M S RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	31st Jan to 22nd May, 2019		**Course Code:**	ISEB1
Course:	Machine Learning		**Semester:**	6 A, B, C
CIE Details:	Test – I	13-03-2019	9.30 to 10.30am	**Max Marks:** 30

Test Portions: Lesson #01-14 **Instructions to Candidates:** Answer **any two** full Qs. Mobiles are banned.

Sl. #	Question	Marks	Bloom's Level	COs
1a	Apply Naive Bayes Classifier to the following dataset to evaluate your activity for tonight given that you have deadlines looming - but none of them are urgent, there is no party on tonight and you are feeling lazy.	6	L3	CO1
1b	Explain with an example each the difference between **any two** of the foll: **a)** ROC and AUC Curves **b)** Bias and Variance of the ML Model **c)** Variance and Covariance of the ML Model **d)** Overfitting and Underfitting of the ML Model	4	L2	CO1
1c	What is the need of 'Perceptron Activation Function'? What are the types?	5	L4	CO2
2a	A Spam Mail Filter Classification Model was developed to filter out spam where 1000 email samples were examined. The number of TP = 100, FP = 30,TN = 700, FN = 170. Do the following for this Classification Model: **a)** Draw the Confusion Matrix **b)** Calculate the Accuracy value & Error Rate **c)** Assess the Sensitivity & Specificity Value **d)** Determine the Precision & Recall Value **e)** Compute the F-Measure **f)** Is this Model a 'High Precision' or 'High Recall' model? Justify your answer **g)** Evaluate if the data sample given to you a Balanced or an Imbalanced one? Justify your answer	10	L4	CO1
2b	Consider the following set of points: {(-2 , -1) , (1 , 1) , (3 , 2)} **a)** Find the least square regression line for the given data points. **b)** Plot the given points and the regression line in the same rectangular system of axes.	5	L3	CO2
3a	What do you mean by Imbalanced Datasets? Why does Accuracy fail as a performance metrics for these datasets?	4	L4	CO1
3b	What do you understand by the following terms? Answer **any two**. Give relevant examples: **a)** Maximum A Posterior (MAP) Classification Rule **b)** Conditional Probability **c)** Bayesian Rule	6	L2	CO1
3c	Write and describe the Perceptron Algorithm. What is the difference between the 'Recall' and 'Training' phase?	5	L2	CO2

Dataset for 1a:

Deadline?	Is there a party?	Lazy?	Activity
Urgent	Yes	Yes	Party
Urgent	No	Yes	Study
Near	Yes	Yes	Party
None	Yes	No	Party
None	No	Yes	Pub
None	Yes	No	Party
Near	No	No	Study
Near	No	Yes	TV
Near	Yes	Yes	Party
Urgent	No	No	Study

ISEB1 (Machine Learning) T1 - Answer Scheme

1a) Answer as below (6 marks):

- Correct statement of Conditional Probability Equations : **1 Mark**
- Calculation of each Conditional Probability : **4 Marks**
- Writing the final inference : **1 Mark**

- P(Party) × P(Near | Party) × P(No Party | Party) × P(Lazy | Party)
- P(Study) × P(Near | Study) × P(No Party | Study) × P(Lazy | Study)
- P(Pub) × P(Near | Pub) × P(No Party | Pub) × P(Lazy | Pub)
- P(TV) × P(Near | TV) × P(No Party | TV) × P(Lazy | TV)

Using the data above these evaluate to:

$$P(\text{Party}|\text{near (not urgent) deadline, no party, lazy}) = \frac{5}{10} \times \frac{2}{5} \times \frac{0}{5} \times \frac{3}{5} = 0$$

$$P(\text{Study}|\text{near (not urgent) deadline, no party, lazy}) = \frac{3}{10} \times \frac{1}{3} \times \frac{3}{3} \times \frac{1}{3} = \frac{1}{30}$$

$$P(\text{Pub}|\text{near (not urgent) deadline, no party, lazy}) = \frac{1}{10} \times \frac{0}{1} \times \frac{1}{1} \times \frac{1}{1} = 0$$

$$P(\text{TV}|\text{near (not urgent) deadline, no party, lazy}) = \frac{1}{10} \times \frac{1}{1} \times \frac{1}{1} \times \frac{1}{1} = \frac{1}{10}$$

So based on this you will be watching TV tonight.

