

NPTEL Live Session

Week 1

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

30/01/24

Live Sessions

- Start Date: 30th January, 2024
- When: Every Tuesday
- Time: 06:00 PM - 08:00 PM
- Link to join: <https://meet.google.com/nrz-gtib-ncf>

Question 1

Which of the following are supervised learning problems? (multiple may be correct)

- a. Learning to ride a bicycle using a reward signal.
- b. Predicting disease from blood sample.
- c. Grouping students in the same class based on similar features.
- d. Face recognition to unlock your phone.

Question 1

Which of the following are supervised learning problems? (multiple may be correct)

- a. Learning to drive using a reward signal.
- b. Predicting disease from blood sample.
- c. Grouping students in the same class based on similar features.
- d. Face recognition to unlock your phone.

Answer : b,d

Question 2

Which of the following are classification problems? (multiple may be correct)

- a. Predict the price of a house.
- b. Predict the disease of a person.
- c. Predict the temperature for tomorrow.
- d. Learn to play a guitar.

Question 2

Which of the following are classification problems? (multiple may be correct)

- a. Predict the price of a house.
- b. **Predict the disease of a patient.**
- c. Predict the temperature for tomorrow.
- d. Learn to play a guitar.

Answer : b

Question 3

Which of the following is not a categorical feature? (multiple may be correct)

- A. Area of a House.
- B. Height of a person
- C. Types of Mountains
- D. Nationality of a person

Question 3

Which of the following is not a categorical feature? (multiple may be correct)

- A. Area of a House**
- B. Height of a person**
- C. Types of Mountains
- D. Nationality of a person

Answer : A, B

Detailed Solution : Categorical variables represent types of data which may be divided into groups. All other features are continuous.

Question 4

Which of the following tasks is NOT a suitable machine learning task?

- A. Finding the shortest path between a pair of nodes in a graph
- B. Predicting if a stock price will rise or fall
- C. Predicting the price of petroleum
- D. Grouping mails as spams or non-spams

Question 4

Which of the following tasks is NOT a suitable machine learning task?

- A. **Finding the shortest path between a pair of nodes in a graph**
- B. Predicting if a stock price will rise or fall
- C. Predicting the price of petroleum
- D. Grouping mails as spams or non-spams

Answer : A. Finding the shortest path between a pair of nodes in a graph

Detailed Solution : Finding the shortest path is a graph theory based task, whereas other options are completely suitable for machine learning.

Question 5

What is the use of Validation dataset in Machine Learning?

- A. To train the machine learning model.
- B. To evaluate the performance of the machine learning model
- C. To tune the hyperparameters of the machine learning model
- D. None of the above.

Question 5

What is the use of Validation dataset in Machine Learning?

- A. To train the machine learning model.
- B. To evaluate the performance of the machine learning model
- C. To tune the hyperparameters of the machine learning model**
- D. None of the above.

Answer : C. To tune the hyperparameters of the machine learning model

Detailed Solution : The validation dataset provides an unbiased evaluation of a model fit on the training dataset while tuning the model's hyperparameters.

Question 6

Which of the following is an unsupervised task?

- (a) Predicting the price of a house based on its area and location.
- (b) Grouping customers based on their preference and history.
- (c) Grouping of handwritten digits from their images.
- (d) Predicting the time (in days) a PhD student will take to complete his/her thesis
- (e) all of the above

Question 6

Which of the following is an unsupervised task?

- (a) Predicting the price of a house based on its area and location.
- (b) Grouping customers based on their preference and history.
- (c) Grouping of handwritten digits from their images.
- (d) Predicting the time (in days) a PhD student will take to complete his/her thesis.
- (e) all of the above

Answer: (b), (c)

Question 7

Let X and Y be uniformly distributed random variables over the interval [0,6] and [2,8] respectively. If X and Y are independent events, then compute the probability, $P(\min(X,Y) < 4)$

- a. 1/9
- b. 2/9
- c. 5/9
- d. 7/9

Question 7

Let X and Y be uniformly distributed random variables over the interval [0,6] and [2,8] respectively. If X and Y are independent events, then compute the probability, $\mathbf{P}(\min(X,Y) < 4)$

- a. 1/9
- b. 2/9
- c. 5/9
- d. 7/9

Answer: d. 7/9

Question 8

Which of the following is an unsupervised learning task? (multiple options may be correct)

- a. Predict a stock market price of a company.
- b. Predict rainfall for a given day.
- c. Group flowers based on similar features.
- d. Predict whether an image contains a Dog or a Cat

Question 8

Which of the following is an unsupervised learning task? (multiple options may be correct)

- a. Predict a stock market price of a company.
- b. Predict rainfall for a given day.
- c. **Group flowers based on similar features.**
- d. Predict whether an image contains a Dog or a Cat

Answer : C

Grouping of Objects is an Unsupervised Learning Task.

Question 9

Which of the following is a reinforcement learning task? (multiple options may be correct)

- A. Learning to swim.
- B. Learning to predict weather.
- C. Learning to play a video game.
- D. Learning to group similar flowers based on features.

Question 9

Which of the following is a reinforcement learning task? (multiple options may be correct)

- A. **Learning to swim.**
- B. Learning to predict weather.
- C. **Learning to play a video game.**
- D. Learning to group similar flowers based on features.

Answer : A,C

These involve learning from trial and error.

Question 10

A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is green?

- a. $2/3$
- b. $10/21$
- c. $2/7$
- d. $4/7$

Question 10

A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is green?

- A. $\frac{2}{3}$
- B. $\frac{10}{21}$
- C. $\frac{2}{7}$
- D. $\frac{4}{7}$

Answer : C. $\frac{2}{7}$

No of ways of drawing 2 balls = 7C_2

No of ways of drawing 2 non-green balls = 4C_2

Probability of drawing 2 non-green balls = ${}^4C_2 / {}^7C_2$

Question 11

Which of the following are false about bias and variance of overfitted and underfitted models?
(multiple options may be correct)

- (a) Underfitted models have high bias.
- (b) Underfitted models have low bias.
- (c) Overfitted models have low variance.
- (d) Overfitted models have high variance.

Question 11

Which of the following are false about bias and variance of overfitted and underfitted models?
(multiple options may be correct)

- (a) Underfitted models have high bias.
- (b) Underfitted models have low bias.
- (c) Overfitted models have low variance.
- (d) Overfitted models have high variance.

Sol. (b), (c)

Question 12

An airplane knows that 5 percent of the people making reservations on a certain flight will not show up. Consequently, their policy is to sell 52 tickets for a flight that can hold only 50 passengers. What is the probability that there will be a seat available for every passenger who shows up?

- a. 0.51
- b. 0.81
- c. 0.63
- d. 0.74

Question 12

An airplane knows that 5 percent of the people making reservations on a certain flight will not show up. Consequently, their policy is to sell 52 tickets for a flight that can hold only 50 passengers. What is the probability that there will be a seat available for every passenger who shows up?

- a. 0.51
- b. 0.81
- c. 0.63
- d. **0.74**

Answer: d. 0.74

THE END

NPTEL Live Session

Week 2

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

06/02/2024

Question 1

Given a training data set of 10,000 instances, with each input instance having 17 dimensions and each output instance having 2 dimensions, the dimensions of the design matrix used in applying linear regression to this data is

- (a) 10000×17
- (b) 10002×17
- (c) 10000×18
- (d) 10000×19

Question 1

Given a training data set of 10,000 instances, with each input instance having 17 dimensions and each output instance having 2 dimensions, the dimensions of the design matrix used in applying linear regression to this data is

- (a) 10000×17
- (b) 10002×17
- (c) 10000×18
- (d) 10000×19

Sol. (c)

Question 2

Which of the following is false about PCA?

- A. PCA is a supervised method
- B. It identifies the directions that data have the largest variance
- C. Maximum number of principal components \leq number of features
- D. All principal components are orthogonal to each other

Question 2

Which of the following is false about PCA?

- A. **PCA is a supervised method**
- B. It identifies the directions that data have the largest variance
- C. Maximum number of principal components \leq number of features
- D. All principal components are orthogonal to each other

Correct Answer : A. PCA is a supervised method

Detailed Solution : PCA is an unsupervised learning algorithm.

Question 3

Given below is your dataset. You are using KNN regression with K=3. What is the prediction for a new input value (3, 2)?

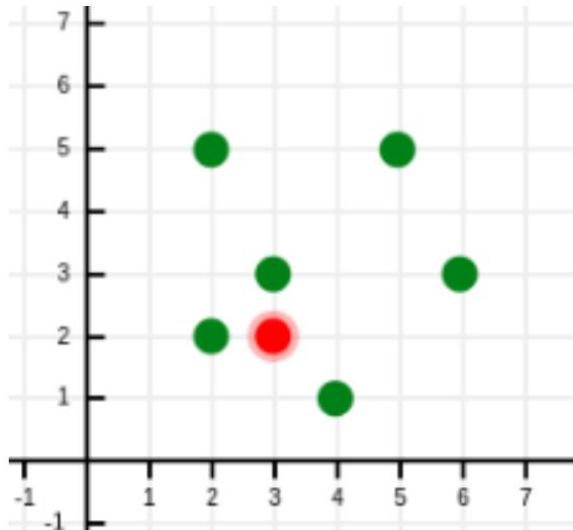
X1	X2	Y
2	5	3.5
5	5	5
3	3	3
6	3	4.5
2	2	2
4	1	2.5

Use Euclidean distance measure for finding closest points.

- (a) 2
- (b) 3
- (c) 3.5
- (d) 2.5

Question 3

Answer : d



$$\text{Closest points} = (2,2), (3,3), (4,1)$$
$$\text{Average over } Y = \frac{2+3+2.5}{3} = 2.5$$

Question 4

Find the mean of squared error for the given predictions:

Y	f(x)
1	2
2	3
4	5
8	9
16	15
32	31

- (a) 1
- (b) 2
- (c) 1.5
- (d) 0

Question 4

Find the mean of squared error for the given predictions:

Y	f(x)
1	2
2	3
4	5
8	9
16	15
32	31

- (a) 1
- (b) 2
- (c) 1.5
- (d) 0

Answer: a

$$\text{Mean Squared Error} = \sum_{i=1}^N \frac{(Y_i - f(X_i))^2}{N}$$

Question 5

Suppose we want to add a regularizer to the linear regression loss function, to control the magnitudes of the weights β . We have a choice between $\Omega_1(\beta) = \sum_{i=1}^p |\beta_i|$ and $\Omega_2(\beta) = \sum_{i=1}^p \beta_i^2$. Which one is more likely to result in sparse weights?

- (a) Ω_1
- (b) Ω_2
- (c) Both Ω_1 and Ω_2 will result in sparse weights
- (d) Neither of Ω_1 or Ω_2 can result in sparse weights

Question 5

Suppose we want to add a regularizer to the linear regression loss function, to control the magnitudes of the weights β . We have a choice between $\Omega_1(\beta) = \sum_{i=1}^p |\beta_i|$ and $\Omega_2(\beta) = \sum_{i=1}^p \beta_i^2$. Which one is more likely to result in sparse weights?

- (a) Ω_1
- (b) Ω_2
- (c) Both Ω_1 and Ω_2 will result in sparse weights
- (d) Neither of Ω_1 or Ω_2 can result in sparse weights

Sol. (a)

Question 6

The model obtained by applying linear regression on the identified subset of features may differ from the model obtained at the end of the process of identifying the subset during

- (a) Forward stepwise selection
- (b) Backward stepwise selection
- (c) Forward stagewise selection
- (d) All of the above

Question 6

The model obtained by applying linear regression on the identified subset of features may differ from the model obtained at the end of the process of identifying the subset during

- (a) Forward stepwise selection
- (b) Backward stepwise selection
- (c) Forward stagewise selection
- (d) All of the above

Sol. (c)

Question 7

Consider forward selection, backward selection and best subset selection with respect to the same data set. Which of the following is true?

- (a) Best subset selection can be computationally more expensive than forward selection
- (b) Forward selection and backward selection always lead to the same result
- (c) Best subset selection can be computationally less expensive than backward selection
- (d) Best subset selection and forward selection are computationally equally expensive
- (e) Both (b) and (d)

Question 7

Consider forward selection, backward selection and best subset selection with respect to the same data set. Which of the following is true?

- (a) Best subset selection can be computationally more expensive than forward selection
- (b) Forward selection and backward selection always lead to the same result
- (c) Best subset selection can be computationally less expensive than backward selection
- (d) Best subset selection and forward selection are computationally equally expensive
- (e) Both (b) and (d)

Answer: a

Question 8

Principal Component Regression (PCR) is an approach to find an orthogonal set of basis vectors which can then be used to reduce the dimension of the input. Which of the following matrices contains the principal component directions as its columns (follow notation from the lecture video)

- (a) X
- (b) S
- (c) X_c
- (d) V
- (e) U

Question 8

Principal Component Regression (PCR) is an approach to find an orthogonal set of basis vectors which can then be used to reduce the dimension of the input. Which of the following matrices contains the principal component directions as its columns (follow notation from the lecture video)

- (a) X
- (b) S
- (c) X_c
- (d) V
- (e) U

Sol. (d)

Question 9

Let $A^{m \times n}$ be a matrix of real numbers. The matrix AA^T has an eigenvector x with eigenvalue b . Then the eigenvector y of A^TA which has eigenvalue b is equal to

- (a) $x^T A$
- (b) $A^T x$
- (c) x
- (d) Cannot be described in terms of x

Question 9

Let $A^{m \times n}$ be a matrix of real numbers. The matrix AA^T has an eigenvector x with eigenvalue b . Then the eigenvector y of A^TA which has eigenvalue b is equal to

- (a) $x^T A$
- (b) $A^T x$
- (c) x
- (d) Cannot be described in terms of x

Sol. (b)

$$(AA^T)x = bx$$

Multiplying by A^T on both sides and rearranging,

$$\begin{aligned}(A^T)(AA^T)x &= A^T(bx) \\ (A^TA)(A^T x) &= b(A^T x)\end{aligned}$$

Hence, $A^T x$ is an eigenvector of A^TA , with eigenvalue b .

Question 10

Let u be a $n \times 1$ vector, such that $u^T u = 1$. Let I be the $n \times n$ identity matrix. The $n \times n$ matrix A is given by $(I - kuu^T)$, where k is a real constant. u itself is an eigenvector of A , with eigenvalue -1 . What is the value of k ?

- (a) -2
- (b) -1
- (c) 2
- (d) 0

Question 10

Let u be a $n \times 1$ vector, such that $u^T u = 1$. Let I be the $n \times n$ identity matrix. The $n \times n$ matrix A is given by $(I - kuu^T)$, where k is a real constant. u itself is an eigenvector of A , with eigenvalue -1 . What is the value of k ?

- (a) -2
- (b) -1
- (c) 2
- (d) 0

Sol. (c)

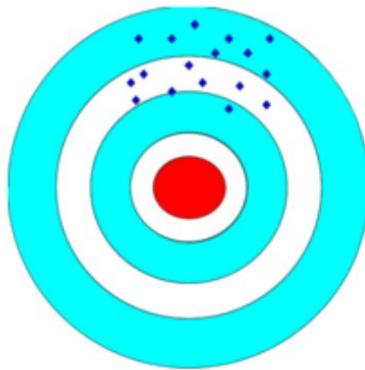
$$\begin{aligned}(I - kuu^T)u &= -u \\ u - ku(u^T u) &= -u \\ 2u - ku &= 0\end{aligned}$$

Hence, $k = 2$

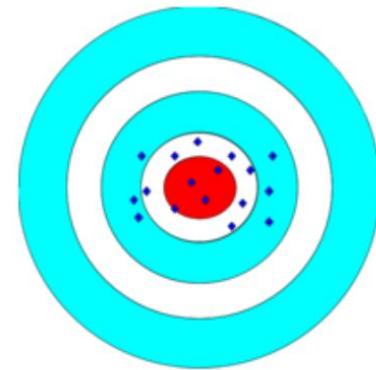
Question 11

Bias and Variance can be visualized using a classic example of a dart game. We can think of the true value of the parameters as the bull's-eye on a target, and the arrow's value as the estimated value from each sample. Consider the following situations, and select the correct option(s)

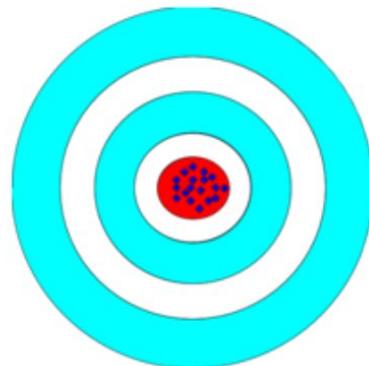
- (a) Player 1 has low variance compared to player 4
- (b) Player 1 has higher variance compared to player 4
- (c) Bias exhibited by player 1 is lesser than that done by player 3.



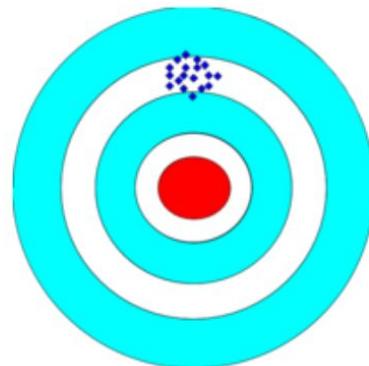
Board of player 1



Board of player 2



Board of player 3



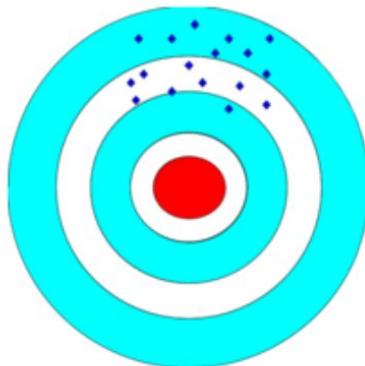
Board of player 4

Question 11

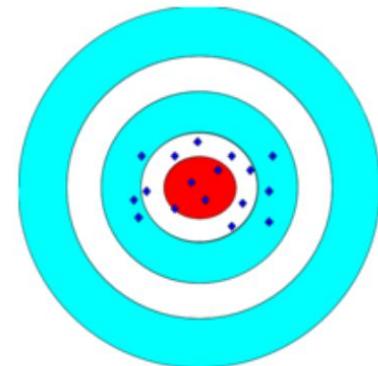
Bias and Variance can be visualized using a classic example of a dart game. We can think of the true value of the parameters as the bull's-eye on a target, and the arrow's value as the estimated value from each sample. Consider the following situations, and select the correct option(s)

- (a) Player 1 has low variance compared to player 4
- (b) Player 1 has higher variance compared to player 4
- (c) Bias exhibited by player 1 is lesser than that done by player 3.

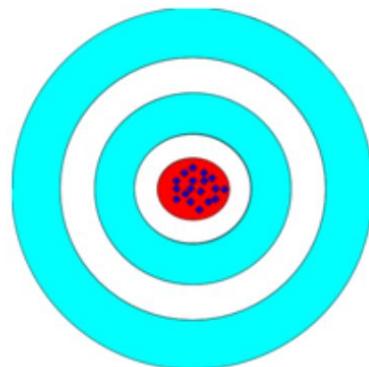
Answer: (b)



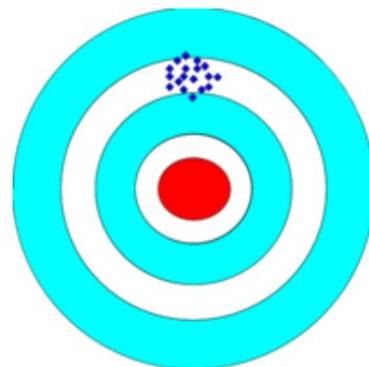
Board of player 1



Board of player 2



Board of player 3



Board of player 4

Question 12

Which of the following are true for any $m \times n$ matrix A of real numbers.

