Unit 2 - Week 0

Assignment 0

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2018-01-22, 00:00 IST.

1) There are n bins of which the k-th bin contains k − 1 blue balls and n − k red balls. You pick a bin at random and remove two balls at random without replacement. Find the probability that: ***1 point***

- the second ball is red;
- the second ball is red, given that the first is red.

- 1/3, 2/3
- 1/2, 1/3
- 1/2, 2/3
- 1/3, 1/3

No, the answer is incorrect.
Score: 0

Accepted Answers:
1/2, 2/3

2) A medical company touts its new test for a certain genetic disorder. The false negative rate is small: if you have the disorder, the probability that the test returns a positive result is 0.999. The false positive rate is also small: if you do not have the disorder, the probability that the test returns a positive result is only 0.005. Assume that 2% of the population has the disorder. If a person chosen uniformly from the population is tested and the result comes back positive, what is the probability that the person has the disorder? ***1 point***

- 0.803
- 0.976
- 0.02
- 0.204

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.803

3) In an experiment, n coins are tossed, with each one showing up heads with probability p independently of the others. Each of the coins which shows up heads is then tossed again. What is the probability of observing 5 heads in the second round of tosses, if we toss 15 coins in the first round and p = 0.4? ***1 point***
(Hint: First find the mass function of the number of heads observed in the second round.)

- 0.372
- 0.055
- 0.0345
- 0.0488

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.055

4) Consider two random variables X and Y having joint density function $f(x,y) = 2e^{-x-y}, 0 < x < y < \infty$. Are X and Y independent? Find the covariance of X and Y . ***1 point***

- Yes, 1/4
- Yes, 1/2
- No, 1/4
- No, 1/2

No, the answer is incorrect.
Score: 0

Accepted Answers:
No, 1/4

s that 5 percent of the people making reservations on a certain flight will not show up. ***1 point***

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?

- 0.5101
- 0.81
- 0.6308
- 0.7405

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.7405

6) Let X have mass function *1 point*

$$f(x) = \begin{cases} \{x(x+1)\}^{-1} & \text{if x = 1, 2, ...,} \\ 0 & \text{otherwise,} \end{cases}$$

and let α ∈ R. For what values of α is it the case that $\mathrm{E}(X^{\alpha}) < \infty$?

- α < 1/2
- α < 1
- α > 1
- α > 3/4

No, the answer is incorrect.
Score: 0

Accepted Answers:
α < 1

7) Is the following a distribution function? *1 point*

$$F(x) = \begin{cases} e^{-1/x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- If so, give the corresponding density function. If not, mention why it is not a distribution function.
- Yes,$x^{-2}e^{-1/x}, x > 0$
- No, not right continuous
- Yes, $x^{-1}e^{-1/x}, x > 0$

No, the answer is incorrect.
Score: 0

Accepted Answers:
Yes,$x^{-2}e^{-1/x}, x > 0$

8) Can the value of a probability density function be greater than one? What about the cumulative distribution function? *1 point*

- PDF: yes, CDF: yes
- PDF: yes, CDF: no
- PDF: no, CDF: yes
- PDF: no, CDF: no

No, the answer is incorrect.
Score: 0

Accepted Answers:
PDF: yes, CDF: no

9) You are given a biased coin with probability of seeing a head is p = 0.6 and probability of seeing a tail is q = 0.4. Suppose you toss the coin 10 times, what is the probability of you getting the head at most 2 times? Also, what is the probability of you getting the head for the first time on your fourth attempt? *1 point*

- 0.012, 0.038
- 0.054, 0.038
- 0.012, 0.064
- 0.054, 0.064

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.012, 0.038

10) Let u be a (n × 1) vector, such that $u^T u = 1$. Let *I* be the nxn identity matrix. The (n x n) matrix A is *1 point*

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?

- -2
- -1
- 1
- 2

No, the answer is incorrect.
Score: 0

Accepted Answers:
2

11) Which of the following are true for any m × n matrix A of real numbers? ***1 point***
Note : There can be more than one correct option.