1b) Answer any two (2 x 2 = 4 Marks : Any two valid points of difference)

a) ROC and AUC Curves
b) Bias and Variance of the ML Model
c) Variance and Covariance of the ML Model
d) Overfitting and Underfitting of the ML Model

2a) Total: 10 Marks

a) **1 mark**

Confusion Matrix

		Folder	
		Spam Folder	Inbox
E-mail	Spam	100 True positives	170 False Negatives
	Not spam	30 False Positives	700 True Negatives

b) 2 x 1 = 2 Marks

$$\text{Accuracy} = \frac{100 + 700}{1000} = 80\%$$

Error Rate = 1 - Accuracy = 0.2 or 20%

c) 2 x 1 = 2 Marks

$$\text{Sensitivity} = \frac{\#TP}{\#TP + \#FN}$$

$$\text{Specificity} = \frac{\#TN}{\#TN + \#FP}$$

$$\text{Precision} = \frac{\#TP}{\#TP + \#FP}$$

$$\text{Recall} = \frac{\#TP}{\#TP + \#FN}$$

- Sensitivity = 100 / (100+170) = 0.3703
- Specificity = 700/(30 + 700) = 0.9589

d)2 x 1 = 2 Marks

2b) Total : 5 Marks

Let us organize the data in a table.

x	y	x y	x^2
-2	-1	2	4
1	1	1	1
3	2	6	9
$\Sigma x = 2$	$\Sigma y = 2$	$\Sigma xy = 9$	$\Sigma x^2 = 14$

- Writing the above values correctly: **1 Mark**
- Calculating the least square regression: **2 Marks**
- Plotting the graph : **1 Mark**

$$\text{Precision} = \frac{100}{100 + 30} = 76.9\%$$

$$\text{Recall} = \frac{100}{100 + 170} = 37\%$$

e) 1 mark

$$\text{F1 Score} = \frac{2 \times 76.9 \times 37}{76.9 + 37} = 49.96\%$$

f) 1 mark

High Precision

Justification: FPR must be as minimum as possible for the Spam Mail Filter system given

g) 1 mark

Imbalanced

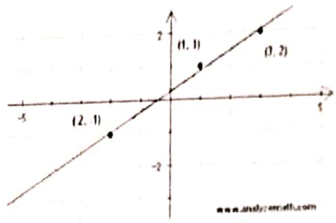
Justification: The number of True Positives are very less (Only 100: 10% of the samples) compared to the number of True Negatives(700 : 70% of the samples)

We now use the above formula to calculate a and b as follows

$a = (n\Sigma x\,y - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (3{*}9 - 2{*}2) / (3{*}14 - 2^2) = 23/38$

$b = (1/n)(\Sigma y - a\ \Sigma x) = (1/3)(2 - (23/38){*}2) = 5/19$

b) We now graph the regression line given by y = a x + b and the given points.



3a)

When TPR / TNR is very low because the number of TP samples are very less, the dataset is imbalanced.

Accuracy fails as a performance metrics because it tends to give a value from the majority sample (TN) which is very high and will fail to detect the minority sample(TP).

3b)

Answering any two below: 3 Marks

a) Maximum A Posterior (MAP) Classification Rule

b) Conditional Probability

c) Bayesian Rule

Explanation : **1 Mark**

One example :**2 Marks**

3c)

The Perceptron Algorithm

- Initialisation
  - set all of the weights w_{ij} to small (positive and negative) random numbers
- Training
  - for T iterations or until all the outputs are correct:
    - for each input vector:
      - compute the activation of each neuron j using activation function g:

$$y_j = g\left(\sum_{i=0}^{m} w_{ij}x_i\right) = \begin{cases} 1 & \text{if } \sum_{i=0}^{m} w_{ij}x_i > 0 \\ 0 & \text{if } \sum_{i=0}^{m} w_{ij}x_i \leq 0 \end{cases} \quad (3.4)$$