- (a) The rowspace of A is the same as the columnspace of A^T
- (b) The rowspace of A is the same as the rowspace of A^T
- (c) The eigenvectors of AA^T are the same as the eigenvectors of A^TA
- (d) The eigenvalues of AA^T are the same as the eigenvalues of A^TA

Question 12

Which of the following are true for any $m \times n$ matrix A of real numbers.

- (a) The rowspace of A is the same as the columnspace of A^T
- (b) The rowspace of A is the same as the rowspace of A^T
- (c) The eigenvectors of AA^T are the same as the eigenvectors of A^TA
- (d) The eigenvalues of AA^T are the same as the eigenvalues of A^TA

Sol. (a), (d)

Since the rows of A are the same as the columns of A^T , the rowspace of A is the same as the columnspace of A^T . The eigenvalues of AA^T are the same as the eigenvalues of A^TA , because if $AA^Tx = \lambda x$ we get $A^TA(A^Tx) = \lambda(A^Tx)$. (b) is clearly not necessary. (c) need not hold, since although the eigenvalues are same, the eigenvectors have a factor of A^T multiplying them.

THE END

NPTEL Live Session

Week 3

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

13/02/2024

Question 1

Which of the following is false about a logistic regression based classifier?

- (a) The logistic function is non-linear in the weights
- (b) The logistic function is linear in the weights
- (c) The decision boundary is non-linear in the weights
- (d) The decision boundary is linear in the weights

Question 1

Which of the following is false about a logistic regression based classifier?

- (a) The logistic function is non-linear in the weights
- (b) The logistic function is linear in the weights
- (c) The decision boundary is non-linear in the weights
- (d) The decision boundary is linear in the weights

Sol. (b), (c)

Question 2

Consider the case where two classes follow Gaussian distribution which are centered at (3, 9) and (-3, 3) and have identity covariance matrix. Which of the following is the separating decision boundary using LDA assuming the priors to be equal?

- (a) $y-x=3$
- (b) $x+y=3$
- (c) $x+y=6$
- (d) both (b) and (c)
- (e) None of the above
- (f) Can not be found from the given information

Question 2

Consider the case where two classes follow Gaussian distribution which are centered at (3, 9) and (-3, 3) and have identity covariance matrix. Which of the following is the separating decision boundary using LDA assuming the priors to be equal?

- (a) $y-x=3$
- (b) $x+y=3$
- (c) $x+y=6$
- (d) both (b) and (c)
- (e) None of the above
- (f) Can not be found from the given information

Sol. (c)

As the distribution is Gaussian and have identity covariance (which are equal), the separating boundary will be linear. The decision boundary will be orthogonal to the line joining the centers and will pass from the midpoint of centers.

Question 3

Logit transformation for $Pr(X = 1)$ for given data is

$$S = [0, 1, 1, 0, 1, 0, 1]$$

- (a) $\frac{3}{4}$
- (b) $\frac{4}{3}$
- (c) $\frac{4}{7}$
- (d) $\frac{3}{7}$

Question 3

Logit transformation for $Pr(X = 1)$ for given data is

$$S = [0, 1, 1, 0, 1, 0, 1]$$

(a) $\frac{3}{4}$

(b) $\frac{4}{3}$

(c) $\frac{4}{7}$

(d) $\frac{3}{7}$

Sol. (b)

$$Pr(X = 1) = \frac{4}{7}$$

$$\text{Logit} = \frac{p(x)}{1-p(x)} = \frac{4 \times 7}{3 \times 7} = \frac{4}{3}$$

Question 4

The output of binary class logistic regression lies in this range.

- (a) $[-\infty, \infty]$
- (b) $[-1, 1]$
- (c) $[0, 1]$
- (d) $[-\infty, 0]$

Question 4

The output of binary class logistic regression lies in this range.

- (a) $[-\infty, \infty]$
- (b) $[-1, 1]$
- (c) $[0, 1]$
- (d) $[-\infty, 0]$

Sol. (c)

Question 5

Logistic regression is robust to outliers. Why?

- (a) The squashing of output values between [0, 1] dampens the affect of outliers.
- (b) Linear models are robust to outliers.
- (c) The parameters in logistic regression tend to take small values due to the nature of the problem setting and hence outliers get translated to the same range as other samples.
- (d) The given statement is false.

Question 5

Logistic regression is robust to outliers. Why?

- (a) The squashing of output values between [0, 1] dampens the affect of outliers.
- (b) Linear models are robust to outliers.
- (c) The parameters in logistic regression tend to take small values due to the nature of the problem setting and hence outliers get translated to the same range as other samples.
- (d) The given statement is false.

Sol. (a)

Question 6

We have two classes in our dataset with mean 0 and 1, and variance 2 and 3.

- (a) LDA may be able to classify them perfectly.
- (b) LDA will definitely be able to classify them perfectly.
- (c) LDA will definitely NOT be able to classify them perfectly.
- (d) None of the above.

Question 6

We have two classes in our dataset with mean 0 and 1, and variance 2 and 3.

- (a) LDA may be able to classify them perfectly.
- (b) LDA will definitely be able to classify them perfectly.
- (c) LDA will definitely NOT be able to classify them perfectly.
- (d) None of the above.

Sol. (c)

The two classes overlap and hence cannot be classified perfectly by LDA.

Question 7

With respect to Linear Discriminant Analysis, which of the following is/are true. (Consider a two class case)

- a. When both the covariance matrices are spherical and equal, the decision boundary will be the perpendicular bisector of the line joining the means.
- b. When both the covariance matrices are spherical and equal, the decision boundary will be perpendicular to the line joining the means.
- c. When both the covariance matrices are spherical and equal and the priors $\pi_1 = \pi_2$ then the decision boundary will be perpendicular bisector of the line joining the means.

Question 7

With respect to Linear Discriminant Analysis, which of the following is/are true. (Consider a two class case)

- a. When both the covariance matrices are spherical and equal, the decision boundary will be the perpendicular bisector of the line joining the means.
- b. When both the covariance matrices are spherical and equal, the decision boundary will be perpendicular to the line joining the means.
- c. When both the covariance matrices are spherical and equal and the priors $\pi_1 = \pi_2$ then the decision boundary will be perpendicular bisector of the line joining the means.

Answer: b,c

Question 8

In general, which of the following classification methods is the most resistant to gross outliers?

- (a) Quadratic Discriminant Analysis (QDA)
- (b) Linear Regression
- (c) Logistic regression
- (d) Linear Discriminant Analysis (LDA)

Question 8

In general, which of the following classification methods is the most resistant to gross outliers?

- (a) Quadratic Discriminant Analysis (QDA)
- (b) Linear Regression
- (c) Logistic regression
- (d) Linear Discriminant Analysis (LDA)

Sol. (c)

In general, a good way to tell if a method is sensitive to outliers is to look at the loss it incurs upon ignoring outliers.

Linear Regression uses a square loss and thus, outliers that are far away from the hyperplane contribute significantly to the loss.

LDA and QDA both use the L2-Norm and, for the same reason, sensitive to outliers.

Logistic Regression weights the points close to the boundary higher than points far away. This is an implication of the Logistic loss function (beyond the boundary, roughly linear instead of quadratic).

Question 9

In a binary classification scenario where x is the independent variable and y is the dependent variable, logistic regression assumes that the conditional distribution $y|x$ follows a

- (a) Bernoulli distribution
- (b) binomial distribution
- (c) normal distribution
- (d) exponential distribution

Question 9

In a binary classification scenario where x is the independent variable and y is the dependent variable, logistic regression assumes that the conditional distribution $y|x$ follows a

- (a) Bernoulli distribution
- (b) binomial distribution
- (c) normal distribution
- (d) exponential distribution

Sol. (a)

The dependent variable is binary, so a Bernoulli distribution is assumed.

Question 10

We have seen methods like Ridge and lasso to reduce variance among the co-efficients. We can use these methods to do feature selection also. Which one of them is more appropriate?

- (a) Ridge
- (b) Lasso

Question 10

We have seen methods like Ridge and lasso to reduce variance among the co-efficients. We can use these methods to do feature selection also. Which one of them is more appropriate?

- (a) Ridge
- (b) Lasso

Sol. (b)

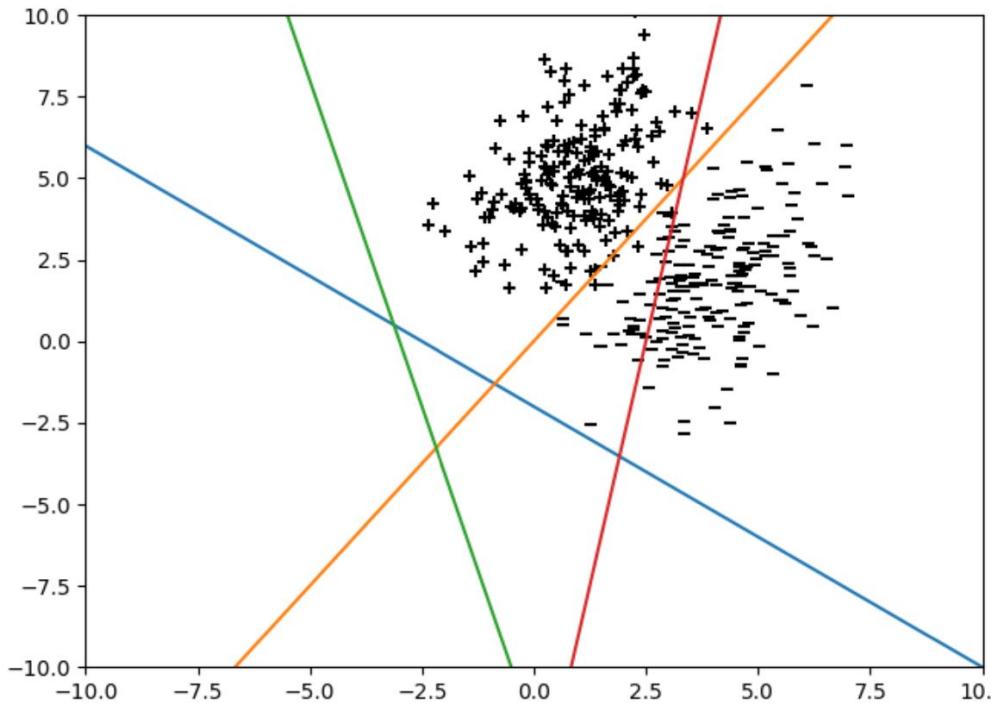
For feature selection, we would prefer to use lasso since solving the optimisation problem when using lasso will cause some of the coefficients to be exactly zero (depending of course on the data) whereas with ridge regression, the magnitude of the coefficients will be reduced, but won't go down to zero.

Question 11

For the two classes '+' and '-' shown below.

While performing LDA on it, which line is the most appropriate for projecting data points?

- (a) Red
- (b) Orange
- (c) Blue
- (d) Green



Question 11

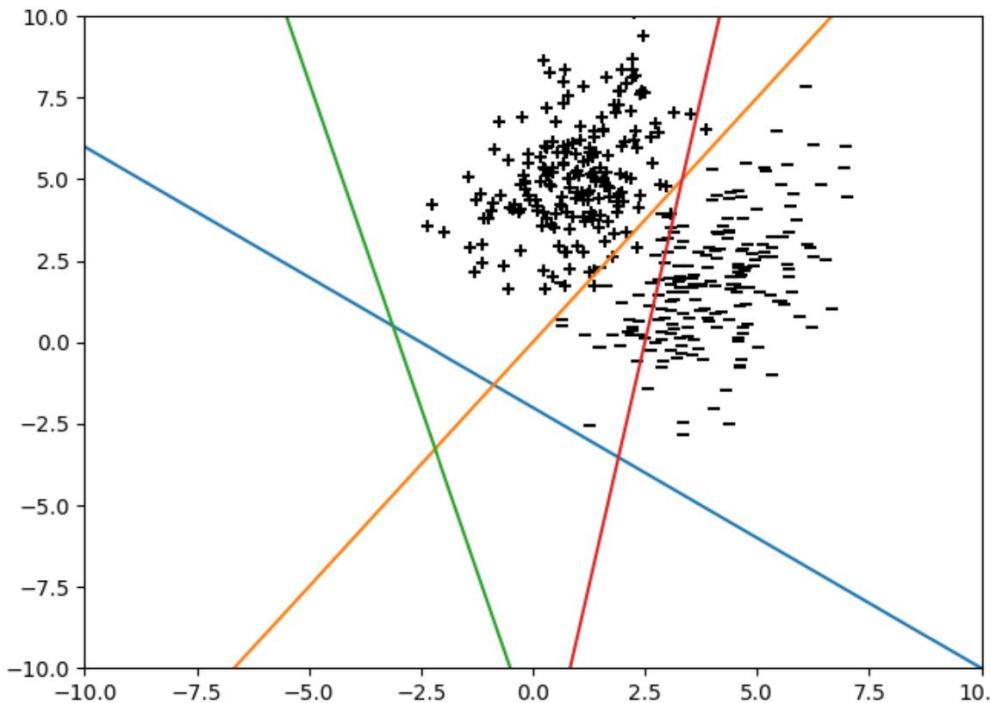
For the two classes '+' and '-' shown below.

While performing LDA on it, which line is the most appropriate for projecting data points?

- (a) Red
- (b) Orange
- (c) Blue
- (d) Green

Answer: (c)

The blue line is parallel to the line joining the mean of the clusters and will therefore maximize the distance between the means.



THE END

NPTEL Live Session

Week 4

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

20/02/2024

Question 1

Consider a Boolean function in three variables, that returns True if two or more variables out of three are True, and False otherwise. Can this function be implemented using the perceptron algorithm?

- (a) no
- (b) yes

Question 1

Consider a Boolean function in three variables, that returns True if two or more variables out of three are True, and False otherwise. Can this function be implemented using the perceptron algorithm?

- (a) no
- (b) yes

Sol. (b)

Question 2

For a support vector machine model, let x_i be an input instance with label y_i . If $y_i(\hat{\beta}_0 + x_i^T \hat{\beta}) > 1$, where $\hat{\beta}_0$ and $\hat{\beta}$ are the estimated parameters of the model, then

- (a) x_i is not a support vector
- (b) x_i is a support vector
- (c) x_i is either an outlier or a support vector
- (d) Depending upon other data points, x_i may or may not be a support vector.

Question 2

For a support vector machine model, let x_i be an input instance with label y_i . If $y_i(\hat{\beta}_0 + x_i^T \hat{\beta}) > 1$, where $\hat{\beta}_0$ and $\hat{\beta}$ are the estimated parameters of the model, then

- (a) x_i is not a support vector
- (b) x_i is a support vector
- (c) x_i is either an outlier or a support vector
- (d) Depending upon other data points, x_i may or may not be a support vector.

Sol. (a)

Question 3

Suppose we use a linear kernel SVM to build a classifier for a 2-class problem where the training data points are linearly separable. In general, will the classifier trained in this manner be always the same as the classifier trained using the perceptron training algorithm on the same training data?

- (a) Yes
- (b) No

Question 3

Suppose we use a linear kernel SVM to build a classifier for a 2-class problem where the training data points are linearly separable. In general, will the classifier trained in this manner be always the same as the classifier trained using the perceptron training algorithm on the same training data?

- (a) Yes
- (b) No

Sol. (b) The hyperplane returned by the SVM approach will have a maximal margin, whereas no such guarantee can be given for the hyperplane identified using the perceptron training algorithm.

Question 4

Which of the following is/are true regarding an SVM?

- (a) For two dimensional data points, the separating hyperplane learnt by a linear SVM will be a straight line.
- (b) In theory, a Gaussian kernel SVM can model any complex separating hyperplane.
- (c) Overfitting in SVM is a function of number of support vectors

Question 4

Which of the following is/are true regarding an SVM? (Multiple options may be correct)

- (a) For two dimensional data points, the separating hyperplane learnt by a linear SVM will be a straight line.
- (b) In theory, a Gaussian kernel SVM can model any complex separating hyperplane.
- (c) Overfitting in SVM is a function of number of support vectors

Answer: a,b,c

Question 5

Consider the 1 dimensional dataset:

x	y
-1	1
0	-1
2	1

(Note: x is the feature, and y is the output)

State true or false: The dataset becomes linearly separable after using basis expansion with the following basis function $\phi(x) = \begin{bmatrix} 1 \\ x^3 \end{bmatrix}$

- (a) True
- (b) False

Question 5

Consider the 1 dimensional dataset:

x	y
-1	1
0	-1
2	1

(Note: x is the feature, and y is the output)

State true or false: The dataset becomes linearly separable after using basis expansion with the following basis function $\phi(x) = \begin{bmatrix} 1 \\ x^3 \end{bmatrix}$

- (a) True
- (b) False

Sol. (b)

After applying basis expansion, $x_1' = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $x_2' = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and $x_3' = \begin{bmatrix} 1 \\ 8 \end{bmatrix}$. Plotting these, we can see that the data points are not linearly separable.

Question 6

State whether True or False.

After training an SVM, we can discard all example which are not support vectors and can still classify new examples.

- A) True
- B) False

Question 6

State whether True or False.

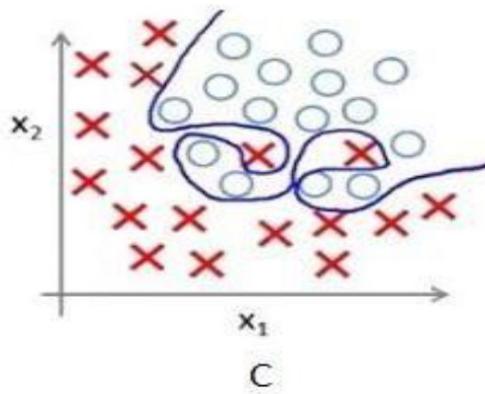
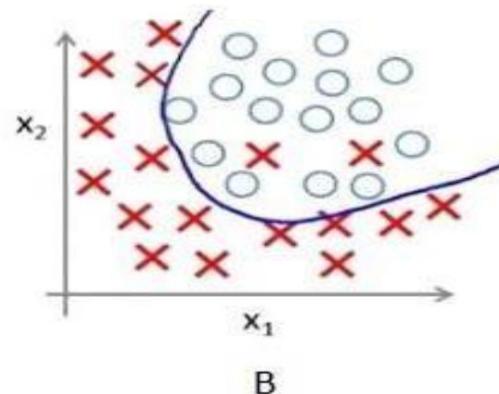
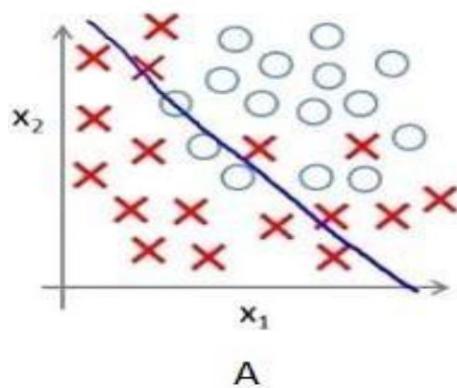
After training an SVM, we can discard all examples which are not support vectors and can still classify new examples.

- A) TRUE**
- B) FALSE**

Answer : A) True

Question 7

Below are the labelled instances of 2 classes and hand drawn decision boundaries. Which of the following figure demonstrates overfitting of the training data?



Question 7

Answer : C

In figure 3, the decision boundary is very complex and unlikely to generalize the data.