- The rowspace of A is the same as the columnspace of A^T
- The rowspace of A is the same as the rowspace of A^T
- The eigenvectors of AA^T are the same as the eigenvectors of A^TA
- The eigenvalues of AA^T are the same as the eigenvalues of A^TA

No, the answer is incorrect.
Score: 0

Accepted Answers:
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

12) The Singular Value Decomposition (SVD) of a matrix R is given by USV^T USV T . Consider an orthogonal matrix Q such that A = QR. The SVD of A is given by $U_1S_1V_1^T$. Which of the following are true? ***1 point***

- $U = U_1$
- $S = S_1$
- $V = V_1$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$S = S_1$
$V = V_1$

13) Let $A_{n\times n}$ be a row stochastic matrix - in other words, all elements are non-negative and the sum of elements in every row is 1. Let b be an eigenvalue of A. Which of the following is true? ***1 point***

- |b|>1
- |b| <= 1
- |b| >= 1
- |b| < 1

No, the answer is incorrect.
Score: 0

Accepted Answers:
|b| <= 1

14) Let $A_{n\times n}$ 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 ***1 point***

- A^TAx
- A^Tx
- x
- Cannot be described in terms of x

No, the answer is incorrect.
Score: 0

Accepted Answers:
A^Tx

Unit 3 - Week 1

Assignment 1

The due date for submitting this assignment has passed. **Due on 2018-02-05, 23:59 IST.**

Assignment submitted on 2018-01-23, 22:42 IST

1) Which of the following is a supervised learning problem? *1 point*
[Note: Multiple options may be correct]

- [ ] Grouping people in a social network.
- [x] Predicting credit approval based on historical data
- [x] Predicting rainfall based on historical data
- [ ] All of the above

Yes, the answer is correct.
Score: 1

Accepted Answers:
Predicting credit approval based on historical data
Predicting rainfall based on historical data

2) Which of the following are classification problems? (multiple options may be correct) *1 point*

- [ ] Predicting the amount of rain fall for a particular day.
- [x] Predicting whether it will rain or not on a particular day.
- [x] Given all the actors in a movie, predicting its genre.
- [x] Filtering of spam messages

Yes, the answer is correct.
Score: 1

Accepted Answers:
Predicting whether it will rain or not on a particular day.
Given all the actors in a movie, predicting its genre.
Filtering of spam messages

3) Which of the following is a regression task? *1 point*

- ○ Predicting whether a given document is related to sports or not
- ○ Predicting the gender of a human
- ◉ Predicting the share price of a company
- ○ Finding clusters in a given data

Yes, the answer is correct.
Score: 1

Accepted Answers:
Predicting the share price of a company

4) Which of the following is an unsupervised learning task? *1 point*

- ○ Learning to ride a bicycle.
- ◉ Grouping related documents from an unannotated corpus.
- ○ Grouping students into following groups- primary, high school, college.
- ○ both (a) and (c)

Yes, the answer is correct.
Score: 1

Accepted Answers:
Grouping related documents from an unannotated corpus.

5) Which of the following is a categorical feature? *1 point*

- ○ Weight of a person
- ◉ Ethnicity of a person
- ○ Height of a person
- ○ Income of a person

Yes, the answer is correct.

Score: 1

Accepted Answers:
Ethnicity of a person

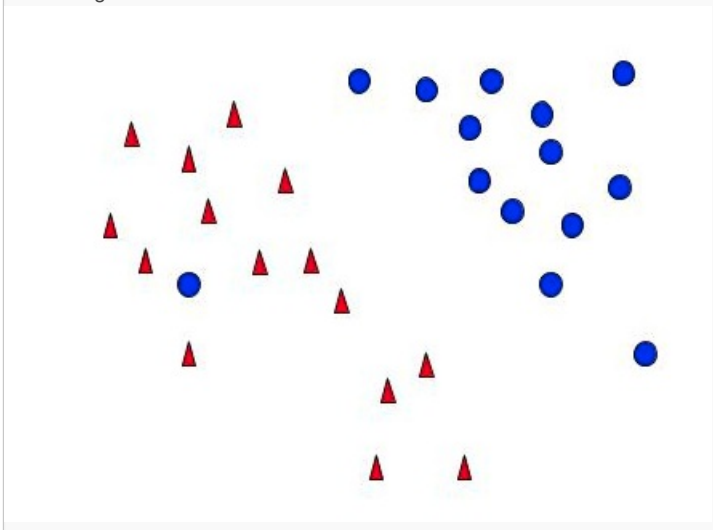
6) A new phone, E-Corp X1 has been announced and it is what you've been waiting for, all along. You decide to read the reviews before buying it. From past experiences, you've figured out that good reviews mean that the product is good 90% of the time and bad reviews mean that it is bad 70% of the time. Upon glancing through the reviews section, you find out that the X1 has been reviewed 1269 times and only 127 of them were bad reviews. What is the probability that, if you order the X1, it is a bad phone? ***1 point***

- ○ 0.1362
- ◉ 0.160
- ○ 0.840
- ○ 0.773

Yes, the answer is correct.
Score: 1

Accepted Answers:
0.160

7) What would be the ideal complexity of the curve which can be used for separating the two classes shown in the image below? ***1 point***

