi_1^T.\phi_1 = \sum a_i^2$ and $K_2 = \phi_2^T.\phi_2 = \sum b_i^2$.
Therefore, $\phi = \phi_1.\phi_2^T = [v_1 \ ... \ v_n]^T$ where $v_n = [a_nb_1 \ ... \ a_nb_m]^T$.
Hence, we get $\phi = \phi^T.\phi = \sum v_i^T v_i = \sum_n \sum_m a_n^2 b_m^2 = \sum_n a_n^2 \sum_m b_m^2 = K_1K_2$.

(ii) [A] $K(x,z) = \phi(x)\phi(z) = [1,x_1,x_2].[1,z_1,z_2]^T = 1 + x_1.z_1 + x_2.z_2 = 1 + x.z$

[B] $K(x,z) = \phi(x)\phi(z) = [x_1^2,x_2^2,\sqrt{2}x_1x_2].[z_1^2,z_2^2,\sqrt{2}z_1z_2]^T$
$= x_1^2.z_1^2 + x_2^2.z_2^2 + 2x_1x_2z_1z_2 = (x_1.z_1 + x_2.z_2)^2 = (x.z)^2$

[C] $K(x,z) = \phi(x)\phi(z)$
$= [1,x_1^2,x_2^2,\sqrt{2}x_1x_2,\sqrt{2}x_1,\sqrt{2}x_2].[1,z_1^2,z_2^2,\sqrt{2}z_1z_2,\sqrt{2}z_1,\sqrt{2}z_2]^T$
$= 1 + x_1^2.z_1^2 + x_2^2.z_2^2 + 2x_1x_2z_1z_2 + 2x_1z_1 + 2x_2z_2$
$= (1 + x_1.z_1 + x_2.z_2)^2 = (1 + x.z)^2$

(iii) [A] $K(x,z) = \phi(x)\phi(z) = [\sqrt{\alpha}.\phi_1(x)].[\sqrt{\alpha}.\phi_1(z)] = \alpha.\phi_1(x)\phi_1(z) = \alpha.K_1(x,z)$

[B] $K(x,z) = \phi(x)\phi(z) = [\ \sqrt{\alpha}.\phi_1(x)\ ,\ \sqrt{\beta}.\phi_2(x)\ ].[\ \sqrt{\alpha}.\phi_1(z)\ ,\ \sqrt{\beta}.\phi_2(z)\ ]^T$
$= \alpha.\phi_1(x)\phi_1(z) + \beta.\phi_2(x)\phi_2(z) = \alpha.K_1(x,z) + \beta.K_2(x,z)$

(iv) RBF kernel generates a 'bump' around the center x. For points z_1 close to the center of the bump, $k(z_1,x)$ will be close to 1, for points away from the center of the bump $K(z_2,x)$ will be close to 0.

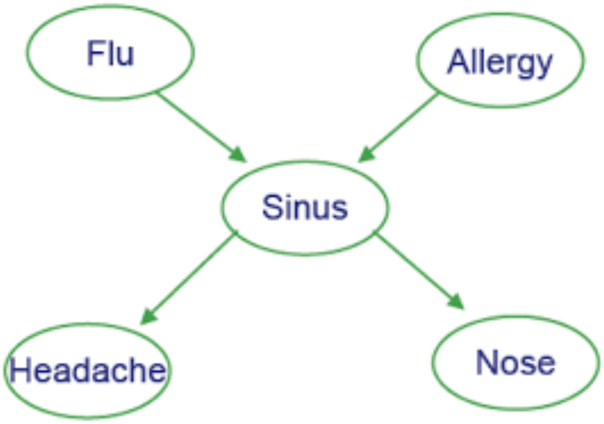
Question-5: [Expectation-Maximization Algorithm] [Marks: 1 + 2 + 3 + 2 = 8]

Consider the following Bayes Net structure (Figure in the right-side of Table). In the figure, we abbreviate as follows: F = Flu, A = Allergy, S = Sinus, H = Headache, N = Nose-running.

We are given the following K = 8 training examples as shown below, where only two examples contain unobserved values, namely, h_7 and n_8. We like to simulate a few steps of the simplified EM algorithm by hand.

k	F	A	S	H	N
k=1	1	0	1	1	1

k=2	0	1	1	1	0
k=3	1	1	1	1	1
k=4	0	0	0	0	1
k=5	0	0	0	1	0
k=6	0	0	0	0	0
k=7	1	1	1	?	1
k=8	1	1	1	1	?