Question 8

A hypothetical SVM model has the following values of lagrange multipliers and support vectors:

α	<i>Support vector</i>	y
1	(1, -1, 1)	+1
0.5	(0, 2, -1)	-1
1	(-1, 0, 2)	-1

Compute the coefficients (β) of the linear separator. (without the intercept term)

Question 8

Answer :

$$\beta = \sum_{i \in SV} \alpha_i y_i X_i$$

$$\beta = (2, -2, -0.5)$$

α	<i>Support vector</i>	y
1	(1, -1, 1)	+1
0.5	(0, 2, -1)	-1
1	(-1, 0, 2)	-1

Question 9

Suppose we have four training examples in two dimensions, positive examples at $X_1 = [0, 0]$, $X_2 = [2, 2]$, and negative examples at $X_3 = [h, 1]$, $X_4 = [0, 3]$, where we treat $0 \leq h \leq 3$ as a parameter.

How large can h be so that the training points are still linearly separable?

Question 9

Suppose we have four training examples in two dimensions, positive examples at $X_1 = [0, 0]$, $X_2 = [2, 2]$, and negative examples at $X_3 = [h, 1]$, $X_4 = [0, 3]$, where we treat $0 \leq h \leq 3$ as a parameter.

How large can h be so that the training points are still linearly separable?

Answer : $h < 1$

Question 10

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 $K(x, z) = \phi(x) \cdot \phi(z)$. Consider the two dimensional input vectors $x = (x_1, x_2)$.

For the kernel function $K(x, z) = 1 + x \cdot z$, what is the corresponding feature transform $\phi(x)$?

- (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) $\phi(x) = (1, x_1^2, x_2^2)$

Question 10

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 $K(x, z) = \phi(x) \cdot \phi(z)$. Consider the two dimensional input vectors $x = (x_1, x_2)$.

For the kernel function $K(x, z) = 1 + x \cdot z$, what is the corresponding feature transform $\phi(x)$?

- (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) $\phi(x) = (1, x_1^2, x_2^2)$

Answer : (B) $\phi(x) = (1, x_1, x_2)$

THE END

NPTEL Live Session

Week 5

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

27/02/2024

Question 1

Which of the following are correct?

- (a) A perceptron will learn the underlying linearly separable boundary with finite number of training steps.
- (b) XOR function can be modelled by a single perceptron.
- (c) Backpropagation algorithm used while estimating parameters of neural networks actually uses gradient descent algorithm.
- (d) The backpropagation algorithm will always converge to global optimum, which is one of the reasons for impressive performance of neural networks.

Question 1

Which of the following are correct?

- (a) A perceptron will learn the underlying linearly separable boundary with finite number of training steps.
- (b) XOR function can be modelled by a single perceptron.
- (c) Backpropagation algorithm used while estimating parameters of neural networks actually uses gradient descent algorithm.
- (d) The backpropagation algorithm will always converge to global optimum, which is one of the reasons for impressive performance of neural networks.

Sol. (a), (c)

Question 2

Recall the XOR(tabulated below) example from class where we did a transformation of features to make it linearly separable. Which of the following transformations can also work?

X_1	X_2	Y
-1	-1	-1
1	-1	1
-1	1	1
1	1	-1

- (a) $X'_1 = X_1^2, X'_2 = X_2^2$
- (b) $X'_1 = 1 + X_1, X'_2 = 1 - X_2$
- (c) $X'_1 = X_1X_2, X'_2 = -X_1X_2$
- (d) $X'_1 = (X_1 - X_2)^2, X'_2 = (X_1 + X_2)^2$

Question 2

Sol. (c), (d)

(c)

X'_1	X'_2	Y
1	-1	-1
-1	1	1
-1	1	1
1	-1	-1

(d)

X'_1	X'_2	Y
0	4	-1
4	0	1
4	0	1
0	4	-1

The two transformations above are linearly separable.

Question 3

What is the effect of using activation function $f(x) = x$ for hidden layers in an ANN?

- (a) No effect. It's as good as any other activation function (sigmoid, tanh etc).
- (b) The ANN is equivalent to doing multi-output linear regression.
- (c) Backpropagation will not work.
- (d) We can model highly complex non-linear functions.

Question 3

What is the effect of using activation function $f(x) = x$ for hidden layers in an ANN?

- (a) No effect. It's as good as any other activation function (sigmoid, tanh etc).
- (b) The ANN is equivalent to doing multi-output linear regression.
- (c) Backpropagation will not work.
- (d) We can model highly complex non-linear functions.

Answer : b. The ANN is equivalent to doing multi-output linear regression.

Question 4

Which of the following functions can be used on the last layer of an ANN for classification?

- (a) Softmax
- (b) Sigmoid
- (c) Tanh
- (d) Linear

Question 4

Which of the following functions can be used on the last layer of an ANN for classification?

- (a) Softmax
- (b) Sigmoid
- (c) Tanh
- (d) Linear

Answer : a,b,c

Question 5

Statement: Threshold function cannot be used as activation function for hidden layers.

Reason: Threshold functions do not introduce non-linearity.

- (a) Statement is true and reason is false.
- (b) Statement is false and reason is true.
- (c) Both are true and the reason explains the statement.
- (d) Both are true and the reason does not explain the statement.

Question 5

Statement: Threshold function cannot be used as activation function for hidden layers.

Reason: Threshold functions do not introduce non-linearity.

- (a) Statement is true and reason is false.
- (b) Statement is false and reason is true.
- (c) Both are true and the reason explains the statement.
- (d) Both are true and the reason does not explain the statement.

Answer : (a) Statement is true and reason is false.

Question 6

We use several techniques to ensure the weights of the neural network are small (such as random initialization around 0 or regularisation). What conclusions can we draw if weights of our ANN are high?

- (a) Model has overfitted.
- (b) It was initialized incorrectly.
- (c) At least one of (a) or (b).
- (d) None of the above.

Question 6

We use several techniques to ensure the weights of the neural network are small (such as random initialization around 0 or regularisation). What conclusions can we draw if weights of our ANN are high?

- (a) Model has overfitted.
- (b) It was initialized incorrectly.
- (c) At least one of (a) or (b).
- (d) None of the above.

Answer : (d)

Overfitting may happen because of high weights but the two are not always associated.

Question 7

Which of the following are true when comparing ANNs and SVMs?

- (a) ANN error surface has multiple local minima while SVM error surface has only one minima
- (b) After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.
- (c) As shown for Perceptron, there are some classes of functions that cannot be learnt by an ANN. An SVM can learn a hyperplane for any kind of distribution.
- (d) In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.

Question 7

Which of the following are true when comparing ANNs and SVMs?

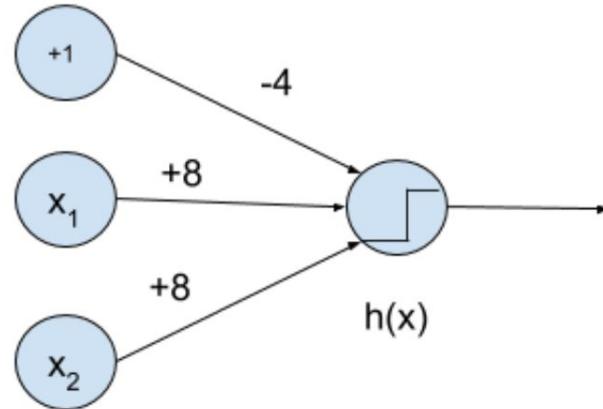
- (a) ANN error surface has multiple local minima while SVM error surface has only one minima
- (b) After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.
- (c) As shown for Perceptron, there are some classes of functions that cannot be learnt by an ANN. An SVM can learn a hyperplane for any kind of distribution.
- (d) In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.

Answer: a,b,d

According to Universal approximation theorem option c is not true.

Question 8

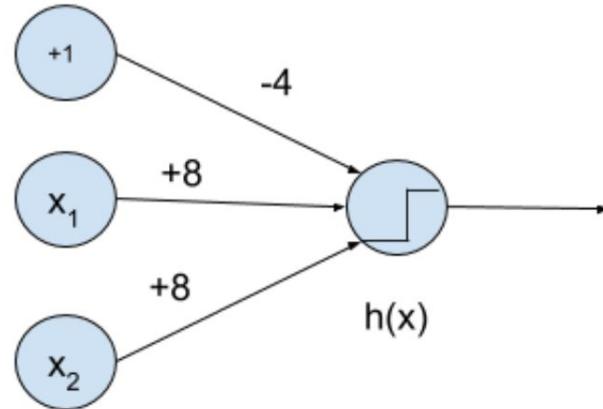
You are given the following neural networks which take two binary valued inputs $x_1, x_2 \in \{0, 1\}$ and the activation function is the threshold function($h(x) = 1$ if $x > 0$; 0 otherwise). Which of the following logical functions does it compute?



- (a) OR
- (b) AND
- (c) NAND
- (d) None of the above.

Question 8

You are given the following neural networks which take two binary valued inputs $x_1, x_2 \in \{0, 1\}$ and the activation function is the threshold function($h(x) = 1$ if $x > 0$; 0 otherwise). Which of the following logical functions does it compute?



- (a) OR
- (b) AND
- (c) NAND
- (d) None of the above.

Sol. (a)

Question 9

Choose the correct statement (multiple may be correct):

- (a) MLE is a special case of MAP when prior is a uniform distribution.
- (b) MLE acts as regularisation for MAP.
- (c) MLE is a special case of MAP when prior is a beta distribution .
- (d) MAP acts as regularisation for MLE.

Question 9

Choose the correct statement (multiple may be correct):

- (a) MLE is a special case of MAP when prior is a uniform distribution.
- (b) MLE acts as regularisation for MAP.
- (c) MLE is a special case of MAP when prior is a beta distribution .
- (d) MAP acts as regularisation for MLE.

Answer : a,d

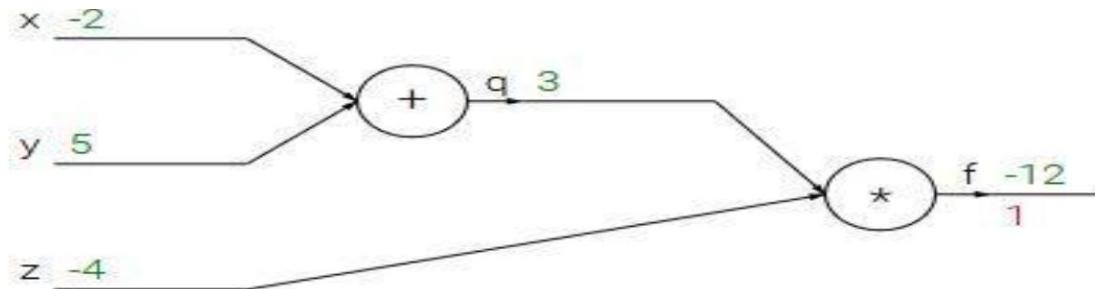
Question 10

Suppose you have inputs as x , y , and z with values -2, 5, and -4 respectively. You have a neuron ' q ' and neuron ' f ' with functions:

$$q = x + y$$

$$f = q * z$$

Graphical representation of the functions is as follows :



What is the gradient of f with respect to x , y , and z ?

Question 10

Answer : (- 4, - 4, 3)

Question 11

In training a neural network, we notice that the loss does not decrease in the first few starting epochs. What is the reason for this?

- A) The learning Rate is low.
- B) Regularization Parameter is High.
- C) Stuck at the Local Minima.
- D) All of these could be the reason.

Question 11

In training a neural network, we notice that the loss does not decrease in the first few starting epochs. What is the reason for this?

- A) The learning Rate is low.
- B) Regularization Parameter is High.
- C) Stuck at the Local Minima.
- D) **All of these could be the reason.**

Answer: D

The problem can occur due to any one of the reasons above.

THE END

NPTEL Live Session

Week 5

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

27/02/2024

Question 1

Which of the following are correct?

- (a) A perceptron will learn the underlying linearly separable boundary with finite number of training steps.
- (b) XOR function can be modelled by a single perceptron.
- (c) Backpropagation algorithm used while estimating parameters of neural networks actually uses gradient descent algorithm.
- (d) The backpropagation algorithm will always converge to global optimum, which is one of the reasons for impressive performance of neural networks.

Question 1

Which of the following are correct?

- (a) A perceptron will learn the underlying linearly separable boundary with finite number of training steps.
- (b) XOR function can be modelled by a single perceptron.
- (c) Backpropagation algorithm used while estimating parameters of neural networks actually uses gradient descent algorithm.
- (d) The backpropagation algorithm will always converge to global optimum, which is one of the reasons for impressive performance of neural networks.

Sol. (a), (c)

Question 2

Recall the XOR(tabulated below) example from class where we did a transformation of features to make it linearly separable. Which of the following transformations can also work?

X_1	X_2	Y
-1	-1	-1
1	-1	1
-1	1	1
1	1	-1

- (a) $X'_1 = X_1^2, X'_2 = X_2^2$
- (b) $X'_1 = 1 + X_1, X'_2 = 1 - X_2$
- (c) $X'_1 = X_1X_2, X'_2 = -X_1X_2$
- (d) $X'_1 = (X_1 - X_2)^2, X'_2 = (X_1 + X_2)^2$

Question 2

Sol. (c), (d)

(c)

X'_1	X'_2	Y
1	-1	-1
-1	1	1
-1	1	1
1	-1	-1

(d)

X'_1	X'_2	Y
0	4	-1
4	0	1
4	0	1
0	4	-1

The two transformations above are linearly separable.

Question 3

What is the effect of using activation function $f(x) = x$ for hidden layers in an ANN?

- (a) No effect. It's as good as any other activation function (sigmoid, tanh etc).
- (b) The ANN is equivalent to doing multi-output linear regression.
- (c) Backpropagation will not work.
- (d) We can model highly complex non-linear functions.

Question 3

What is the effect of using activation function $f(x) = x$ for hidden layers in an ANN?

- (a) No effect. It's as good as any other activation function (sigmoid, tanh etc).
- (b) The ANN is equivalent to doing multi-output linear regression.
- (c) Backpropagation will not work.
- (d) We can model highly complex non-linear functions.

Answer : b. The ANN is equivalent to doing multi-output linear regression.

Question 4

Which of the following functions can be used on the last layer of an ANN for classification?

- (a) Softmax
- (b) Sigmoid
- (c) Tanh
- (d) Linear

Question 4

Which of the following functions can be used on the last layer of an ANN for classification?

- (a) Softmax
- (b) Sigmoid
- (c) Tanh
- (d) Linear

Answer : a,b,c

Question 5

Statement: Threshold function cannot be used as activation function for hidden layers.

Reason: Threshold functions do not introduce non-linearity.

- (a) Statement is true and reason is false.
- (b) Statement is false and reason is true.
- (c) Both are true and the reason explains the statement.
- (d) Both are true and the reason does not explain the statement.

Question 5

Statement: Threshold function cannot be used as activation function for hidden layers.

Reason: Threshold functions do not introduce non-linearity.

- (a) Statement is true and reason is false.
- (b) Statement is false and reason is true.
- (c) Both are true and the reason explains the statement.
- (d) Both are true and the reason does not explain the statement.

Answer : (a) Statement is true and reason is false.

Question 6

We use several techniques to ensure the weights of the neural network are small (such as random initialization around 0 or regularisation). What conclusions can we draw if weights of our ANN are high?

- (a) Model has overfitted.
- (b) It was initialized incorrectly.
- (c) At least one of (a) or (b).
- (d) None of the above.

Question 6

We use several techniques to ensure the weights of the neural network are small (such as random initialization around 0 or regularisation). What conclusions can we draw if weights of our ANN are high?

- (a) Model has overfitted.
- (b) It was initialized incorrectly.
- (c) At least one of (a) or (b).
- (d) None of the above.

Answer : (d)

Overfitting may happen because of high weights but the two are not always associated.

Question 7

Which of the following are true when comparing ANNs and SVMs?

- (a) ANN error surface has multiple local minima while SVM error surface has only one minima
- (b) After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.
- (c) As shown for Perceptron, there are some classes of functions that cannot be learnt by an ANN. An SVM can learn a hyperplane for any kind of distribution.
- (d) In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.

Question 7

Which of the following are true when comparing ANNs and SVMs?

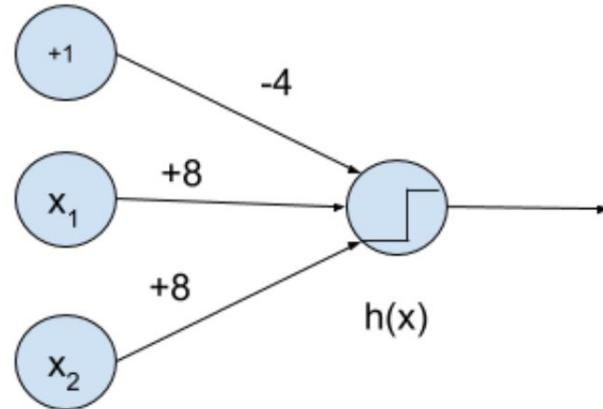
- (a) ANN error surface has multiple local minima while SVM error surface has only one minima
- (b) After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.
- (c) As shown for Perceptron, there are some classes of functions that cannot be learnt by an ANN. An SVM can learn a hyperplane for any kind of distribution.
- (d) In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.

Answer: a,b,d

According to Universal approximation theorem option c is not true.

Question 8

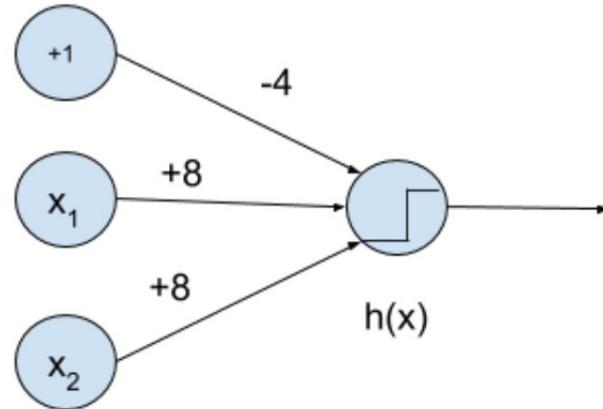
You are given the following neural networks which take two binary valued inputs $x_1, x_2 \in \{0, 1\}$ and the activation function is the threshold function($h(x) = 1$ if $x > 0$; 0 otherwise). Which of the following logical functions does it compute?



- (a) OR
- (b) AND
- (c) NAND
- (d) None of the above.

Question 8

You are given the following neural networks which take two binary valued inputs $x_1, x_2 \in \{0, 1\}$ and the activation function is the threshold function($h(x) = 1$ if $x > 0$; 0 otherwise). Which of the following logical functions does it compute?



- (a) OR
- (b) AND
- (c) NAND
- (d) None of the above.

Sol. (a)

Question 9

Choose the correct statement (multiple may be correct):

- (a) MLE is a special case of MAP when prior is a uniform distribution.
- (b) MLE acts as regularisation for MAP.
- (c) MLE is a special case of MAP when prior is a beta distribution .
- (d) MAP acts as regularisation for MLE.

Question 9

Choose the correct statement (multiple may be correct):

- (a) MLE is a special case of MAP when prior is a uniform distribution.
- (b) MLE acts as regularisation for MAP.
- (c) MLE is a special case of MAP when prior is a beta distribution .
- (d) MAP acts as regularisation for MLE.

Answer : a,d

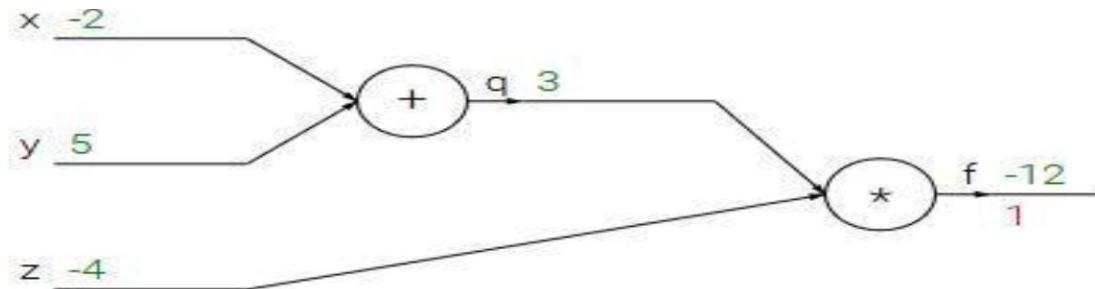
Question 10

Suppose you have inputs as x , y , and z with values -2, 5, and -4 respectively. You have a neuron ' q ' and neuron ' f ' with functions:

$$q = x + y$$

$$f = q * z$$

Graphical representation of the functions is as follows :



What is the gradient of f with respect to x , y , and z ?