      - update each of the weights individually using:

$$w_{ij} \leftarrow w_{ij} - \eta(y_j - t_j) \cdot x_i \quad (3.5)$$

- Recall
  - compute the activation of each neuron j using:

$$y_j = g\left(\sum_{i=0}^{m} w_{ij}x_i\right) = \begin{cases} 1 & \text{if } w_{ij}x_i > 0 \\ 0 & \text{if } w_{ij}x_i \leq 0 \end{cases} \quad (3.6)$$

M S RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute, affiliated to VTU)

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

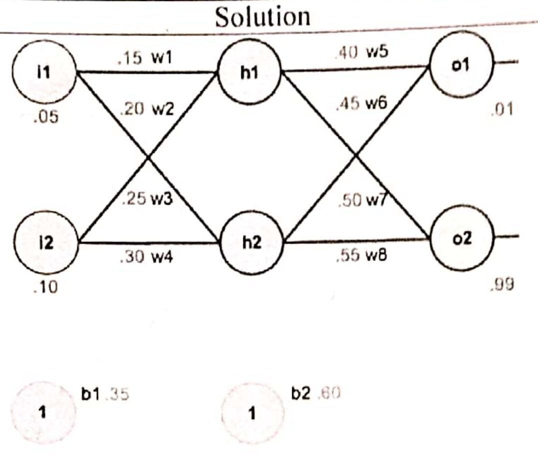
Term:	31st Jan to 22nd May, 2019		**Course Code:**	ISE81
Course:	Machine Learning		**Semester:**	6 A, B, C
CIE Details:	Test – II	10-04-2019	9.30 to 10.30am	**Max Marks:** 30

Test Portions: Lesson #14-36 **Instructions to Candidates:** Answer **any two** full Qs. Mobiles are banned.

Sl. #	Question	Marks	Bloom's Level	CO
1a	Apply MLP to train the neural network to output 0.01 and 0.99 for given inputs 0.05 and 0.10 with bias values 1 by using weights as w1=0.15, w2=0.20, w3=0.25, w4=0.30, b1=0.35, w5=0.40, w6=0.45, w7=0.50, w8=0.55 and b2=0.60.	7	L3	CO2
1b	Show example transaction data where for the rule $X \rightarrow Y$: (a) Both support and confidence are high. (b) Support is high and confidence is low. (c) Support is low and confidence is high. (d) Both support and confidence are low.	4	L3	CO3
1c	In a two-class, two-action problem, if the loss function is $\lambda11 = \lambda22 = 0$, $\lambda12 = 10$, and $\lambda21 = 1$, write the optimal decision rule.	4	L3	CO4
2a	What are the limitations of Single Layer Perceptrons? Why do we need MLP? Explain with an example how multilayer perceptron handles nonlinearity.	7	L3	CO2
2b	In PCA, obtain the **Eigenvalues** for the feature vector and covariance matrix given below. Which 'Eigenvalue' will you consider for PC1 and why? **Original Feature Vector**: $\begin{pmatrix} X1 & X2 \\ 2 & 4 \\ 1 & 3 \\ 0 & 1 \\ -1 & 0.5 \end{pmatrix}$ **Covariance Matrix**: $\begin{pmatrix} 1.667 & 2.08 \\ 2.08 & 2.727 \end{pmatrix}$	4	L3	CO3
2c	Compare the approaches of ID3 and CART	4	L4	CO4
3a	Write the Backpropagation Algorithm in MLP. Explain how MLP overcomes the limitations of single layered perceptron.	5	L3	CO2
3b	Answer **any two** of the following questions: **(i)** In LDA what is Fischer Ratio? What is its significance? **(ii)** Two differences between PCA and ICA **(iii)** Probably Approximately Correct (PAC) learning	6	L2	CO3
3c	Write a brief note on Classification and Regression Trees.	4	L2	CO4