Notation:
Here, f_k, a_k, s_k, h_k, n_k indicate the values of F, A, S, H, N, respectively, as seen in the k-th example/row. For example, $f_1 = 1$, $a_1 = 0$, $s_1 = 1$, $h_1 = 1$, $n_1 = 1$.

Answer the following questions:

Note: For REAL valued answers, give the approximated results upto THREE places after the decimal point.

(i) Given that all variables are Boolean, how many basic parameters we need to estimate for the given Bayes Net? For example, one parameter will be $\theta(s|11)$, which stands for Prob(S = 1 | F = 1, A = 1). [1 mark]

(ii) Now, we like to simulate the first E-step of the EM algorithm. Before we start, we initialize all the parameters as 0.5, and then proceed to execute the E-step. What are the following expectation values that will get calculated in this E-step? [2 x 1 = 2 marks]
(Note that, only two examples (k=7 and k=8) contains unobserved variables, where h_7 = ?, but $f_7 = a_7 = s_7 = n_7 = 1$; and n_8 = ?, but $f_8 = a_8 = s_8 = h_8 = 1$, respectively.)
[A] Calculate: $E(h_7=1 \mid f_7, a_7, s_7, n_7, \theta)$ = ?
[B] Calculate: $E(n_8=1 \mid f_8, a_8, s_8, h_8, \theta)$ = ?

(iii) Now, we like to simulate the first M-step of the EM algorithm. What will be the estimated values of the following model parameters that we obtain in this M-step? [3 x 1 = 3 marks]
(Note that, we use the expected count only when the variable is unobserved in an example)
[A] Calculate: $\theta(s|11)$ = Prob(S = 1 | F = 1, A = 1) = ?
[B] Calculate: $\theta(h|0)$ = Prob(H = 1 | S = 0) = ?
[C] Calculate: $\theta(n|0)$ = Prob(N = 1 | S = 0) = ?

(iv) Last, let us (again) simulate the second E-step of the EM algorithm. What are the following expectation values that will get calculated in this E-step? [2 x 1 = 2 marks]
[A] Calculate: $E(h_7=1 \mid f_7, a_7, s_7, n_7, \theta)$ = ?
[B] Calculate: $E(n_8=1 \mid f_8, a_8, s_8, h_8, \theta)$ = ?