Question 10

Answer : (- 4, - 4, 3)

Question 11

In training a neural network, we notice that the loss does not decrease in the first few starting epochs. What is the reason for this?

- A) The learning Rate is low.
- B) Regularization Parameter is High.
- C) Stuck at the Local Minima.
- D) All of these could be the reason.

Question 11

In training a neural network, we notice that the loss does not decrease in the first few starting epochs. What is the reason for this?

- A) The learning Rate is low.
- B) Regularization Parameter is High.
- C) Stuck at the Local Minima.
- D) All of these could be the reason.

Answer: D

The problem can occur due to any one of the reasons above.

THE END

NPTEL Live Session

Week 7

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

12/3/2024

Question 1

You have 2 binary classifiers A and B. A has accuracy=0% and B has accuracy=50%. Which classifier is more useful?

- (a) A
- (b) B
- (c) Both are good
- (d) Cannot say

Question 1

You have 2 binary classifiers A and B. A has accuracy=0% and B has accuracy=50%. Which classifier is more useful?

- (a) A
- (b) B
- (c) Both are good
- (d) Cannot say

Sol. (a)

Flip the labels and get 100% accuracy!

Question 2

Sensitivity is same as:

- a. Precision
- b. Recall
- c. Specificity
- d. None of the above

Question 2

Sensitivity is same as:

- a. Precision
- b. Recall
- c. Specificity
- d. None of the above

Answer : Recall

Question 3

Using the bootstrap approach for sampling, the new dataset will have _____ of the original samples on expectation.

- (a) 50.0%
- (b) 56.8%
- (c) 63.2%
- (d) 73.6%

Question 3

Using the bootstrap approach for sampling, the new dataset will have _____ of the original samples on expectation.

- (a) 50.0%
- (b) 56.8%
- (c) 63.2%
- (d) 73.6%

Sol. (c)

Refer to lecture.

Question 4

You have a special case where your data has 10 classes and is sorted according to target labels. You attempt 5-fold cross validation by selecting the folds sequentially. What can you say about your resulting model?

- (a) It will have 100% accuracy.
- (b) It will have 0% accuracy.
- (c) Accuracy will depend on how good the model does.
- (d) Accuracy will depend on the compute power available for training.

Question 4

You have a special case where your data has 10 classes and is sorted according to target labels. You attempt 5-fold cross validation by selecting the folds sequentially. What can you say about your resulting model?

- (a) It will have 100% accuracy.
- (b) It will have 0% accuracy.
- (c) Accuracy will depend on how good the model does.
- (d) Accuracy will depend on the compute power available for training.

Sol. (b)

Due to the nature of the splits, each class part of the test set will not be included in the training sets. So the classifier will not even know the class exists and therefore never predict it.

Question 5

Given the following information

	Actual Positive	Actual Negative
Predicted Positive	35	15
Predicted Negative	45	20

What is the precision and recall?

- (a) 0.5, 0.4375
- (b) 0.7, 0.636
- (c) 0.6, 0.636
- (d) 0.7, 0.4375
- (e) None of the above

Question 5

Given the following information

	Actual Positive	Actual Negative
Predicted Positive	35	15
Predicted Negative	45	20

What is the precision and recall?

Sol. (d)

$$\text{Precision} = 35 / (35+15) = 0.7$$

$$\text{Recall} = 35 / (35+45) = 0.4375$$

Question 6

Suppose I have 10,000 emails in my mailbox out of which 200 are spams. The spam detection system detects 150 mails as spams, out of which 50 are actually spams. What is the precision and recall of my spam detection system?

- a. Precision = 33.333%, Recall = 25%
- b. Precision = 25%, Recall = 33.33%
- c. Precision = 33.33%, Recall = 75%
- d. Precision = 75%, Recall = 33.33%

Question 6

Suppose I have 10,000 emails in my mailbox out of which 200 are spams. The spam detection system detects 150 mails as spams, out of which 50 are actually spams. What is the precision and recall of my spam detection system?

- a. **Precision = 33.333%, Recall = 25%**
- b. Precision = 25%, Recall = 33.33%
- c. Precision = 33.33%, Recall = 75%
- d. Precision = 75%, Recall = 33.33%

Answer : a

$$\begin{aligned}\text{Precision} &= \frac{Tp}{Tp+Fp} \\&= \frac{50}{150} \\&= 33.333\%\end{aligned}$$

$$\begin{aligned}\text{Recall} &= \frac{Tp}{Tp+Fn} \\&= \frac{50}{200} \\&= 25\%\end{aligned}$$

Question 7

What is the effect of using bagging on weak classifiers for variance?

- (a) Increases variance
- (b) Reduces variance
- (c) Does not change

Question 7

What is the effect of using bagging on weak classifiers for variance?

- (a) Increases variance
- (b) Reduces variance
- (c) Does not change

Sol. (b)

Refer to lecture.

Question 8

You are building a model to detect cancer. Which metric will you prefer for evaluating your model?

- (a) Accuracy
- (b) Sensitivity
- (c) Specificity
- (d) MSE

Question 8

You are building a model to detect cancer. Which metric will you prefer for evaluating your model?

- (a) Accuracy
- (b) Sensitivity
- (c) Specificity
- (d) MSE

Answer : (b)

You want to make sure that if a person has cancer, it is detected.

Question 9

Regarding bias and variance, which of the following statements are true? (Here ‘high’ and ‘low’ are relative to the ideal model.)

- a. Models which overfit have a high bias.
- b. Models which overfit have a low bias.
- c. Models which underfit have a high variance.
- d. Models which underfit have a low variance.

Question 9

Regarding bias and variance, which of the following statements are true? (Here ‘high’ and ‘low’ are relative to the ideal model.)

- a. Models which overfit have a high bias.
- b. Models which overfit have a low bias.
- c. Models which underfit have a high variance.
- d. Models which underfit have a low variance.

Answer : b,d

In supervised learning, underfitting happens when a model is unable to capture the underlying pattern of the data. These models usually have high bias and low variance. Overfitting happens when our model captures the noise along with the underlying pattern in data. These models have low bias and high variance.

Question 10

Considering the AdaBoost algorithm, which among the following statements are false?

- a. In each stage, we try to train a classifier which makes accurate predictions on any subset of the data points where the subset size is at least half the size of the data set
- b. In each stage, we try to train a classifier which makes accurate predictions on a subset of the data points where the subset contains more of the data points which were misclassified in earlier stages
- c. The weight assigned to an individual classifier depends upon the number of data points correctly classified by the classifier
- d. The weight assigned to an individual classifier depends upon the weighted sum error of misclassified points for that classifier

Question 10

Considering the AdaBoost algorithm, which among the following statements are false?

- a. In each stage, we try to train a classifier which makes accurate predictions on any subset of the data points where the subset size is at least half the size of the data set
- b. In each stage, we try to train a classifier which makes accurate predictions on a subset of the data points where the subset contains more of the data points which were misclassified in earlier stages
- c. The weight assigned to an individual classifier depends upon the number of data points correctly classified by the classifier
- d. The weight assigned to an individual classifier depends upon the weighted sum error of misclassified points for that classifier

Answer: a,c

Question 11

We have a dataset with 1000 samples and 5 classes for classification. What would be the training size for a 20 fold cross validation?

- (a) 50
- (b) 200
- (c) 800
- (d) 950

Question 11

We have a dataset with 1000 samples and 5 classes for classification. What would be the training size for a 20 fold cross validation?

- (a) 50
- (b) 200
- (c) 800
- (d) 950

Sol. (d)

The training size does not depend on the number of classes. 20 fold means that the dataset will be divided equally 20 and 19 of those will be used for training. Which results in the answer to be 950.

Question 12

Which of the following are true?

TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative

(a) Precision = $\frac{TP}{TP+FP}$

(b) Recall = $\frac{TP}{TP+FN}$

(c) Accuracy = $\frac{2(TP+TN)}{TP+TN+FP+FN}$

(d) Recall = $\frac{FP}{TP+FP}$

Question 12

Which of the following are true?

TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative

(a) Precision = $\frac{TP}{TP+FP}$

(b) Recall = $\frac{TP}{TP+FN}$

(c) Accuracy = $\frac{2(TP+TN)}{TP+TN+FP+FN}$

(d) Recall = $\frac{FP}{TP+FP}$

Sol. (a), (b)

THE END

NPTEL Live Session

Week 8

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

19/3/2024

Question 1

A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the number obtained is 4. Find the probability that the number obtained is actually 4 :

- A. $2/3$
- B. $3/4$
- C. $5/22$
- D. $2/7$

Question 1

A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the number obtained is 4. Find the probability that the number obtained is actually 4 :

- A. 2/3
- B. 3/4
- C. 5/22
- D. 2/7

Correct Answer : D. 2/7

Detailed Solution : Suppose,

A : The man reports that 4 is obtained.

B : Number 4 is obtained

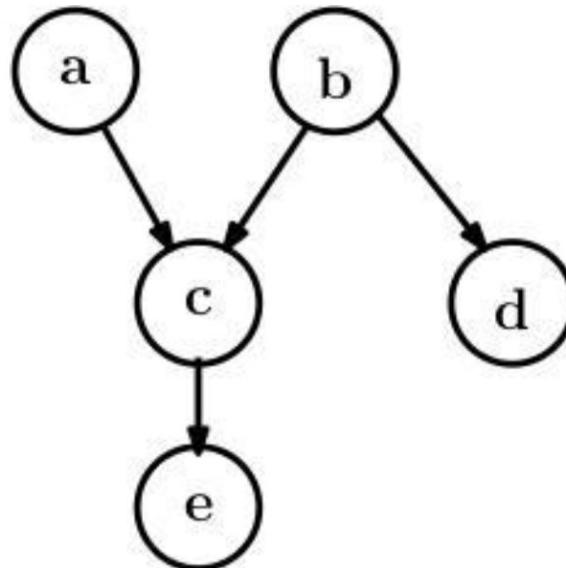
$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})} \text{ here, } P(A|B) = \frac{2}{3}, P(B) = \frac{1}{6}, P(A|\bar{B}) = \frac{1}{3}, P(\bar{B}) = \frac{5}{6}$$

$$P(B|A) = \frac{2}{7}$$

Question 2

Consider the following graphical model, mark which of the following pair of random variables are independent given no evidence?

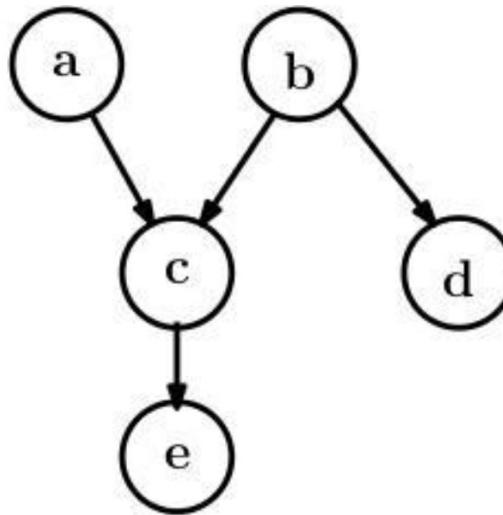
- a. a,b
- b. c,d
- c. e,d
- d. c,e



Question 2

Consider the following graphical model, mark which of the following pair of random variables are independent given no evidence?

- a. a,b
- b. c,d
- c. e,d
- d. c,e



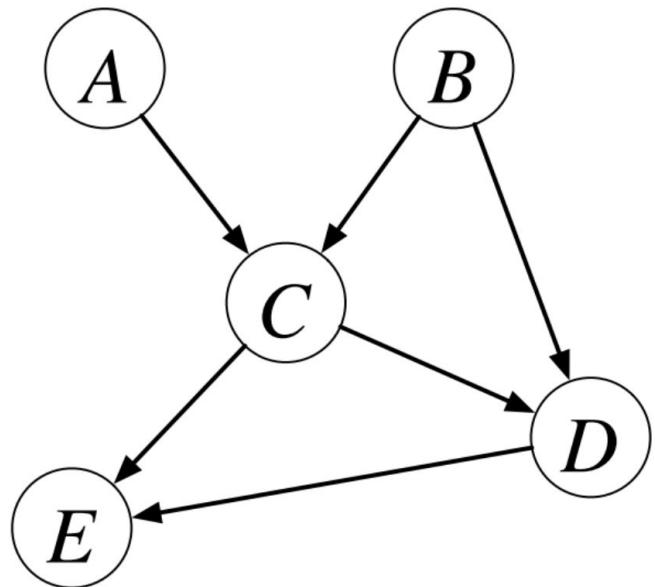
Correct Answer : A. a,b

Detailed Solution : Nodes a and b don't have any predecessor nodes. As they don't have any common parent nodes, a and b are independent.

Question 3

Consider the following graphical model, which of the following are false about the model? (multiple options may be correct)

- (a) A is independent of B when C is known
- (b) D is independent of A when C is known
- (c) D is not independent of A when B is known
- d) D is not independent of A when C is known

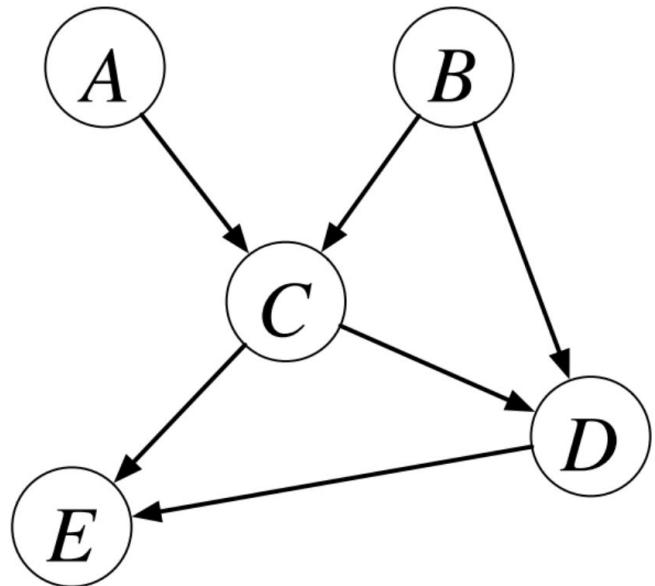


Question 3

Consider the following graphical model, which of the following are false about the model? (multiple options may be correct)

- (a) A is independent of B when C is known
- (b) D is independent of A when C is known
- (c) D is not independent of A when B is known
- d) D is not independent of A when C is known

Sol. (a), (b)



Question 4

Consider the following data for 20 budget phones, 30 mid-range phones, and 20 high-end phones:

Type	Single SIM	5G Comaptability	NFC	Total
Budget	15	5	0	20
Mid-Range	20	20	15	30
High End	15	15	15	20

Consider a phone with 2 SIM card slots and NFC but no 5G compatibility. Calculate the probabilities of this phone being a budget phone, a mid-range phone, and a high-end phone using the Naive Bayes method. The correct ordering of the phone type from the highest to the lowest probability is?

- (a) Budget, Mid-Range, High End
- (b) Budget, High End, Mid-Range
- (c) Mid-Range, High End, Budget
- (d) High End, Mid-Range, Budget

Question 4

Type	Single SIM	5G Comaptability	NFC	Total
Budget	15	5	0	20
Mid-Range	20	20	15	30
High End	15	15	15	20

Sol. (c)

$$P(\text{Class} | x_1, x_2, x_3) \sim P(\text{Class}) * P(x_1 | \text{Class}) * P(x_2 | \text{Class}) * P(x_3 | \text{Class})$$

$$P(\text{Budget} | \text{!SSIM, !5G, NFC}) \sim 20/70 * 5/20 * 15/20 * 0/20 = 0$$

$$P(\text{Mid-range} | \text{!SSIM, !5G, NFC}) \sim 30/70 * 10/30 * 10/30 * 15/30 = 0.0238$$

$$P(\text{High-end} | \text{!SSIM, !5G, NFC}) \sim 20/70 * 5/20 * 5/20 * 15/20 = 0.0134$$

Question 5

A drug test (random variable T) has 1% false positives (i.e., 1% of those not taking drugs show positive in the test), and 5% false negatives (i.e., 5% of those taking drugs test negative). Suppose that 2% of those tested are taking drugs. Determine the probability that somebody who tests positive is actually taking drugs (random variable D).

- A. 0.66
- B. 0.34
- C. 0.50
- D. 0.91

Question 5

A drug test (random variable T) has 1% false positives (i.e., 1% of those not taking drugs show positive in the test), and 5% false negatives (i.e., 5% of those taking drugs test negative). Suppose that 2% of those tested are taking drugs. Determine the probability that somebody who tests positive is actually taking drugs (random variable D).

- A. 0.66
- B. 0.34
- C. 0.50
- D. 0.91

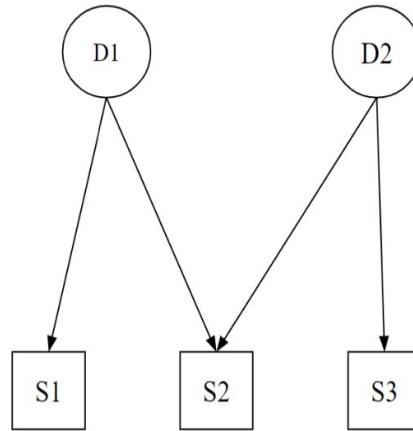
Correct Answer : A. 0.66

Detailed Solution :

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}, P(T|D) = \frac{95}{100}, P(T|\bar{D}) = \frac{1}{100}, P(D) = \frac{2}{100}$$
$$P(D|T) = 0.66$$

Question 6

A patient goes to a doctor with symptoms S1, S2 and S3. The doctor suspects disease D1 and D2 and constructs a Bayesian network for the relation among the disease and symptoms as the following:



What is the joint probability distribution in terms of conditional probabilities?

- A. $P(D1) * P(D2|D1) * P(S1|D1) * P(S2|D1) * P(S3|D2)$
- B. $P(D1) * P(D2) * P(S1|D1) * P(S2|D1) * P(S3|D1, D2)$
- C. $P(D1) * P(D2) * P(S1|D2) * P(S2|D2) * P(S3|D2)$
- D. $P(D1) * P(D2) * P(S1|D1) * P(S2|D1, D2) * P(S3|D2)$

Question 6

Answer : D

From the figure, we can see that D1 and D2 are not dependent on any variable as they don't have any incoming directed edges. S1 has an incoming edge from D1, hence S1 depends on D1. S2 has 2 incoming edges from D1 and D2, hence S2 depends on D1 and D2. S3 has an incoming edge from D2, S3 depends on D2. Hence, (D) is the answer.

Question 7

In a Bayesian network a node with only outgoing edge(s) represents

- a. a variable conditionally independent of the other variables.
- b. a variable dependent on its siblings.
- c. a variable whose dependency is uncertain.
- d. None of the above.

Question 7

In a Bayesian network a node with only outgoing edge(s) represents

- a. **a variable conditionally independent of the other variables.**
- b. a variable dependent on its siblings.
- c. a variable whose dependency is uncertain.
- d. None of the above.

Answer: A. a variable conditionally independent of the other variables.

As there is no incoming edge for the node, the node is not conditionally dependent on any other node.

Question 8

Which among Gradient Boosting and AdaBoost is less susceptible to outliers considering their respective loss functions?