IA test 2

Machine Learning (ISEB1) - Scheme

Q. no	Solution	Marks
1a	 **The Forward Pass** i. Evaluate activation function (one sample for each layer) ii. Calculate the total error for the sample considered in step i **The Backward Pass** **i.** Update the weights at output layer **ii.** Update the weights at hidden layer	2 2 3
1b	Example: (a) A rule that has high support and high confidence. Answer: Milk --> Bread. Such obvious rule tends to be uninteresting. (b) A rule that has reasonably high support but low confidence. Answer: Milk --> Tuna. While the sale of tuna and milk may be higher than the support threshold, not all transactions that contain milk also contain tuna. Such low-confidence rule tends to be uninteresting. (c) A rule that has low support and low confidence. Answer: Cooking oil --> Laundry detergent. Such low confidence rule tends to be uninteresting. (d) A rule that has low support and high confidence. Answer: Vodka --> Caviar. Such rule tends to be interesting	4x1 = 4
1c	*In a two-class, two-action problem, if the loss function is $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = 10$, and $\lambda_{21} = 1$, write the optimal decision rule.* We calculate the expected risks of the two actions: $R(\alpha_1 \mid x) = \lambda_{11}P(C_1 \mid x) + \lambda_{12}P(C_2 \mid x) = 10P(C_2 \mid x)$ $R(\alpha_2 \mid x) = \lambda_{21}P(C_1 \mid x) + \lambda_{22}P(C_2 \mid x) = P(C_1 \mid x)$ and we choose C_1 if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$, or if $P(C_1 \mid x) > 10P(C_2 \mid x)$, $P(C_1 \mid x) > 10/11$. Assigning accidentally an instance of C_2 to C_1 is so bad that we need to be very sure before assigning an instance to C_1.	1 1 1 1
2a	Limitations of Single layer perceptron Need for MLP Non-linearity handling with an example	2 2 3
2b	Dimensionality reduction from 2D to 1D Explanation for selecting PC1	2 2
2c	Minimum of 2 comparison between ID3 and CART	2x2=4

3a MLP with backpropagation — 4

- **Initialisation**
  - initialise all weights to small (positive and negative) random values
- **Training**
  - repeat:
    - for each input vector:

Forwards phase:

- compute the activation of each neuron j in the hidden layer(s) using:

$$h_\zeta = \sum_{i=0}^{L} x_i v_{i\zeta}$$

$$a_\zeta = g(h_\zeta) = \frac{1}{1+\exp(-\beta h_\zeta)}$$

- work through the network until you get to the output layer neurons which have activations (although see also Section 4.2.3):

$$h_\kappa = \sum_j a_j w_{j\kappa}$$

$$y_\kappa = g(h_\kappa) = \frac{1}{1+\exp(-\beta h_\kappa)}$$

Backwards phase:

- compute the error at the output using:

$$\delta_o(\kappa) = (y_\kappa - t_\kappa) y_\kappa (1 - y_\kappa)$$

- compute the error in the hidden layer(s) using:

$$\delta_h(\zeta) = a_\zeta(1 - a_\zeta) \sum_{k=1}^{N} w_\zeta \delta_o(k)$$

- update the output layer weights using:

$$w_{\zeta\kappa} \leftarrow w_{\zeta\kappa} - \eta \delta_o(\kappa) a_\zeta^{\text{hidden}}$$

- update the hidden layer weights using:

$$v_\iota \leftarrow v_\iota - \eta \delta_h(\kappa) x_\iota$$

(if using sequential updating) randomise the order of the input vectors so that you don't train in exactly the same order each iteration

- **Recall**
  - use the Forwards phase in the training section above

MLP overcomes the limitations of single layer perceptron — 1

3b

i. Fisher ratio: — 1

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

Explain the significance — 2

ii. Minimum of 2 differences between LDA and PCA with example for each — 2+1

iii. Explain PAC learning algorithm steps — 3

3c Explain CART with an example — 4

M S RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	31st Jan to 22nd May, 2019			**Course Code:**	ISE81
Course:	Machine Learning			**Semester:**	6 A, B, C
CIE Details:	Test – III	11-05-2019	9.30 to 10.30am	**Max Marks:**	30

Test Portions: Lesson #37-56**Instructions to Candidates:** Answer any two full Qs. Mobiles are banned.