\~

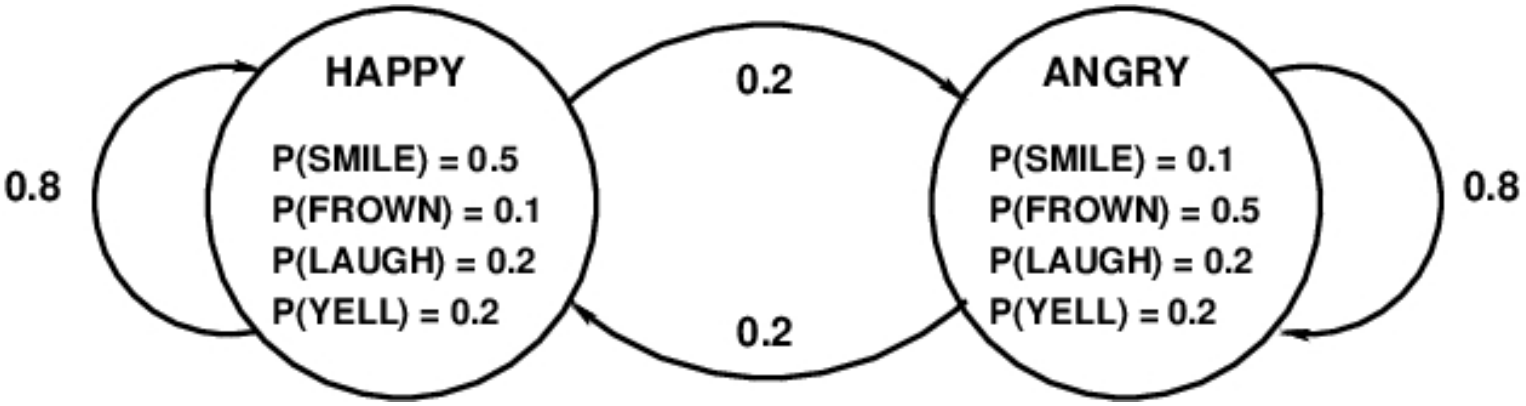
Answers:
(i) 10
(ii) [A] $E(h_7=1 \mid f_7, a_7, s_7, n_7, \theta) = 0.5$
     [B] $E(n_8=1 \mid f_8, a_8, s_8, h_8, \theta) = 0.5$
(iii) [A] $\theta(s|11)$ = Prob(S = 1 | F = 1, A = 1) = 1.0
      [B] $\theta(h|0)$ = Prob(H = 1 | S = 0) = 0.333

```
      [C] θ(n|0) = Prob(N = 1 | S = 0) = 0.333
(iv) [A] E(h₇=1 | f₇, a₇, s₇, n₇, θ) = 0.9
     [B] E(n₈=1 | f₈, a₈, s₈, h₈, θ) = 0.7
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Explanations:
(i) 10 parameters are as follows:
    θ(f) = Prob(F = 1)
    θ(a) = Prob(A = 1)
    θ(s|00) = Prob(S = 1|F = 0, A = 0)
    θ(s|01) = Prob(S = 1|F = 0, A = 1)
    θ(s|10) = Prob(S = 1|F = 1, A = 0)
    θ(s|11) = Prob(S = 1|F = 1, A = 1)
    θ(h|0) = Prob(H = 1|S = 0)
    θ(h|1) = Prob(H = 1|S = 1)
    θ(n|0) = Prob(N = 1|S = 0)
    θ(n|1) = Prob(N = 1|S = 1)

(ii) [A] E(h₇=1 | f₇, a₇, s₇, n₇, θ)
                        Prob(h₇ = 1, f₇, a₇, s₇, n₇ | θ)
     = -------------------------------------------------------------------
       Prob(h₇ = 1, f₇, a₇, s₇, n₇ | θ) + Prob(h₇ = 0, f₇, a₇, s₇, n₇ | θ)
                      θ(h₇=1|s₇).θ(s₇|f₇,a₇).θ(f₇).θ(a₇)
     = -----------------------------------------------------------------------
       θ(h₇=1|s₇).θ(s₇|f₇,a₇).θ(f₇).θ(a₇) + θ(h₇=0|s₇).θ(s₇|f₇,a₇).θ(f₇).θ(a₇)
         0.5 x 0.5 x 0.5 x 0.5
     = ------------------------- = 0.5                = E(h₇=1 | s₇=1, θ(h|1))
       2 x 0.5 x 0.5 x 0.5 x 0.5

     [B] E(n₈=1 | f₈, a₈, s₈, h₈, θ)
                        Prob(n₈ = 1, f₈, a₈, s₈, h₈ | θ)
     = -------------------------------------------------------------------
       Prob(n₈ = 1, f₈, a₈, s₈, h₈ | θ) + Prob(n₈ = 0, f₈, a₈, s₈, h₈ | θ)
                      θ(n₈=1|s₈).θ(s₈|f₈,a₈).θ(f₈).θ(a₈)
     = -----------------------------------------------------------------------
       θ(n₈=1|s₈).θ(s₈|f₈,a₈).θ(f₈).θ(a₈) + θ(n₈=0|s₈).θ(s₈|f₈,a₈).θ(f₈).θ(a₈)
         0.5 x 0.5 x 0.5 x 0.5
     = ------------------------- = 0.5                = E(n₈=1 | s₈=1, θ(n|1))
       2 x 0.5 x 0.5 x 0.5 x 0.5

(iii) 10 parameters will get the updated values as follows:
    θ(f) = Prob(F = 1) = #{F=1} / #K = 4/8 = 0.5
    θ(a) = Prob(A = 1) = #{A=1} / #K = 4/8 = 0.5
    θ(s|00) = Prob(S = 1|F = 0, A = 0) = #{S=1,F=0,A=0} / #{F=0,A=0} = 0/3 = 0.0