- (a) AdaBoost
- (b) Gradient Boost
- (c) On average, both are equally susceptible.

Question 8

Which among Gradient Boosting and AdaBoost is less susceptible to outliers considering their respective loss functions?

- (a) AdaBoost
- (b) Gradient Boost
- (c) On average, both are equally susceptible.

Sol. (b)

Gradient Boosting (discussed in the lecture) uses a least squares loss function, while AdaBoost uses an exponential loss function. AdaBoost penalizes outliers to an exponential amount, whereas Gradient Boost penalizes them to a lesser extent and, thus, cares less about them.

Question 9

How do you prevent overfitting in random forest models?

- (a) Increasing Tree Depth.
- (b) Increasing the number of variables sampled at each split.
- (c) Increasing the number of trees.
- (d) All of the above.

Question 9

How do you prevent overfitting in random forest models?

- (a) Increasing Tree Depth.
- (b) Increasing the number of variables sampled at each split.
- (c) Increasing the number of trees.
- (d) All of the above.

Sol. (c)

Refer to the lecture.

Question 10

Ensembling in random forest classifier helps in achieving:

- (a) reduction of bias error
- (b) reduction of variance error
- (c) reduction of data dimension
- (d) none of the above

Question 10

Ensembling in random forest classifier helps in achieving:

- (a) reduction of bias error
- (b) reduction of variance error
- (c) reduction of data dimension
- (d) none of the above

Answer : b

THE END

NPTEL Live Session

Week 9

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

26/03/2024

Question 1

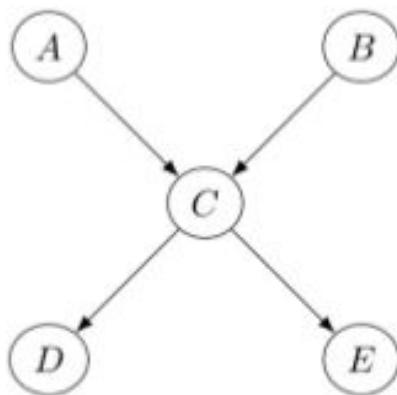


Figure 1: Q1

Consider the bayesian network shown in the Figure. Two students - Student A and Student B make the following claims:

- Student A claims $P(A|\{B,D\})=P(A|D)$
- Student B claims $P(A|B)=P(A)$

Where $P(X|Y)$ denotes probability of event X given Y . Please note that Y can be a set.

Which of the following is true?

- a. Student A and Student B are correct.
- b. Student A is correct and Student B is incorrect.
- c. Student A is incorrect and Student B is correct.
- d. Both are incorrect.

Question 1

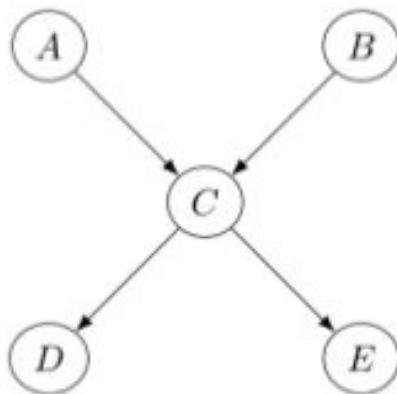


Figure 1: Q1

Consider the bayesian network shown in the Figure. Two students - Student A and Student B make the following claims:

- Student A claims $P(A|\{B, D\}) = P(A|D)$
- Student B claims $P(A|B) = P(A)$

Where $P(X|Y)$ denotes probability of event X given Y . Please note that Y can be a set.

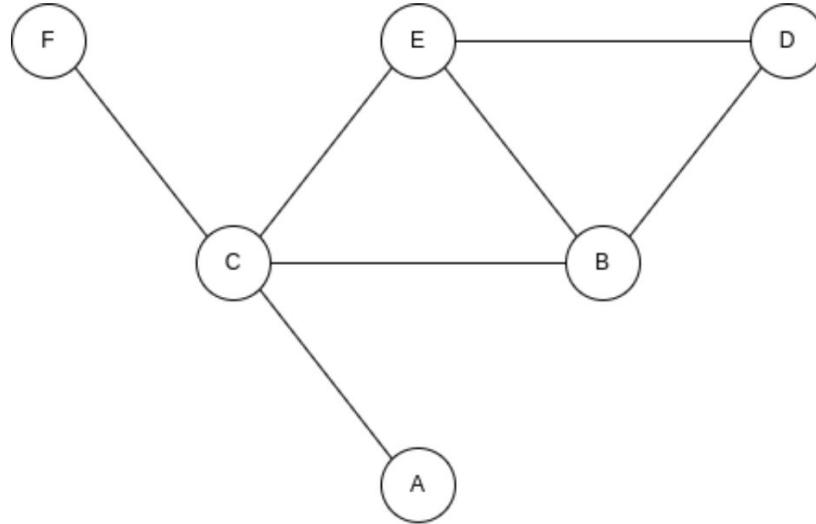
Which of the following is true?

- a. Student A and Student B are correct.
- b. Student A is correct and Student B is incorrect.
- c. Student A is incorrect and Student B is correct.
- d. Both are incorrect.

Answer: c.

Question 2

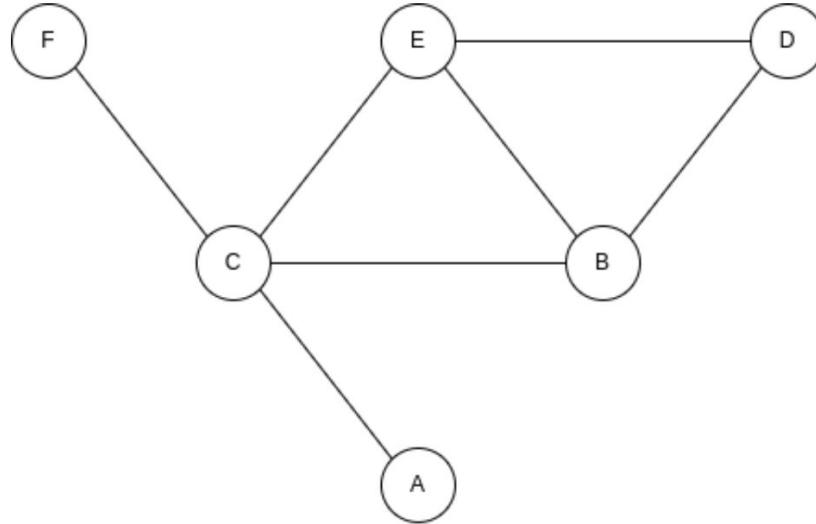
In the undirected graph given below, which nodes are conditionally independent of each other given B? Select all that apply.



- (a) A, D
- (b) D, E
- (c) C, D
- (d) A, F
- (e) None of the above

Question 2

In the undirected graph given below, which nodes are conditionally independent of each other given B? Select all that apply.



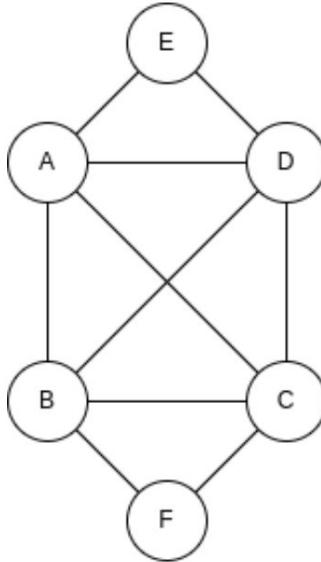
- (a) A, D
- (b) D, E
- (c) C, D
- (d) A, F
- (e) None of the above

Sol. (e)

All pairs have an alternate route to each other that does not pass through B.

Question 3

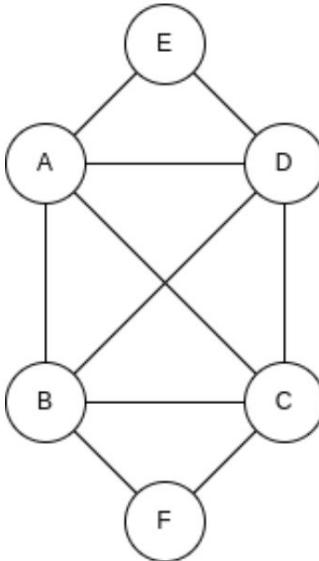
In the undirected graph given below, how many terms will be there in its potential function factorization?



- (a) 7
- (b) 3
- (c) 5
- (d) 9

Question 3

In the undirected graph given below, how many terms will be there in its potential function factorization?



- (a) 7
- (b) 3
- (c) 5
- (d) 9

Sol. (b)

It'll be the same as the number of max-cliques in the graph: (A, B, C, D), (A, E, D), (B, C, F)

Question 4

Consider the following Bayesian network. The random variables given in the model are modeled as discrete variables (Rain = R, Sprinkler = S and Wet Grass = W) and the corresponding probability values are given below.

$$P(R) = 0.1$$

$$P(S) = 0.2$$

$$P(W | R, S) = 0.8$$

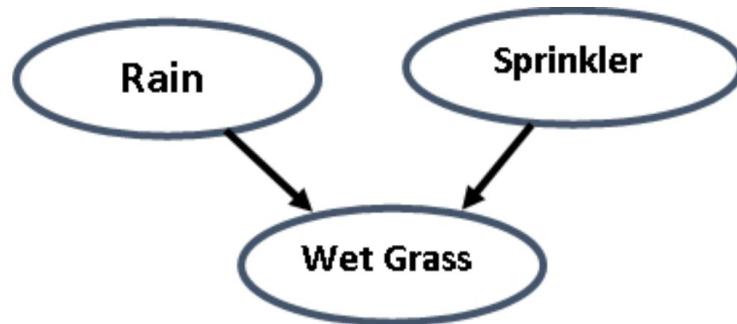
$$P(W | R, \neg S) = 0.7$$

$$P(W | \neg R, S) = 0.6$$

$$P(W | \neg R, \neg S) = 0.5$$

Calculate $P(S | W, R)$.

- A. 1
- B. 0.5
- C. 0.22
- D. 0.78



Question 4

Correct Answer : C. 0.22

Detailed Solution : $P(S|W, R) = \frac{P(W, S, R)}{P(W, R)} = \frac{P(WSR)}{P(WSR) + P(W\bar{S}R)}$

$$P(WSR) = P(W|S, R) * P(R) * P(S) = 0.8 * 0.1 * 0.2 = 0.016$$

$$P(W\bar{S}R) = P(W|\bar{S}, R) * P(R) * P(\bar{S}) = 0.7 * 0.1 * 0.8 = 0.056$$

Question 5

What is the naive assumption in a Naive Bayes Classifier?

- a. All the classes are independent of each other
- b. All the features of a class are independent of each other
- c. The most probable feature for a class is the most important feature to be considered for classification
- d. All the features of a class are conditionally dependent on each other.

Question 5

What is the naive assumption in a Naive Bayes Classifier?

- a. All the classes are independent of each other
- b. All the features of a class are independent of each other
- c. The most probable feature for a class is the most important feature to be considered for classification
- d. All the features of a class are conditionally dependent on each other.

Answer : b

Question 6

A spam filtering system has a probability of 0.95 to correctly classify a mail as spam and 0.10 probability of giving false positives. It is estimated that 1% of the mails are actual spam mails.

Suppose that the system is now given a new mail to be classified as spam/ not-spam, what is the probability that the mail will be classified as spam?

- A. 0.89575
- B. 0.10425
- C. 0.1085
- D. 0.0995

Question 6

Correct Answer: C. 0.1085

Detailed Solution:

Let S = ‘Mails marked spam by the system’, M = ‘Spam mails’.

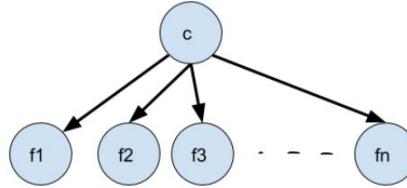
$$P(S|M) = 0.95, P(S|M') = 0.10, P(M) = 0.01$$

We have to find the probability of mail being classified as spam which can either be if a spam mail is correctly classified as spam or if a mail is misclassified as spam.

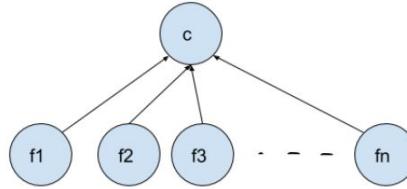
$$P(S) = P(S|M)*P(M) + P(S|M')*P(M') = 0.95 * 0.01 + 0.10 * 0.99 = 0.1085$$

Question 7

Which of the following graphical models capture the Naive Bayes assumption, where c represents the class label and f_i are the features?



(a)

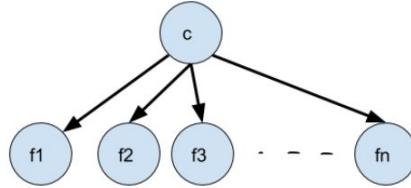


(b)

- (c) It cannot be captured by a graphical model.
- (d) Graphical model can capture the assumption, but the given models don't do it.

Question 7

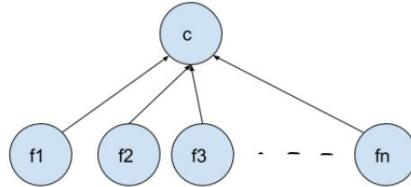
Which of the following graphical models capture the Naive Bayes assumption, where c represents the class label and f_i are the features?



(a)

Solution: A

The bayes assumption states that given the class label, the features are independent. This is captured when the class label is the parent node for all the feature nodes.

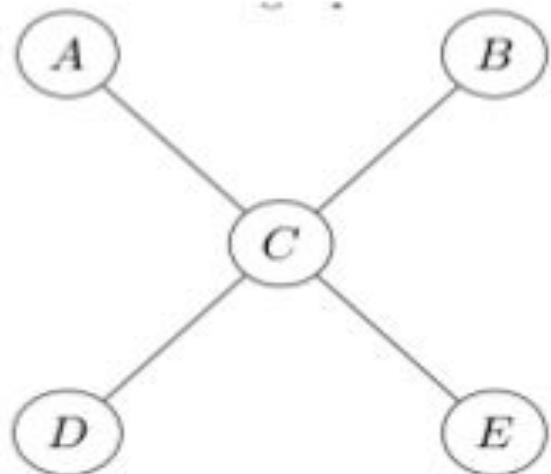


(b)

- (c) It cannot be captured by a graphical model.
- (d) Graphical model can caputre the assumption, but the given models don't do it.

Question 8

Consider the Markov network shown below in the Figure.



Two students - Student A and Student B make the following claims:

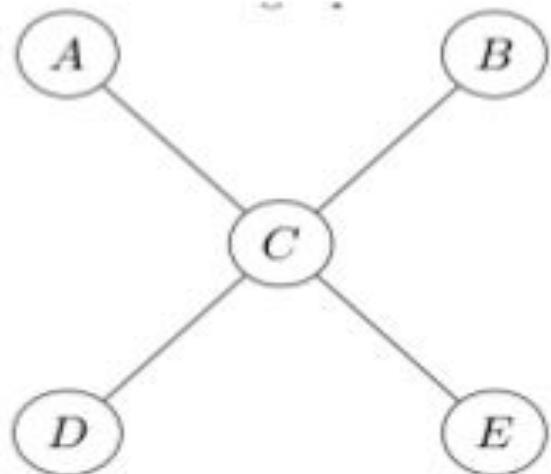
- Student A claims $P(A|B)=P(A)$
- Student B claims $P(A|B,C)=P(A|C)$

Which of the following is true?

- a. Student A and Student B are correct.
- b. Student A is incorrect and Student B is correct.
- c. Student A is correct and Student B is incorrect.
- d. Both are incorrect.
- e. Insufficient information to make any conclusion. Probability distributions of each variable should be given.

Question 8

Consider the Markov network shown below in the Figure.



Two students - Student A and Student B make the following claims:

- Student A claims $P(A|B)=P(A)$
- Student B claims $P(A|\{B,C\})=P(A|C)$

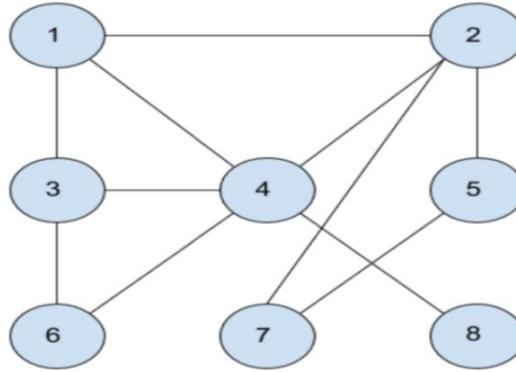
Which of the following is true?

- a. Student A and Student B are correct.
- b. Student A is incorrect and Student B is correct.
- c. Student A is correct and Student B is incorrect.
- d. Both are incorrect.
- e. Insufficient information to make any conclusion. Probability distributions of each variable should be given.

Answer: b.

Question 9

Consider the Markov network shown below in Figure 4

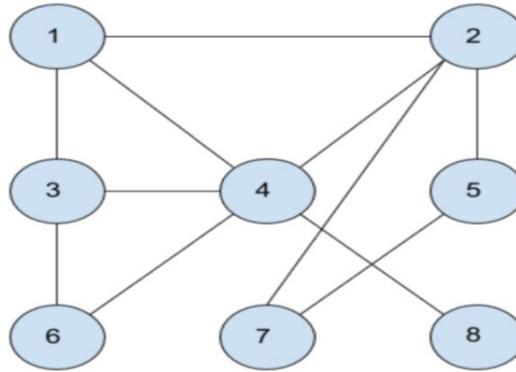


Which of the following variables are NOT in the markov blanket of variable “4” shown in the above Figure 4 ? (multiple answers may be correct)

- (a) 1
- (b) 8
- (c) 2
- (d) 5
- (e) 6
- (f) 4
- (g) 7

Question 9

Consider the Markov network shown below in Figure 4



Which of the following variables are NOT in the markov blanket of variable “4” shown in the above Figure 4 ? (multiple answers may be correct)

- (a) 1
- (b) 8
- (c) 2
- (d) 5
- (e) 6
- (f) 4
- (g) 7

Sol. (d) and (g)

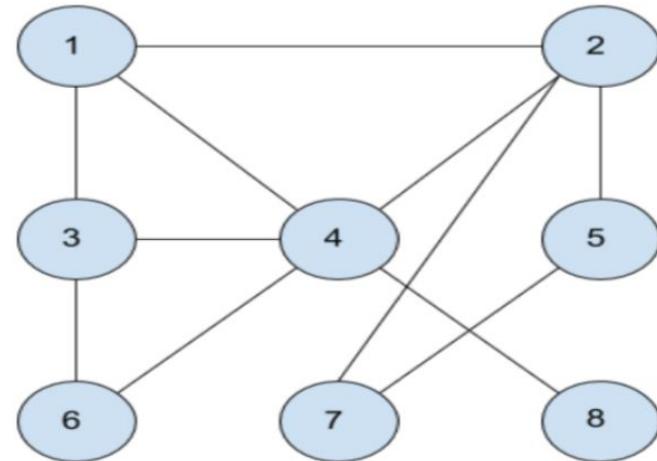
Question 10

In the Markov network given in the figure, two students make the following claims

- Manish claims variable “1” is dependent on variable “7” given variable “2”.
- Trina claims variable “2” is independent of variable “6” given variable “3”.

Which of the following is true?

- (a) Both the students are correct.
- (b) Trina is incorrect and Manish is correct.
- (c) Trina is correct and Manish is incorrect.
- (d) Both the students are incorrect.
- (e) Insufficient information to make any conclusion.