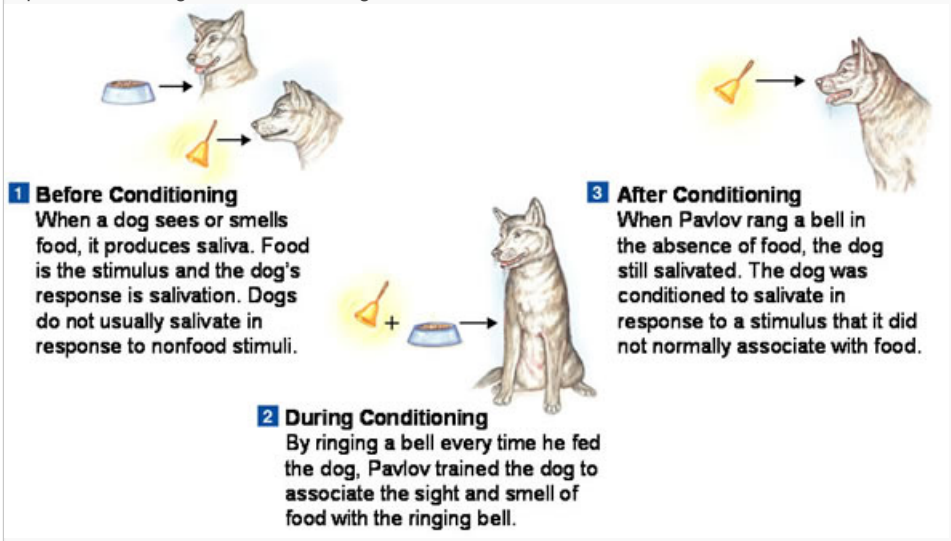


- ○ Linear
- ○ Quadratic
- ○ Cubic
- ◉ Insufficient data to draw conclusion

No, the answer is incorrect.
Score: 0

Accepted Answers:
Linear

8) Pavlov's experiment is a classic experiment conducted by the Russian physiologist Ivan Pavlov. He experiments on dogs shown in the image below. ***1 point***



Before conditioning, dog responds with saliva only in presence of the food but after conditioning, it starts salivating

just with the bell. Select the correct option(s) about the experiment.
[Multiple options may be correct]

- [x] In this experiment, the dog learns in a supervised setting
- [x] In this experiment, the dog acts as a Reinforcement learning agent
- [x] Comparing this experiment to Reinforcement learning theory, the various states are:
  - Presence of just the bell
  - Presence of just food
  - Presence of both food and bell

No, the answer is incorrect.
Score: 0

Accepted Answers:
In this experiment, the dog acts as a Reinforcement learning agent
Comparing this experiment to Reinforcement learning theory, the various states are:

- *Presence of just the bell*
- *Presence of just food*
- *Presence of both food and bell*

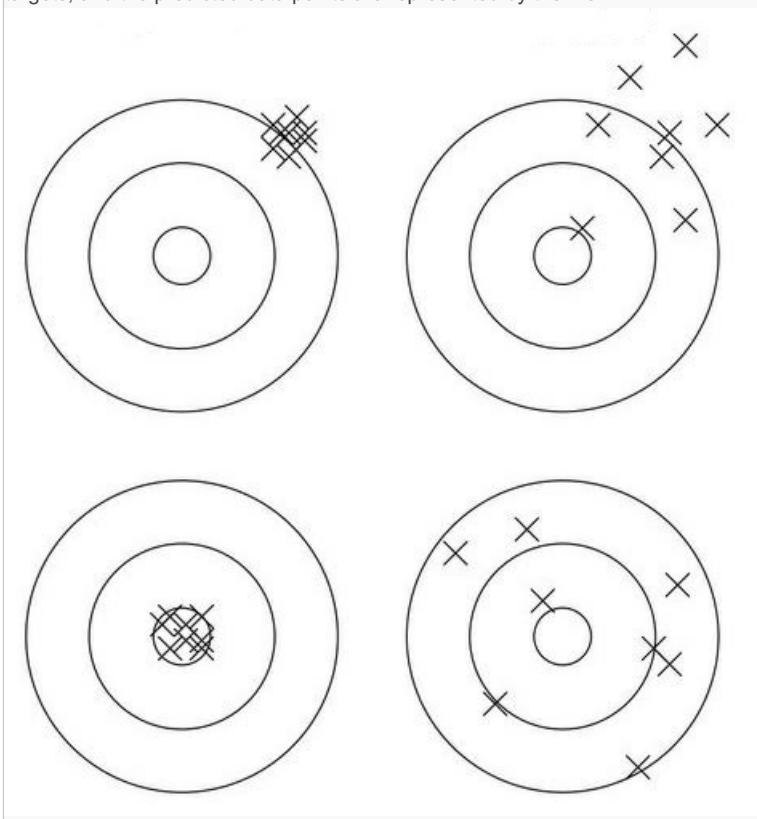
9) One of the most common uses of Machine Learning today is in the domain of Robotics. Robotic tasks include a multitude of ML methods tailored towards navigation, robotic control and a number of other tasks. Robotic control includes controlling the actuators available to the robotic system. An example of this is control of a painting arm in automotive industries. The robotic arm must be able to paint every corner in the automotive parts while minimizing the quantity of paint wasted in the process. Which of the following learning paradigms would you select for training such a robotic arm? ***1 point***