    θ(s|01) = Prob(S = 1|F = 0, A = 1) = #{S=1,F=0,A=1} / #{F=0,A=1} = 1/1 = 1.0
    θ(s|10) = Prob(S = 1|F = 1, A = 0) = #{S=1,F=1,A=0} / #{F=1,A=0} = 1/1 = 1.0
    θ(s|11) = Prob(S = 1|F = 1, A = 1) = #{S=1,F=1,A=1} / #{F=1,A=1} = 3/3 = 1.0
    θ(h|0) = Prob(H = 1|S = 0) = #{S=0}.E[H=1] / #{S=0}
                               = (1 x 1.0 + 2 x 0.0) / 3 = 0.333
    θ(h|1) = Prob(H = 1|S = 1) = #{S=1}.E[H=1] / #{S=1}
                               = (4 x 1.0 + 1 x 0.5) / 5 = 0.9
    θ(n|0) = Prob(N = 1|S = 0) = #{S=0}.E[N=1] / #{S=0}
                               = (1 x 1.0 + 2 x 0.0) / 3 = 0.333
    θ(n|1) = Prob(N = 1|S = 1) = #{S=1}.E[N=1] / #{S=1}
                               = (3 x 1.0 + 1 x 0.0 + 1 x 0.0) / 5 = 0.7

(iv) [A] E(h₇=1 | f₇, a₇, s₇, n₇, θ)
                        Prob(h₇ = 1, f₇, a₇, s₇, n₇ | θ)
     = -------------------------------------------------------------------
       Prob(h₇ = 1, f₇, a₇, s₇, n₇ | θ) + Prob(h₇ = 0, f₇, a₇, s₇, n₇ | θ)
                      θ(h₇=1|s₇).θ(s₇|f₇,a₇).θ(f₇).θ(a₇)
```

```
    = ---------------------------------------------------------------------
      θ(h₇=1|s₇).θ(s₇|f₇,a₇).θ(f₇).θ(a₇) + θ(h₇=0|s₇).θ(s₇|f₇,a₇).θ(f₇).θ(a₇)
        0.9 x 1.0 x 0.5 x 0.5
    = ------------------------- = 0.9                = E(h₇=1 | s₇=1, θ(h|1))
        1.0 x 1.0 x 0.5 x 0.5

    [B] E(n₈=1 | f₈, a₈, s₈, h₈, θ)
                       Prob(n₈ = 1, f₈, a₈, s₈, h₈ | θ)
    = -----------------------------------------------------------------
      Prob(n₈ = 1, f₈, a₈, s₈, h₈ | θ) + Prob(n₈ = 0, f₈, a₈, s₈, h₈ | θ)
                     θ(n₈=1|s₈).θ(s₈|f₈,a₈).θ(f₈).θ(a₈)
    = ---------------------------------------------------------------------
      θ(n₈=1|s₈).θ(s₈|f₈,a₈).θ(f₈).θ(a₈) + θ(n₈=0|s₈).θ(s₈|f₈,a₈).θ(f₈).θ(a₈)
        0.7 x 1.0 x 0.5 x 0.5
    = ------------------------- = 0.7                = E(n₈=1 | s₈=1, θ(n|1))
        1.0 x 1.0 x 0.5 x 0.5
================================================================================
```

```
================================================================================
Question-6:  [ Hidden Markov Model ]            [ Marks: 1 + 2 + 2 + 2 + 1 = 8 ]
================================================================================
```