Question 10

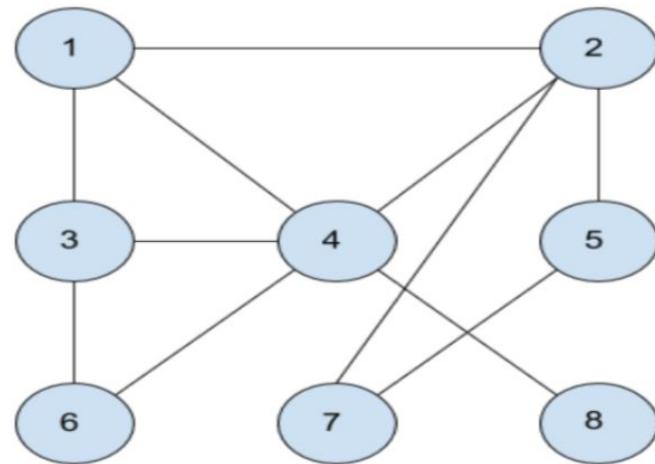
In the Markov network given in the figure, two students make the following claims

- Manish claims variable “1” is dependent on variable “7” given variable “2”.
- Trina claims variable “2” is independent of variable “6” given variable “3”.

Which of the following is true?

- (a) Both the students are correct.
- (b) Trina is incorrect and Manish is correct.
- (c) Trina is correct and Manish is incorrect.
- (d) Both the students are incorrect.
- (e) Insufficient information to make any conclusion.

Sol. (d)



THE END

NPTEL Live Session

Week 10

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

02/04/24

Question 1

Suppose you have a single cluster of data points. The data points are $(-2,-2), (-1,-2), (2,1), (1,2)$. Find the data point x which has the highest average L_2 distance with respect to other data points.

- (a) $(-2,-2)$
- (b) $(-1,-2)$
- (c) $(2,1)$
- (d) $(1,2)$

Question 1

Suppose you have a single cluster of data points. The data points are $(-2,-2), (-1,-2), (2,1), (1,2)$. Find the data point x which has the highest average L_2 distance with respect to other data points.

- (a) $(-2,-2)$
- (b) $(-1,-2)$
- (c) $(2,1)$
- (d) $(1,2)$

Sol. (a)

Question 2

Which of the following statements is/are **not true** about k -means clustering?

- (a) It is an unsupervised learning algorithm
- (b) Overlapping of clusters is allowed in k -means clustering
- (c) It is a hard-clustering technique
- (d) k is a hyperparameter

Question 2

Which of the following statements is/are **not true** about k -means clustering?

- (a) It is an unsupervised learning algorithm
- (b) Overlapping of clusters is allowed in k -means clustering
- (c) It is a hard-clustering technique
- (d) k is a hyperparameter

Sol. (b)

Refer to the lecture

Question 3

Which among the following is/are some of the assumptions made by the k-means algorithm (assuming Euclidean distance measure)?

- (a) Clusters are spherical in shape
- (b) Clusters are of similar sizes
- (c) Data points in one cluster are well separated from data points of other clusters
- (d) There is no wide variation in density among the data points

Question 3

Which among the following is/are some of the assumptions made by the k-means algorithm (assuming Euclidean distance measure)?

- (a) **Clusters are spherical in shape**
- (b) **Clusters are of similar sizes**
- (c) Data points in one cluster are well separated from data points of other clusters
- (d) There is no wide variation in density among the data points

Sol. (a) & (b)

The Euclidean distance measure ensures that areas around a cluster centroid comprising points closest to that centroid (which is a cluster) is spherical in shape. Also, this particular distance measure prevents arbitrarily sized clusters since this typically violates the clustering criterion.

Question 4

Considering single-link and complete-link hierarchical clustering, is it possible for a point to be closer to points in other clusters than to points in its own cluster? If so, in which approach will this tend to be observed?

- (a) No
- (b) Yes, single-link clustering
- (c) Yes, complete-link clustering
- (d) Yes, both single-link and complete-link clustering

Question 4

Considering single-link and complete-link hierarchical clustering, is it possible for a point to be closer to points in other clusters than to points in its own cluster? If so, in which approach will this tend to be observed?

- (a) No
- (b) Yes, single-link clustering
- (c) Yes, complete-link clustering
- (d) Yes, both single-link and complete-link clustering

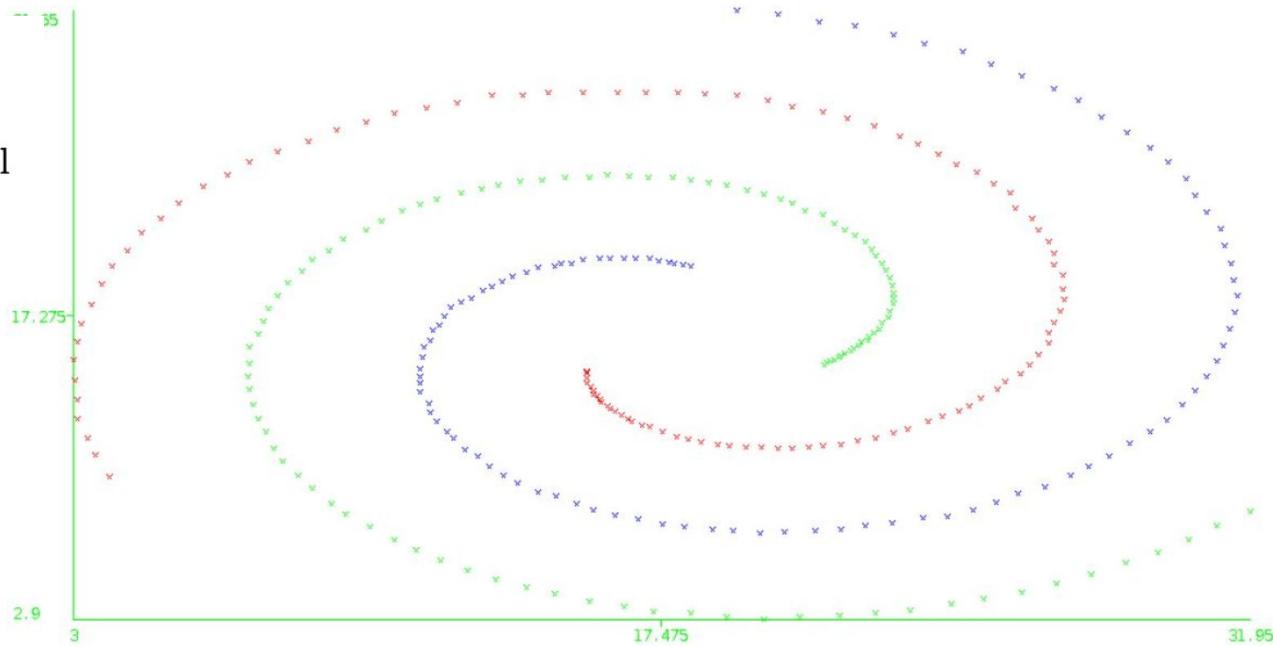
Sol. (d)

This is possible in both single-link and complete-link clustering. In the single-link case, an example would be two parallel chains where many points are closer to points in the other chain/cluster than to points in their own cluster. In the complete-link case, this notion is more intuitive due to the clustering constraint (measuring distance between two clusters by the distance between their farthest points).

Question 5

Consider the following image showing data points belonging to three different clusters (indicated by the colours of the points). Which among the following clustering algorithms will perform well in accurately clustering the given data?

- (a) K-means
- (b) Single-link hierarchical
- (c) Complete-link hierarchical
- (d) DBSCAN



Question 5

Consider the following image showing data points belonging to three different clusters (indicated by the colours of the points). Which among the following clustering algorithms will perform well in accurately clustering the given data?

- (a) K-means
- (b) Single-link hierarchical
- (c) Complete-link hierarchical
- (d) DBSCAN

Sol. (b) & (d)

K-means will clearly not work for this data set. Complete-link clustering will also not be able to recover the desired clustering since in the spiral structure in which the data points lie, points in the same cluster are actually quite far from other points in their own clusters. Single-link clustering is ideally suited for this data set as well as DBSCAN, since there is enough distance between points belonging to the different clusters.

Question 6

Suppose while performing DBSCAN we randomly choose a point which has less than MinPts number of points in its neighbourhood. Which among the following is true for such a point?

- (a) It is treated as noise, and not considered further in the algorithm
- (b) It becomes part of its own cluster
- (c) Depending upon other points, it may later turn out to be a core point
- (d) Depending upon other points, it may be density connected to other points

Question 6

Suppose while performing DBSCAN we randomly choose a point which has less than MinPts number of points in its neighbourhood. Which among the following is true for such a point?

- (a) It is treated as noise, and not considered further in the algorithm
- (b) It becomes part of its own cluster
- (c) Depending upon other points, it may later turn out to be a core point
- (d) Depending upon other points, it may be density connected to other points

Sol. (d)

Since there are less than MinPts number of points in its neighbourhood, it is not a core point. However, it is not necessarily a noise point, since if there exists a core point in whose neighbourhood this point lies in, it can be a boundary point.

Question 7

Which of the following can act as possible termination conditions in K-Means?

- a. For a fixed number of iterations.
- b. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
- c. Centroids do not change between successive iterations,
- d. Terminate when RSS falls below a threshold

Question 7

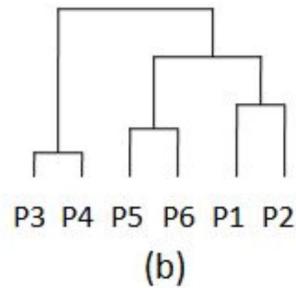
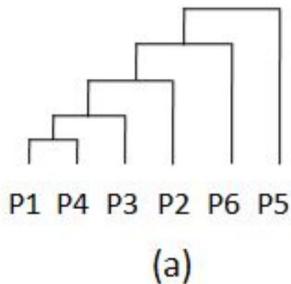
Which of the following can act as possible termination conditions in K-Means?

- a. For a fixed number of iterations.
- b. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
- c. Centroids do not change between successive iterations,
- d. Terminate when RSS falls below a threshold

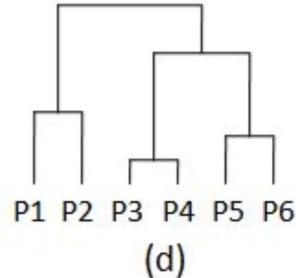
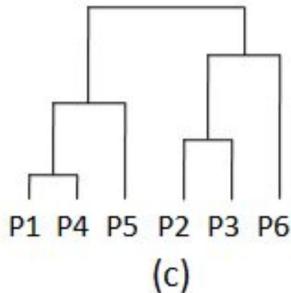
Answer : a,b,c,d

Question 8

Consider the similarity matrix given below: Which of the following shows the hierarchy of clusters created by the single link clustering algorithm.

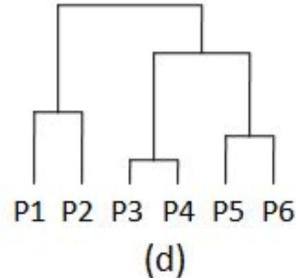
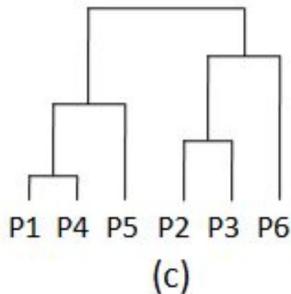
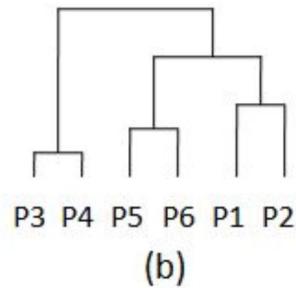
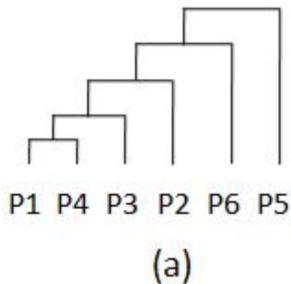


	P1	P2	P3	P4	P5	P6
P1	1.0000	0.7895	0.1579	0.0100	0.5292	0.3542
P2	0.7895	1.0000	0.3684	0.2105	0.7023	0.5480
P3	0.1579	0.3684	1.0000	0.8421	0.5292	0.6870
P4	0.0100	0.2105	0.8421	1.0000	0.3840	0.5573
P5	0.5292	0.7023	0.5292	0.3840	1.0000	0.8105
P6	0.3542	0.5480	0.6870	0.5573	0.8105	1.0000



Question 8

Consider the similarity matrix given below: Which of the following shows the hierarchy of clusters created by the single link clustering algorithm.



	P1	P2	P3	P4	P5	P6
P1	1.0000	0.7895	0.1579	0.0100	0.5292	0.3542
P2	0.7895	1.0000	0.3684	0.2105	0.7023	0.5480
P3	0.1579	0.3684	1.0000	0.8421	0.5292	0.6870
P4	0.0100	0.2105	0.8421	1.0000	0.3840	0.5573
P5	0.5292	0.7023	0.5292	0.3840	1.0000	0.8105
P6	0.3542	0.5480	0.6870	0.5573	0.8105	1.0000

Answer: (b)

Question 9

In which of the following cases will K-Means clustering fail to give good results?

- a. Data points with outliers
- b. Data points with different densities
- c. Data points with round shapes
- d. Data points with non-convex shapes

Question 9

In which of the following cases will K-Means clustering fail to give good results?

- a. Data points with outliers
- b. Data points with different densities
- c. Data points with round shapes
- d. Data points with non-convex shapes

Answer : a,b,d

Question 10

Consider the following 3 points in a cluster : (2,1), (3,4), (5,7)

Computer the BIRCH Cluster Features (CF) for this cluster.

- a. (4,5,15)
- b. (5,13,200)
- c. (3,(6,6),125)
- d. (3,(10,12),104)

Question 10

Consider the following 3 points in a cluster : (2,1), (3,4), (5,7)

Computer the BIRCH Cluster Features (CF) for this cluster.

- a. (4,5,15)
- b. (5,13,200)
- c. (3,(6,6),125)
- d. **(3,(10,12),104)**

Answer: d

THE END

NPTEL Live Session

Week 11

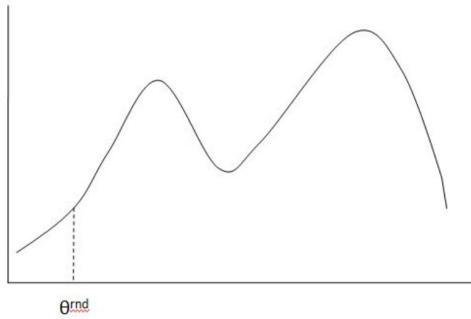
Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

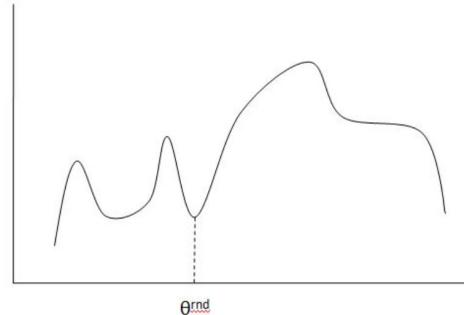
09/04/24

Question 1

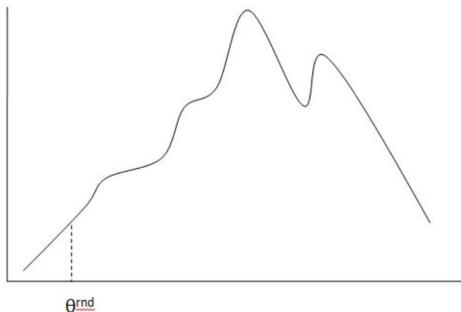
In each of the options the diagram contains a log likelihood function for a particular problem as well as an initial value of the parameter used in the execution of the EM algorithm. In which of the given scenarios will the EM algorithm be able to achieve the global maximum?



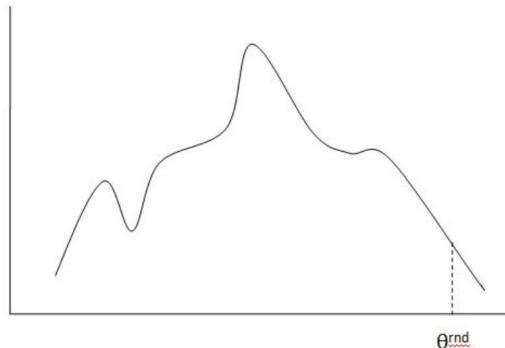
Q1 (a)



Q1 (b)



Q1 (c)



Q1 (d)

Question 1

Sol. (c)

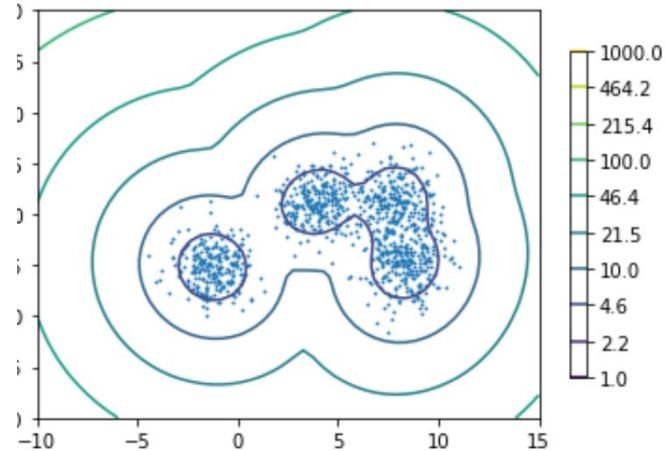
In each scenario other than (c), the EM algorithm will get stuck in a local minima.

Question 2

For the given GMM model in the figure:

What is value of k ?

- (a) 3
- (b) 4
- (c) 5
- (d) 6

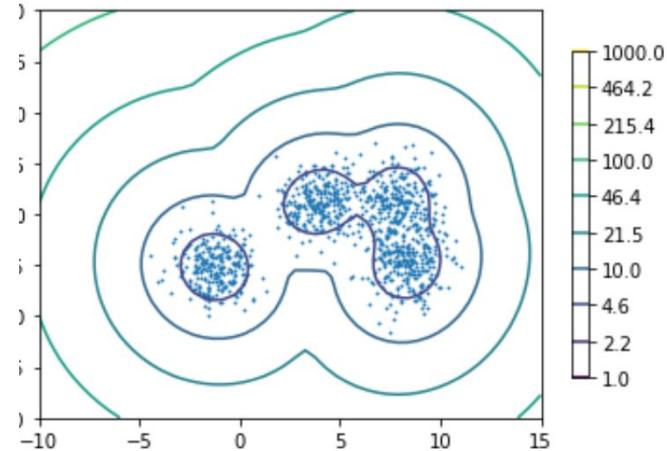


Question 2

For the given GMM model in the figure:

What is value of k ?

- (a) 3
- (b) 4
- (c) 5
- (d) 6



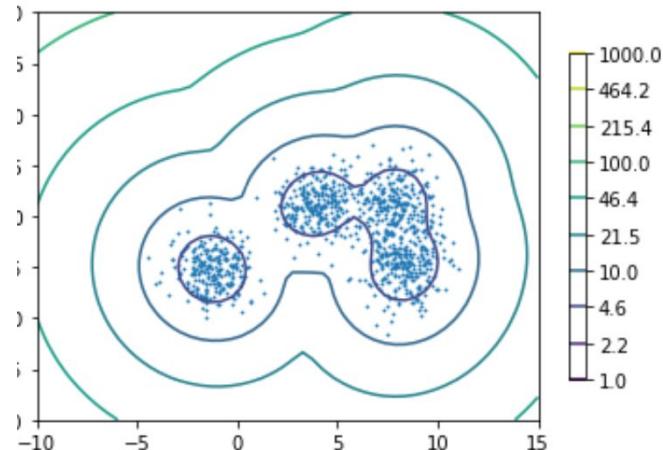
Sol. (b)

Equal to the number of clusters you can see in the figure.

Question 3

What is the minimum value of $k' \neq k$, where k is from previous question, for which you will get a very similar density estimation?