Sl. #	Question	Marks	Bloom's Level	CO
1a	Cluster the dataset = {2,3,4,10,11,12, 20, 25,30} using k-means algorithm. We need to group into two clusters. Assume the initial centroids as 2 and 12.	5	L3	CO5
1b	Explain Reinforcement learning cycle with suitable example.	5	L2	CO5
1c	Explain how extensions of SVM is applied to any one of the below scenarios: (i) Multi-class classification (ii) Regression	5	L2	CO4
2a	Write and explain the SOM Algorithm. Explain with proper example why does it fall under the category of *'competitive learning'* algorithms?	5	L2	CO5
2b	suppose that the following are the given positively labeled data points in □2: $\left\{\begin{pmatrix}3\\1\end{pmatrix},\begin{pmatrix}3\\-1\end{pmatrix},\begin{pmatrix}6\\1\end{pmatrix},\begin{pmatrix}6\\-1\end{pmatrix}\right\}$ and the following negatively labeled data points in □2: $\left\{\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}-1\\0\end{pmatrix}\right\}$ Obtain a separating hyperplane that accurately discriminates the two classes using SVM.	6	L3	CO4
2c	Compare K-means and Hierarchical clustering	4	L4	CO4
3a	Explain the following with appropriate examples: (i) Criteria for choosing the Number of Clusters (ii) Mixture Densities	5	L2	CO5
3b	With any example of your choice illustrate the use of k-Nearest Neighbour as a probabilistic learning algorithm.	5	L2	CO4
3c	Explain The Gaussian Mixture Model EM Algorithm steps and list few application of EM	5	L2	CO4

SOLUTION

1a. Cluster the dataset = {2,3,4,10,11,12, 20, 25,30} using k-means algorithm. We need to group into two clusters. Assume the initial centroids as 2 and 12.

	Centroid1 = 2	Centroid2 = 12
Iteration 1	{2,3,4}	{10,11,12,20,25,30}
	Centroid1 = 3	Centroid2 = 18
Iteration 2	{2,3,4,10}	{11,12,20,25,30}
	Centroid1 = 4.75~5	Centroid2 = 19.6~20
Iteration 3	{2,3,4,10,11,12}	{20,25,30}
	Centroid1 = 7	Centroid2 = 25
Iteration 4	{2,3,4,10,11,12}	{20,25,30}
	Centroid1 = 7	Centroid2 = 25
Iteration 5	{2,3,4,10,11,12}	{20,25,30}

Clusters in Iteration 4 = Clusters in Iteration 5 and centroids for these iteration converges. (5X1=5)

1b. Explain Reinforcement learning cycle with suitable example.

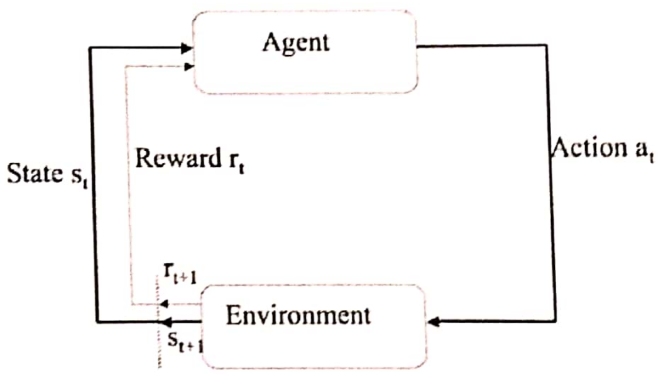


Figure explanation with example (2+3)

1c. Explain how extensions of SVM is applied to any one of the below scenarios:

(i) Multi-class classification (2.5)

(ii) Regression (2.5)

2a.Write and explain the SOM Algorithm. (3)

- **Initialisation**
  - choose a size (number of neurons) and number of dimensions d for the map
  - either:
    - choose random values for the weight vectors so that they are all different OR
    - set the weight values to increase in the direction of the first d principal components of the dataset

- **Learning**
  - repeat:
    - * for each datapoint:
      - select the best-matching neuron n_b using the minimum Euclidean distance between the weights and the input,

$$n_b = \min_j \|\mathrm{x} - \mathrm{w}_j^T\|. \tag{14.8}$$

    - * update the weight vector of the best-matching node using:

$$\mathrm{w}_j^T \leftarrow \mathrm{w}_j^T + \eta(t)(\mathrm{x} - \mathrm{w}_j^T), \tag{14.9}$$

where $\eta(t)$ is the learning rate.