Suppose you live a very simple life and have only two emotional states, ANGRY and HAPPY. Some days you are ANGRY and some days you remain HAPPY. However, you hide your emotional state and others can only observe this from whether you SMILE, FROWN, LAUGH or YELL. Suppose, you start on Day-1 at a HAPPY state and there is one transition per day. So, your emotional model (states and transitions with probabilities) and the probabilities of the observations at every state are given in the following Figure.



In the above figure, the probabilities of transition from HAPPY to HAPPY and ANGRY to ANGRY are both 0.8, and that of transition from HAPPY to ANGRY and ANGRY to HAPPY are both 0.2. The observation output probabilities are also mentioned at each state.

Let us also define, (a) Q(t) = State at Day-t, and (b) O(t) = Observations on Day-t.

Answer the following questions:

Note: For REAL valued answers, give the approximated results upto THREE places after the decimal point.

(i) Calculate: Prob[Q(2)=HAPPY] = ? [1 mark]

```
(ii) Calculate: Prob[O(2)=FROWN] = ?  [2 marks]

(iii) Calculate: Prob[Q(2)=HAPPY | O(2)=FROWN] = ?  [2 marks]

(iv) Calculate: Prob[O(100)=YELL] = ?  [2 marks]

(v) Assume O(1) = FROWN, O(2) = FROWN, O(3) = FROWN, O(4) = FROWN, O(5) = FROWN,
 what is the most likely sequence of states?  [1 mark]
Choose the correct option:
  (a) ANGRY, ANGRY, ANGRY, ANGRY, ANGRY
  (b) HAPPY, ANGRY, ANGRY, ANGRY, ANGRY
  (c) HAPPY, HAPPY, ANGRY, ANGRY, ANGRY
  (d) HAPPY, ANGRY, HAPPY, ANGRY, ANGRY
  (e) HAPPY, ANGRY, ANGRY, HAPPY, ANGRY
  (f) HAPPY, ANGRY, ANGRY, ANGRY, HAPPY
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Answers:
(i) 0.8
(ii) 0.18
(iii) 0.444
(iv) 0.2
(v) Option-(b) HAPPY, ANGRY, ANGRY, ANGRY, ANGRY
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Explanations:
(i) If Q(1) = HAPPY (i.e. you start from HAPPY state), then on Day-2, you can re
ach to HAPPY state by direct transition having 0.8 probability value.
Hence, Prob[ Q(2) = HAPPY ] = 0.8.

(ii) O(2) = FROWN can occur in Day-2 by either staying in HAPPY state or transit
ing to ANGRY state. Hence, Prob[ O(2) = FROWN] = 0.8 x 0.1 + 0.2 x 0.5 = 0.18

(iii)                           Prob[O(2)=FROWN | Q(2)=HAPPY] x Prob[Q(2)=HAPPY]

Prob[Q(2)=HAPPY | O(2)=FROWN] = ------------------------------------------------
                                               Prob[ O(2) = FROWN ]
= (0.1 x 0.8) / 0.18 = 0.444

(iv) Prob[O(100)=YELL] = Prob[O(100)=YELL | Q(100)=HAPPY] x Prob[Q(100)=HAPPY] +
                         Prob[O(100)=YELL | Q(100)=ANGRY] x Prob[Q(100)=ANGRY]
                       = 0.2 x Prob[Q(100)=HAPPY] + 0.2 x Prob[Q(100)=ANGRY]
                       = 0.2 (Prob[Q(100)=HAPPY] + Prob[Q(100)=ANGRY])
                       = 0.2 x 1 = 0.2

(v) Left as an exercise!
================================================================================



================================================================================
Question-7: [ Principal Component Analysis ]                      [ Marks: 2 ]
================================================================================
Given three data points in two-dimensional space: (1,1), (2,2), and (3,3), the f
irst principal component is given by (1/√2,a) or (0.70711,a). What is the value 
of a?  [2 marks]

Note: For REAL valued answers, give the approximated results upto THREE places a
fter the decimal point.
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Answers:
  1/√2 = 0.70711
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Explanations:
The 2-D points given and their mean value is computed are as follows.

	x1	x2
P1	1	1
P2	2	2
P3	3	3
Mean	2	2

Now, the co-variance matrix is shown below.

```
CoVar = [ CoVar(x1,x1) CoVar(x1,x2) ] = [ 1 1 ]
        [ CoVar(x2,x1) CoVar(x2,x2) ] = [ 1 1 ]
```

because with N=3 points,

```
CoVar(x1,x1) = Var(x1)
             = [(P1(x1)-Mean)² + (P2(x1)-Mean)² + (P3(x1)-Mean)²] / (N-1)
             = [(1-2)² + (2-2)² + (3-2)²] / 2 = 1
CoVar(x2,x2) = Var(x2)
             = [(P1(x2)-Mean)² + (P2(x2)-Mean)² + (P3(x2)-Mean)²] / (N-1)
             = [(1-2)² + (2-2)² + (3-2)²] / 2 = 1
CoVar(x1,x2) = CoVar(x2,x1)
             = [(P1(x1)-Mean)x(P1(x2)-Mean)
                 + (P2(x1)-Mean)x(P2(x2)-Mean)
                   + (P3(x1)-Mean)x(P3(x2)-Mean)] / (N-1)
             = [(1-2)x(1-2) + (2-2)x(2-2) + (3-2)x(3-2)]/2 = 1
```

Eigenvalue computed as, det(CoVar-λI) = 0
that gives, (1-λ)² - 1 = 0; implies, λ = 0, 2

Eigenvector (x1,x2) corresponding to the highest eigenvalue will be the principal component here.

```
Hence, [ 1 1 ] [x1] = 2 [x1]  implies, x1 = x2 = a = 1/√2.
       [ 1 1 ] [x2]     [x2]
```

==