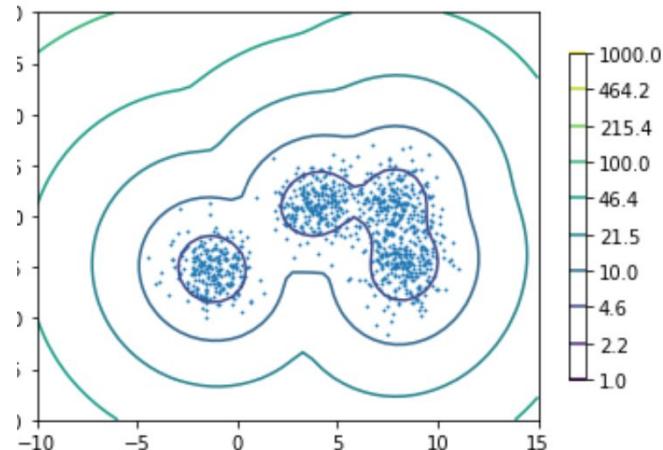
- (a) 3
- (b) 4
- (c) 5
- (d) 6



Question 3

What is the minimum value of $k' \neq k$, where k is from previous question, for which you will get a very similar density estimation?

- (a) 3
- (b) 4
- (c) 5
- (d) 6



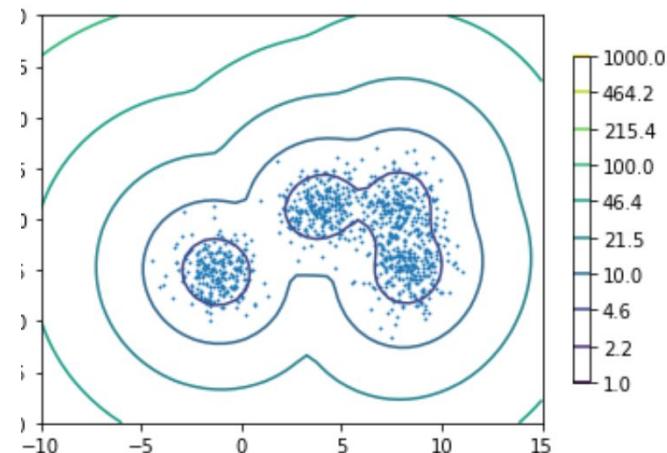
Sol. (a)

Should be clear from the image.

Question 4

Assume equal π_i for each gaussian model after convergence as in Q2. What would (approximately) be π_i 's for the model you'll get with k' as in Q3?

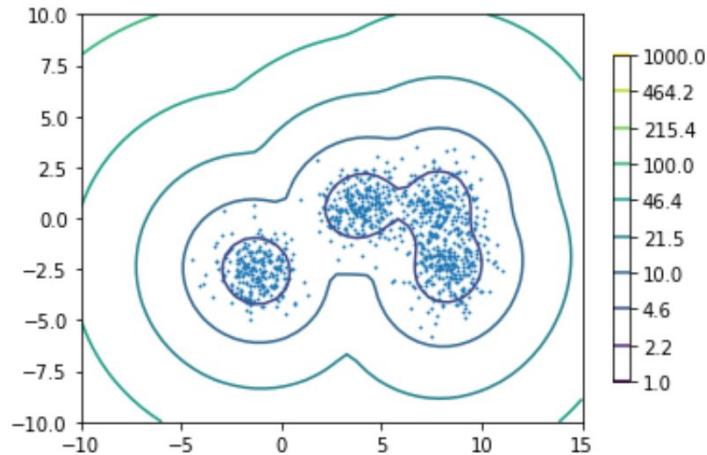
- (a) [0.33, 0.33, 0.17, 0.17]
- (b) [0.2, 0.2, 0.6]
- (c) [0.25, 0.25, 0.5]
- (d) [0.2, 0.2, 0.2, 0.4]



Question 4

Assume equal π_i for each gaussian model after convergence as in Q2. What would (approximately) be π_i 's for the model you'll get with k' as in Q3?

- (a) [0.33, 0.33, 0.17, 0.17]
- (b) [0.2, 0.2, 0.6]
- (c) [0.25, 0.25, 0.5]
- (d) [0.2, 0.2, 0.2, 0.4]



Sol. (c)

For $k = 4$, π_i 's = [0.25, 0.25, 0.25, 0.25]

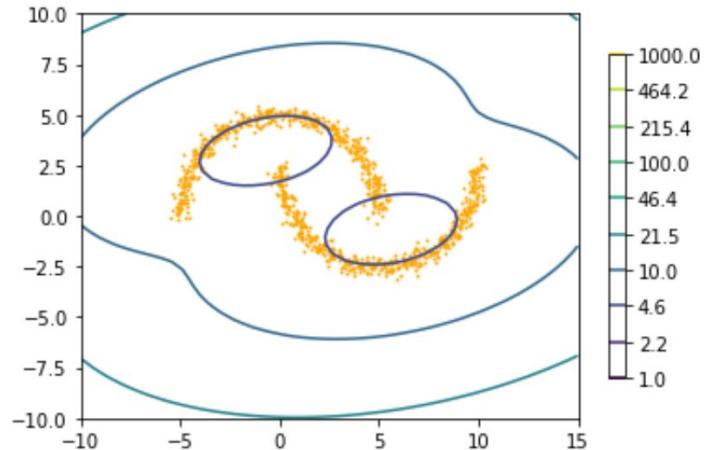
For $k = 3$, two of the clusters are merged $\implies \pi_i$'s = [0.25, 0.25, 0.5]

Question 5

For a set of points (given in orange), the density estimation of a GMM is given below :

What is the problem evident in the image?

- (a) π_i 's are too big
- (b) The clusters are not sampled from a gaussian distribution.
- (c) The GMM has not converged yet.
- (d) There is no problem

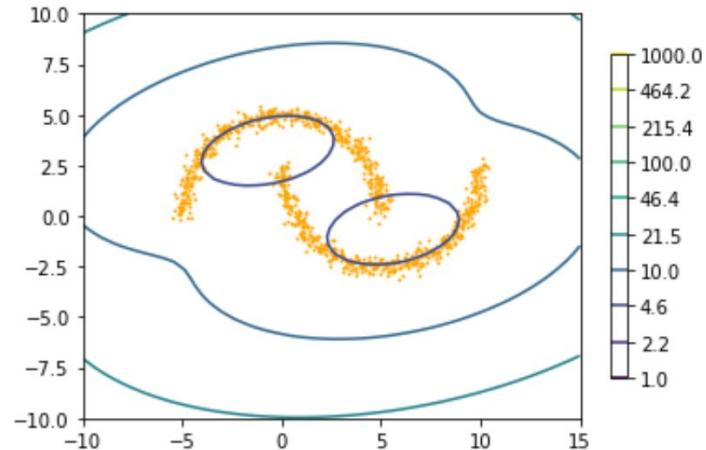


Question 5

For a set of points (given in orange), the density estimation of a GMM is given below :

What is the problem evident in the image?

- (a) π_i 's are too big
- (b) The clusters are not sampled from a gaussian distribution.
- (c) The GMM has not converged yet.
- (d) There is no problem



Sol. (b)

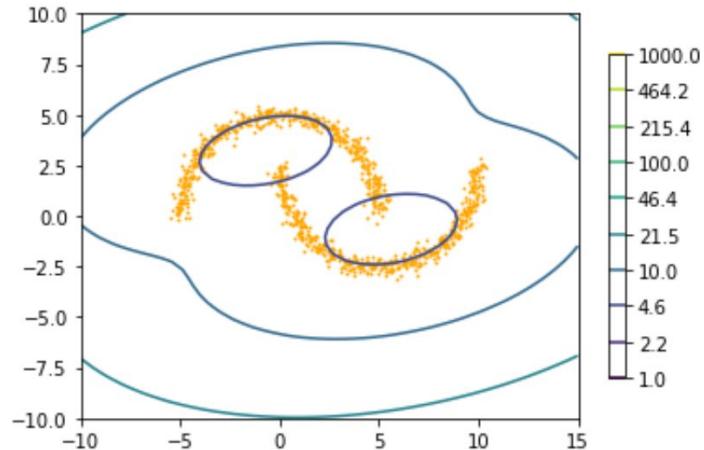
We would have gotten a nice density estimate if the clusters were sampled from a gaussian.

Question 6

For a set of points (given in orange), the density estimation of a GMM is given below :

What can be done to get a better fit?

- (a) Increase k
- (b) Use a better initialisation
- (c) Learn for more iterations
- (d) There is no problem

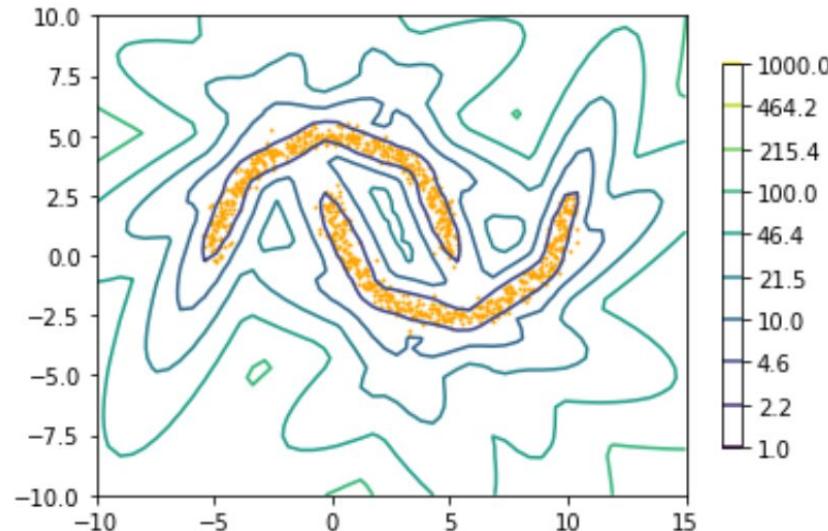


Question 6

Sol. (a)

You can increase k to allow the model to fit a larger number of gaussians to approximate a clearly non-gaussian distribution.

An example is shown below with $k=8$.



Question 7

Given n samples x_1, x_2, \dots, x_N drawn independently from an Exponential distribution unknown parameter λ , find the MLE of λ .

- (a) $\lambda_{MLE} = \sum_{i=1}^n x_i$
- (b) $\lambda_{MLE} = n \sum_{i=1}^n x_i$
- (c) $\lambda_{MLE} = \frac{n}{\sum_{i=1}^n x_i}$
- (d) $\lambda_{MLE} = \frac{\sum_{i=1}^n x_i}{n}$
- (e) $\lambda_{MLE} = \frac{n-1}{\sum_{i=1}^n x_i}$
- (f) $\lambda_{MLE} = \frac{\sum_{i=1}^n x_i}{n-1}$

Question 7

Sol. (c)

$$\mathcal{L}(\lambda, x_1, \dots, x_n) = \prod_{i=1}^n f(x_i, \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

$$\begin{aligned}\frac{d \ln (\mathcal{L}(\lambda, x_1, \dots, x_n))}{d \lambda} &= \frac{d \ln (\lambda^n e^{-\lambda \sum_{i=1}^n x_i})}{d \lambda} \\&= \frac{d \ln (n \ln(\lambda) - \lambda \sum_{i=1}^n x_i)}{d \lambda} \\&= \frac{n}{\lambda} - \sum_{i=1}^n x_i\end{aligned}$$

Set the above term to zero to obtain MLE of λ

$$\lambda = \frac{n}{\sum_{i=1}^n x_i}$$

Question 8

Given n samples x_1, x_2, \dots, x_N drawn independently from a Poisson distribution unknown parameter λ , find the MLE of λ .

(a) $\lambda_{MLE} = \sum_{i=1}^n x_i$

(b) $\lambda_{MLE} = n \sum_{i=1}^n x_i$

(c) $\lambda_{MLE} = \frac{n}{\sum_{i=1}^n x_i}$

(d) $\lambda_{MLE} = \frac{\sum_{i=1}^n x_i}{n}$

(e) $\lambda_{MLE} = \frac{n-1}{\sum_{i=1}^n x_i}$

(f) $\lambda_{MLE} = \frac{\sum_{i=1}^n x_i}{n-1}$

Question 8

Sol. (d)

Write the likelihood:

$$l(\lambda; x) = \prod_i e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}}{x_1! x_2! \cdots x_n!} \lambda^{x_1+x_2+\cdots+x_n}$$

Take the log and differentiate the log-likelihood with respect to λ and set it to 0.

Question 9

Suppose we have a Gaussian mixture model of 3-dimensional data with 4 components, and we use a model with full covariance matrices. How many parameters are in the model?

- a. 30
- b. 40
- c. 50
- d. 60

Question 9

Suppose we have a Gaussian mixture model of 3-dimensional data with 4 Gaussians, and we use a model with full covariance matrices. How many parameters are in the model?

- a. 30
- b. 40
- c. 50
- d. 60

Answer : b. 40 : Number of parameters: - # of parameters of pi per gaussian = 1 - # of parameters of mu per gaussian = V = # dimensions - # of parameters of Sigma per gaussian for full Covariance matrix = $V*(V+1)/2$ or - # of parameters of Sigma per gaussian for just diagonal matrix = V

For each gaussian: - # of parameters of pi: 1 - # of parameters of mu: 3 - # of parameters of Sigma for full Covariance matrix = $3(3+1)/2 = 6$

Total # of parameters for each gaussians : $1 + 3 + 6 = 10$

Total # of parameters: $10 * \text{num of gaussians} = 40$

Question 10

Which of the following correctly describes the differences between EM for mixtures of Gaussians and k-means? Choose all that apply.

- a. k-means often gets stuck in a local minimum, while EM tends not to
- b. EM is better at capturing clusters of different sizes and orientations
- c. EM is better at capturing clusters with overlaps
- d. EM is less prone to overfitting than k-means

Question 10

Which of the following correctly describes the differences between EM for mixtures of GMM and k-means? Choose all that apply.

- a. k-means often gets stuck in a local minimum, while EM tends not to
- b. EM is better at capturing clusters of different sizes and orientations
- c. EM is better at capturing clusters with overlaps
- d. EM is less prone to overfitting than k-means

Answer : b,c

THE END

NPTEL Live Session

Week 12

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

16/04/24

Question 1

Generalization error is given by:

$$\mathcal{E}(h) = P_{(x,y) \sim D}(h(x) \neq y)$$

What is D ?

- (a) The underlying distribution we are trying to estimate
- (b) A sample of the underlying distribution
- (c) A different distribution that we are using to estimate the true underlying distribution
- (d) The universe of distributions

Question 1

Generalization error is given by:

$$\mathcal{E}(h) = P_{(x,y) \sim D}(h(x) \neq y)$$

What is D ?

- (a) The underlying distribution we are trying to estimate
- (b) A sample of the underlying distribution
- (c) A different distribution that we are using to estimate the true underlying distribution
- (d) The universe of distributions

Sol. (a)

Refer to lecture

Question 2

The hypothesis function obtained by empirical risk minimization is denoted as:

$$\hat{h} = \arg \min_{h \in \mathcal{H}} \mathcal{E}(h)$$

If \mathcal{H} is a class of neural networks, which aspects constitute its space? (Select all that apply)

- (a) Different number of layers
- (b) Different number of neurons per layer
- (c) Different values for weights
- (d) Different inputs

Question 2

The hypothesis function obtained by empirical risk minimization is denoted as:

$$\hat{h} = \arg \min_{h \in \mathcal{H}} \mathcal{E}(h)$$

If \mathcal{H} is a class of neural networks, which aspects constitute its space? (Select all that apply)

- (a) Different number of layers
- (b) Different number of neurons per layer
- (c) Different values for weights
- (d) Different inputs

Sol. (c)

Question 3

As discussed in the lecture, most of the classifiers minimize the empirical risk. Which among the following is an exceptional case?

- (a) Perceptron learning algorithm
- (b) Artificial Neural Network
- (c) Support Vector Machines
- (d) both (a) and (b)
- (e) None of the above

Question 3

As discussed in the lecture, most of the classifiers minimize the empirical risk. Which among the following is an exceptional case?

- (a) Perceptron learning algorithm
- (b) Artificial Neural Network
- (c) Support Vector Machines
- (d) both (a) and (b)
- (e) None of the above

Sol. (c)

In addition to minimizing empirical risk, the SVMs also maximize the margin. This is known as structural risk minimization.

Question 4

What happens when your model complexity (such as interaction terms in linear regression, order of polynomial in SVM, etc.) increases?

- (a) Model Bias increases
- (b) Model Bias decreases
- (c) Variance of the model increases
- (d) Variance of the model decreases

Question 4

What happens when your model complexity (such as interaction terms in linear regression, order of polynomial in SVM, etc.) increases?

- (a) Model Bias increases
- (b) Model Bias decreases
- (c) Variance of the model increases
- (d) Variance of the model decreases

Sol. (b) and (c)

Question 5

Aliens challenge you to a complex game that no human has seen before. They give you time to learn the game and develop strategies before the final showdown. You choose to use machine learning because an intelligent machine is your only hope. Which machine learning paradigm should you choose for this?

- (a) Supervised learning
- (b) Unsupervised learning
- (c) Reinforcement learning
- (d) Use a random number generator and hope for the best

Question 5

Aliens challenge you to a complex game that no human has seen before. They give you time to learn the game and develop strategies before the final showdown. You choose to use machine learning because an intelligent machine is your only hope. Which machine learning paradigm should you choose for this?

- (a) Supervised learning
- (b) Unsupervised learning
- (c) Reinforcement learning
- (d) Use a random number generator and hope for the best

Sol. (c)

Reinforcement learning is the optimal for building agents for complex games where no expert trajectories exist.

Question 6

In the Pavlov's dog experiment, what is the input signal to the dog and what is the reward?

- (a) Bell, digestion
- (b) Food, praise
- (c) Food, digestion
- (d) Bell, food

Question 6

In the Pavlov's dog experiment, what is the input signal to the dog and what is the reward?

- (a) Bell, digestion
- (b) Food, praise
- (c) Food, digestion
- (d) Bell, food

Sol. (d)

Refer to lecture.

Question 7

Which of the following measure best analyze the performance of a classifier?

- (a) Precision
- (b) Recall
- (c) Accuracy
- (d) Time complexity
- (e) Depends on the application

Question 7

Which of the following measure best analyze the performance of a classifier?

- (a) Precision
- (b) Recall
- (c) Accuracy
- (d) Time complexity
- (e) Depends on the application

Sol. (e)

Different applications might need to optimize different performance measures. Applications of machine learning span over playing games to very critical domains(such as health and security). Measures like accuracy for instance cannot be reliable when we have a dataset with significant class imbalance. So there cannot be a single measure to analyze the effectiveness of a classifier in all environments.

Question 8

Which type of feedback does an agent in reinforcement learning receive?

- A. Predictions
- B. Labels
- C. Clusters
- D. Rewards or penalties

Question 8

Which type of feedback does an agent in reinforcement learning receive?

- A. Predictions
- B. Labels
- C. Clusters
- D. Rewards or penalties

Answer : D

Question 9

You are designing a Reinforcement Learning agent for a racing game. Among the following reward schemes, which one leads to the best performance of the agent?

- a. +5 for reaching the finish line, -1 for going off the road
- b. +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line
- c. +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line, +1 for the agent going off the road.
- d. -5 for reaching the finish line, +0.1 for every second that passes before the agent reaches the finish line.

Question 9

You are designing a Reinforcement Learning agent for a racing game. Among the following reward schemes, which one leads to the best performance of the agent?

- a. +5 for reaching the finish line, -1 for going off the road
- b. +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line
- c. +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line, +1 for the agent going off the road.
- d. -5 for reaching the finish line, +0.1 for every second that passes before the agent reaches the finish line.