* update the weight vector of all other neurons using:

$$\mathrm{w}_j^T \leftarrow \mathrm{w}_j^T + \eta_n(t)h(n_b, t)(\mathrm{x} - \mathrm{w}_j^T), \tag{14.10}$$

Explain with proper example why does it fall under the category of *'competitive learning'* algorithms?

(2)

2b. Obtain a separating hyperplane that accurately discriminates the two classes using SVM.

$$\left\{\begin{pmatrix}3\\1\end{pmatrix}, \begin{pmatrix}3\\-1\end{pmatrix}, \begin{pmatrix}6\\1\end{pmatrix}, \begin{pmatrix}6\\-1\end{pmatrix}\right\} \left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix}\right\}$$

(1+1+2+1+1)

Soln:

$$\left\{s_1 = \begin{pmatrix}1\\0\end{pmatrix}, s_2 = \begin{pmatrix}3\\1\end{pmatrix}, s_3 = \begin{pmatrix}3\\-1\end{pmatrix}\right\}$$

SV hyperplane representation

$$\begin{aligned}
\alpha_1\Phi(s_1)\cdot\Phi(s_1) + \alpha_2\Phi(s_2)\cdot\Phi(s_1) + \alpha_3\Phi(s_3)\cdot\Phi(s_1) &= -1\\
\alpha_1\Phi(s_1)\cdot\Phi(s_2) + \alpha_2\Phi(s_2)\cdot\Phi(s_2) + \alpha_3\Phi(s_3)\cdot\Phi(s_2) &= +1\\
\alpha_1\Phi(s_1)\cdot\Phi(s_3) + \alpha_2\Phi(s_2)\cdot\Phi(s_3) + \alpha_3\Phi(s_3)\cdot\Phi(s_3) &= +1
\end{aligned}$$

Substitute values for s1, s2 and s3 and solve for α

$$\begin{aligned}
2\alpha_1 + 4\alpha_2 + 4\alpha_3 &= -1\\
4\alpha_1 + 11\alpha_2 + 9\alpha_3 &= +1\\
4\alpha_1 + 9\alpha_2 + 11\alpha_3 &= +1
\end{aligned}$$

$\alpha_1 = 3{:}5$; $\alpha_2 = 0{:}75$ and $\alpha_3 = 0{:}75$

$y = wx + b$ with $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $b = -2$.

2c. Compare K-means and Hierarchical clustering (minimum 2 differences)

3a. Explain the following with appropriate examples: (2x2.5 = 5)

(i) Criteria for choosing the Number of Clusters

(ii) Mixture Densities

3b. With any example of your choice illustrate the use of k-Nearest Neighbour as a probabilistic learning algorithm.

Explain working of algorithm with example (3+2=5)

3c. Explain The Gaussian Mixture Model EM Algorithm steps and list few application of EM(4+1=5)

GMM EM algorithm steps

- Initialisation
  - set $\hat{\mu}_1$ and $\hat{\mu}_2$ to be randomly chosen values from the dataset
  - set $\hat{\sigma}_1 = \hat{\sigma}_2 = \sum_{i=1}^{N} (y_i - \bar{y})^2 / N$ (where $\bar{y}$ is the mean of the entire dataset)
  - set $\hat{\pi} = 0.5$
- Repeat until convergence:
  - (E-step) $\hat{\gamma}_i = \frac{\hat{\pi}\phi(y_i;\hat{\mu}_1,\hat{\sigma}_1)}{\hat{\pi}\phi(y_i;\hat{\mu}_1,\hat{\sigma}_1)+(1-\hat{\pi})\phi(y_i;\hat{\mu}_2,\hat{\sigma}_2)}$ for $i = 1 \dots N$
  - (M-step 1) $\hat{\mu}_1 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i)y_i}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)}$
  - (M-step 2) $\hat{\mu}_2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i y_i}{\sum_{i=1}^{N}\hat{\gamma}_i}$
  - (M-step 3) $\hat{\sigma}_1 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i)(y_i-\hat{\mu}_1)^2}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)}$
  - (M-step 4) $\hat{\sigma}_2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i(y_i-\hat{\mu}_2)^2}{\sum_{i=1}^{N}\hat{\gamma}_i}$
  - (M-step 5) $\hat{\pi} = \sum_{i=1}^{N} \frac{\hat{\gamma}_i}{N}$

List any 2 applications of EM