Answer : b

Question 10

You want to create a self-driving car software. In the context of the standard Reinforcement Learning framework, what would you classify as the **state** and **actions**? Note that your system does not have access to previous states.

- a. **State**-(Current steering wheel position, Current pedal positions, Current speed) **Actions**-(Turn the steering wheel, Press pedals)
- b. **State**-(Current steering wheel position, Current pedal positions, Current acceleration) **Actions**-(Turn the steering wheel, Press pedals)
- c. **State**-(Current steering wheel position, Current pedal positions, Current speed) **Actions**-(Change direction, Change speed)
- d. **State**-(Current steering wheel position, Current pedal positions, Current acceleration)
Actions-(Change direction, Change speed)

Question 10

You want to create a self-driving car software. In the context of the standard Reinforcement Learning framework, what would you classify as the **state** and **actions**? Note that your system does not have access to previous states.

- a. **State**-(Current steering wheel position, Current pedal positions, Current speed) **Actions**-(Turn the steering wheel, Press pedals)
- b. **State**-(Current steering wheel position, Current pedal positions, Current acceleration) **Actions**-(Turn the steering wheel, Press pedals)
- c. **State**-(Current steering wheel position, Current pedal positions, Current speed) **Actions**-(Change direction, Change speed)
- d. **State**-(Current steering wheel position, Current pedal positions, Current acceleration)
Actions-(Change direction, Change speed)

Answer : a

Question 11

Which of these classifiers do not require any additional modifications to use them when we have more than 2 classes? (Note that more than one answers could be correct)

- a. Decision Trees
- b. Linear Classifier
- c. Support Vector Machines (SVMs)
- d. k Nearest Neighbors (kNN)

Question 11

Which of these classifiers do not require any additional modifications to use them when we have more than 2 classes? (Note that more than one answers could be correct)

- a. Decision Trees
- b. Linear Classifier
- c. Support Vector Machines (SVMs)
- d. k Nearest Neighbors (kNN)

Answer: a,d

THE END

NPTEL Live Session 13

Extra Session

Introduction to Machine Learning

Ayan Maity
PhD Scholar,
CSE,IIT Kharagpur

18/04/24

Question 1

Suppose we want to add a regularizer to the linear regression loss function, to control the magnitudes of the weights β . We have a choice between $\Omega_1(\beta) = \sum_{i=1}^p |\beta_i|$ and $\Omega_2(\beta) = \sum_{i=1}^p \beta_i^2$. Which one is more likely to result in sparse weights?

- (a) Ω_1
- (b) Ω_2
- (c) Both Ω_1 and Ω_2 will result in sparse weights
- (d) Neither of Ω_1 or Ω_2 can result in sparse weights

Question 1

Suppose we want to add a regularizer to the linear regression loss function, to control the magnitudes of the weights β . We have a choice between $\Omega_1(\beta) = \sum_{i=1}^p |\beta_i|$ and $\Omega_2(\beta) = \sum_{i=1}^p \beta_i^2$. Which one is more likely to result in sparse weights?

- (a) Ω_1
- (b) Ω_2
- (c) Both Ω_1 and Ω_2 will result in sparse weights
- (d) Neither of Ω_1 or Ω_2 can result in sparse weights

Sol. (a)

Question 2

Which of the following statements are true:

- (a) The chances of overfitting decreases with increasing the number of hidden nodes and increasing the number of hidden layers.
- (b) A neural network with one hidden layer can represent any Boolean function given sufficient number of hidden units and appropriate activation functions.
- (c) Two hidden layer neural networks can represent any continuous functions (within a tolerance) as long as the number of hidden units is sufficient and appropriate activation functions used.

Question 2

Which of the following statements are true:

- (a) The chances of overfitting decreases with increasing the number of hidden nodes and increasing the number of hidden layers.
- (b) A neural network with one hidden layer can represent any Boolean function given sufficient number of hidden units and appropriate activation functions.
- (c) Two hidden layer neural networks can represent any continuous functions (within a tolerance) as long as the number of hidden units is sufficient and appropriate activation functions used.

Sol. (b), (c)

By increasing the number of hidden nodes or hidden layers we are increasing the number of parameters. Increased set of parameters is more capable to memorize the training data. Hence it may result in overfitting.

Question 3

For the given confusion matrix, compute the recall

	True Positive	True Negative
Predicted Positive	6	4
Predicted Negative	3	7

- (a) 0.73
- (b) 0.7
- (c) 0.6
- (d) 0.67
- (e) 0.78
- (f) None of the above

Question 3

For the given confusion matrix, compute the recall

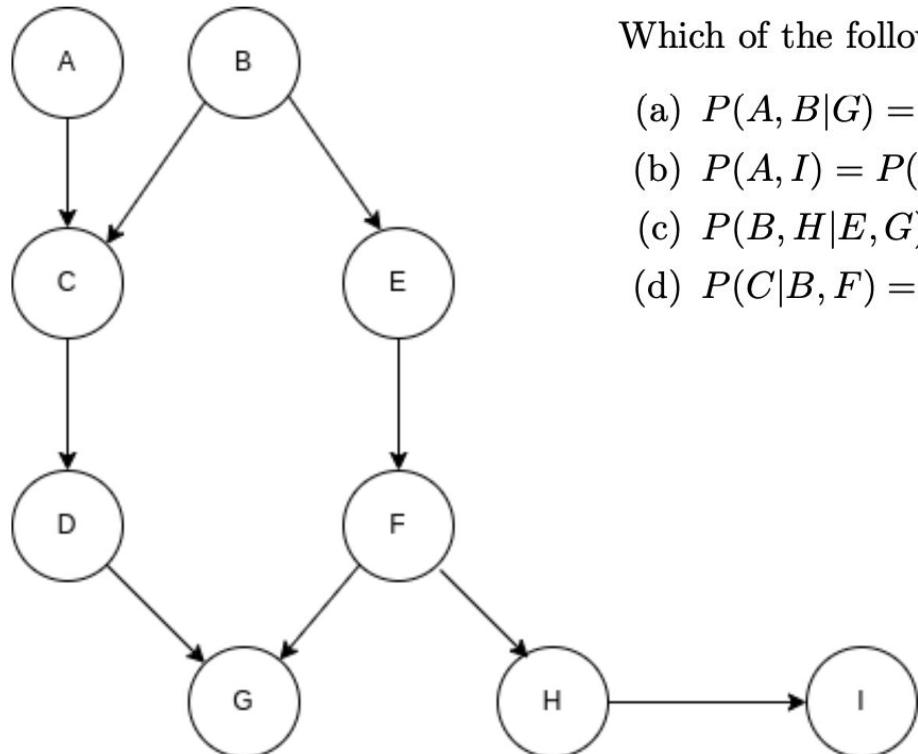
	True Positive	True Negative
Predicted Positive	6	4
Predicted Negative	3	7

- (a) 0.73
- (b) 0.7
- (c) 0.6
- (d) 0.67
- (e) 0.78
- (f) None of the above

Sol. (d)

Question 4

The figure below shows a Bayesian Network with 9 variables, all of which are binary.

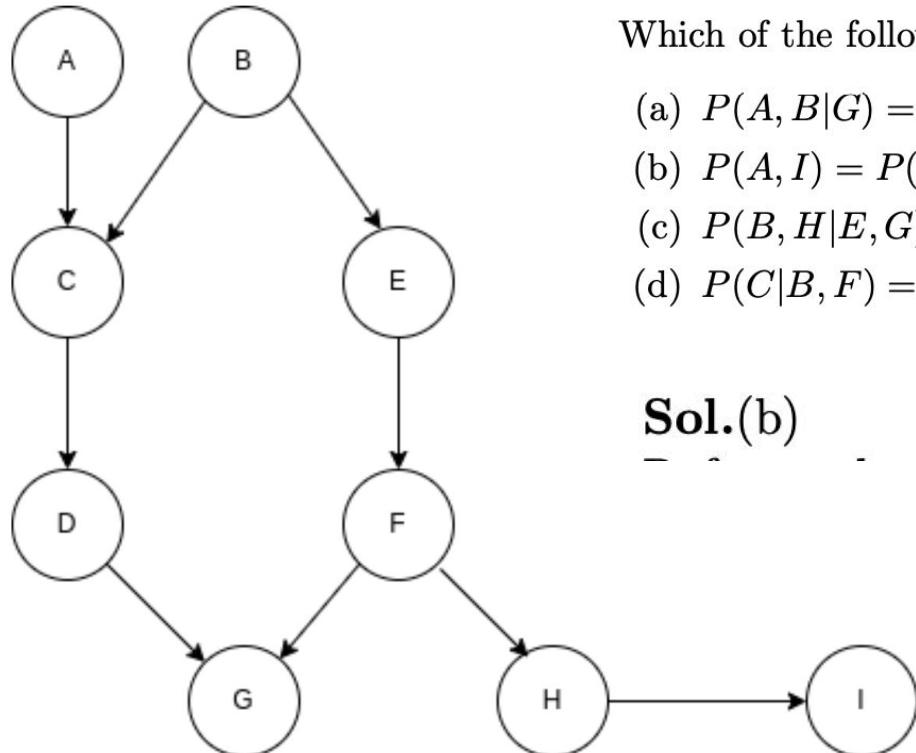


Which of the following is/are always true for the above Bayesian Network?

- (a) $P(A, B|G) = P(A|G)P(B|G)$
- (b) $P(A, I) = P(A)P(I)$
- (c) $P(B, H|E, G) = P(B|E, G)P(H|E, G)$
- (d) $P(C|B, F) = P(C|F)$

Question 4

The figure below shows a Bayesian Network with 9 variables, all of which are binary.



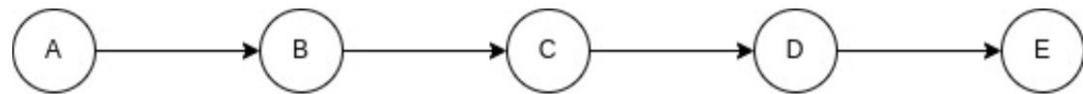
Which of the following is/are always true for the above Bayesian Network?

- (a) $P(A, B|G) = P(A|G)P(B|G)$
- (b) $P(A, I) = P(A)P(I)$
- (c) $P(B, H|E, G) = P(B|E, G)P(H|E, G)$
- (d) $P(C|B, F) = P(C|F)$

Sol.(b)

Question 5

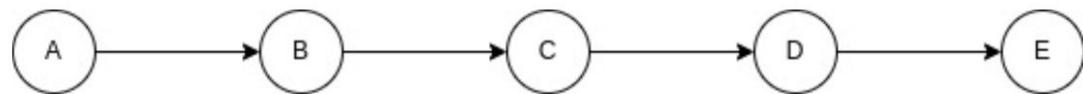
For the given graphical model, what is the optimal variable elimination order when trying to calculate $P(E=e)$?



- (a) A,B,C,D
- (b) D,C,B,A
- (c) A,D,B,C
- (d) D,A,C,A

Question 5

For the given graphical model, what is the optimal variable elimination order when trying to calculate $P(E=e)$?



(a) A,B,C,D

Sol. (a)

(b) D,C,B,A

(c) A,D,B,C

(d) D,A,C,A

$$P(E = e) = \sum_d \sum_c \sum_b \sum_a P(a, b, c, d, e)$$

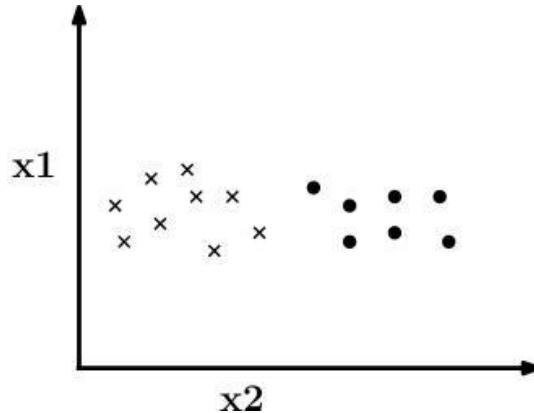
$$P(E = e) = \sum_d \sum_c \sum_b \sum_a P(a)P(b|a)P(c|b)P(d|c)P(e|d)$$

$$P(E = e) = \sum_d P(e|d) \sum_c P(d|c) \sum_b P(c|b) \sum_a P(a)P(b|a)$$

Question 6

The data points shown in the figure below have two features (x_1 and x_2) and each data point belongs to one of the two classes shown in the figure by cross and circle. If you have to select one feature to use in a classification problem, which feature will you choose?

- a. x_1
- b. x_2
- c. Any one of the above
- d. None of the above are suitable.

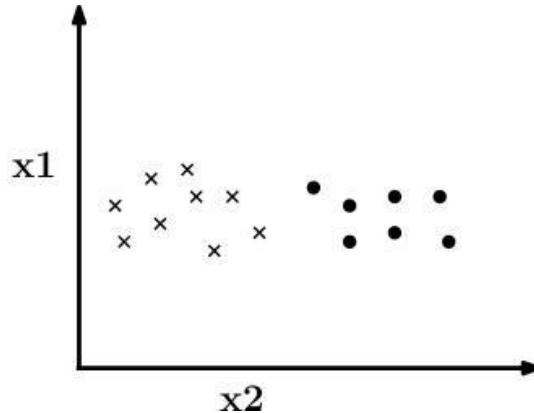


Question 6

The data points shown in the figure below have two features (x_1 and x_2) and each data point belongs to one of the two classes shown in the figure by cross and circle. If you have to select one feature to use in a classification problem, which feature will you choose?

- a. x_1
- b. x_2
- c. Any one of the above
- d. None of the above are suitable.

Answer : b



Question 7

Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable. Suppose, you want to predict the class of new data point $x=1$ and $y=1$ using Euclidean distance in 3-NN. In which class this data point belong to?

- a. + Class
- b. - Class
- c. Can't Say
- d. None of these

x	y	Class
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

Question 7

Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable. Suppose, you want to predict the class of new data point $x=1$ and $y=1$ using Euclidean distance in 3-NN. In which class this data point belong to?

- a. + Class
- b. - Class
- c. Can't Say
- d. None of these

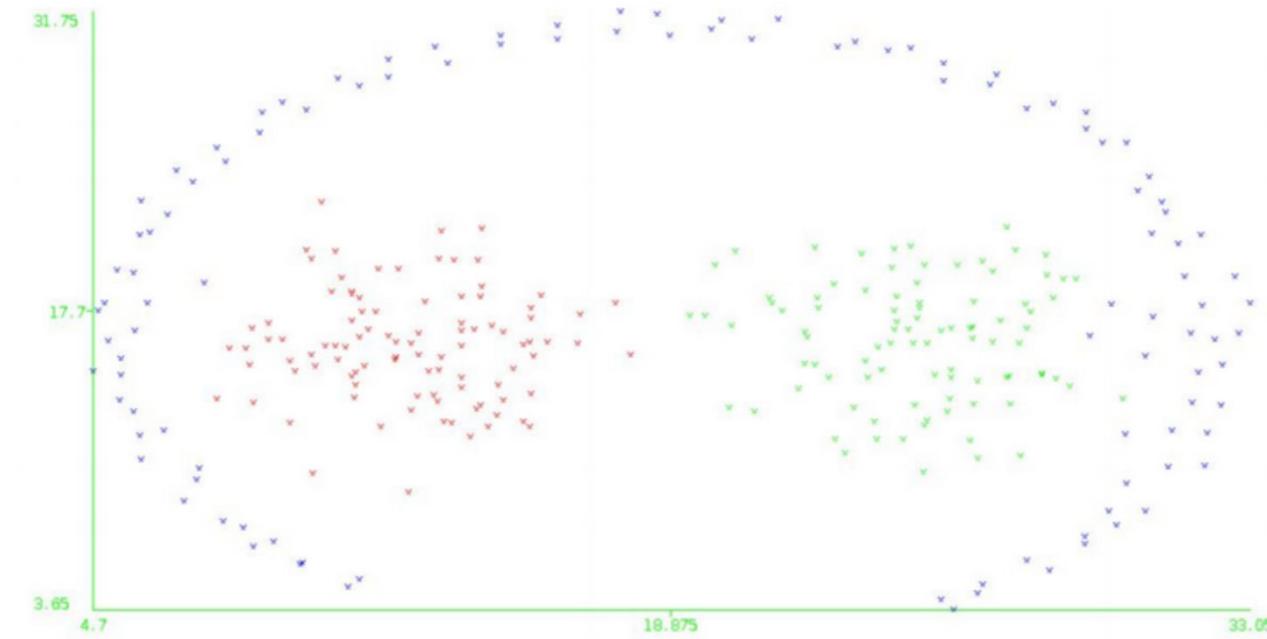
x	y	Class
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

Answer: a

Question 8

Consider the following image showing data points belonging to three different clusters (indicated by the colours of the points). Which among the following clustering algorithms will perform well in accurately clustering the given data?

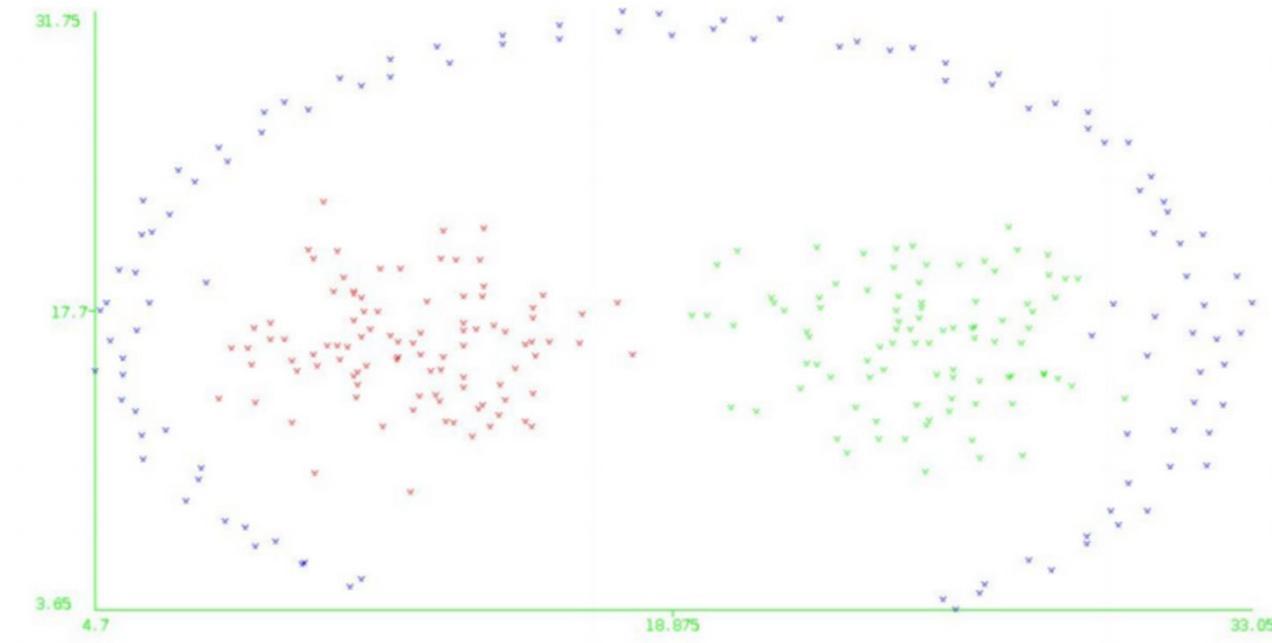
- A. K-means
- B. Single-link hierarchical
- C. Complete-link hierarchical
- D. DBSCAN



Question 8

Consider the following image showing data points belonging to three different clusters (indicated by the colours of the points). Which among the following clustering algorithms will perform well in accurately clustering the given data?

- A. K-means
- B. Single-link hierarchical
- C. Complete-link hierarchical
- D. DBSCAN



THE END