- ( ) Supervised learning
- ( ) Unsupervised learning
- ( ) Combination of supervised and unsupervised learning
- (•) Reinforcement learning

Yes, the answer is correct.
Score: 1

Accepted Answers:
Reinforcement learning

10) Consider the following diagram where we assume that the target data points lie on the bullseye of the targets, and the predicted data points are represented by the X's. ***1 point***



Based on the above diagram, which of the following statements are true?

- [x] The top left diagram represents a model with high bias and low variance
- [ ] The variance in the top right diagram is less than the variance in the bottom left diagram
- [x] The variance in the top left diagram is less than the variance in the bottom right diagram

☑ To improve the model on the bottom right diagram, we should focus more on reducing variance rather than bias

Yes, the answer is correct.
Score: 1

Accepted Answers:
The top left diagram represents a model with high bias and low variance
The variance in the top left diagram is less than the variance in the bottom right diagram
To improve the model on the bottom right diagram, we should focus more on reducing variance rather than bias

Previous Page End

Unit 4 - Week 2

Assignment 2

The due date for submitting this assignment has passed. **Due on 2018-02-09, 23:59 IST.**

Assignment submitted on 2018-02-09, 23:31 IST

1) The parameters obtained in linear regression *1 point*

- (•) can take any value in the real space
- ( ) are strictly integers
- ( ) always lie in the range [0,1]
- ( ) can take only non zero values

Yes, the answer is correct.
Score: 1

Accepted Answers:
can take any value in the real space

2) Given a set of n data points, $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, the best least squares fit $f(x)$ is obtained by minimization of: *1 point*

- ( ) $\sum_{i=1}^{n}[y_i - f(x_i)]$
- ( ) $\min\ (y_i - f(x_i))$
- (•) $\sum_{i=1}^{n}[y_i - f(x_i)]^2$
- ( ) $\max\ (y_i - f(x_i))$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$\sum_{i=1}^{n}[y_i - f(x_i)]^2$

3) Consider forward selection, backward selection and best subset selection with respect to the same data set. Which of the following is true? *1 point*

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

Yes, the answer is correct.
Score: 1

Accepted Answers:
Best subset selection can be computationally more expensive than forward selection

4) Adding interaction terms (such as products of two dimensions) along with original features in linear regression *1 point*

- (•) reduces training error.
- ( ) increases training error.
- ( ) doesn't affect training error.

Yes, the answer is correct.
Score: 1

Accepted Answers:
reduces training error.

5) Consider the following five training examples *1 point*

$X = [2\ 3\ 4\ 5\ 6]$
$Y = [12.8978\ 17.7586\ 23.3192\ 28.3129\ 32.1351]$
We want to learn a function $f(x)$ of the form $f(x) = ax + b$ which is parameterized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

- ( ) (4,3)
- (•) (5,3)
- ( ) (5,1)
- ( ) (1,5)

Yes, the answer is correct.
Score: 1

Accepted Answers:
(5,3)

6) A study was conducted to understand the effect of number of hours the students spent studying to their performance in the *1 point*

final exams. You are given the following 8 samples from the study. What is the best linear fit on this dataset?

Number of hours spent studying (x)	Score in the final exam (0-100) (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

- ○ $y = -3.39x + 11.62$
- ◉ $y = 4.59x + 12.58$
- ○ $y = 3.39x + 10.58$
- ○ $y = 4.69x + 11.62$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$y = 4.59x + 12.58$

7) kNN regressor outputs the average of the k nearest neighbours of a query point. Consider a variant of kNN regressor where instead of returning the average we fit a linear regression model on the k neighbours. Which of the following **do not hold true** for this new variant? ***1 point***
[more than one option can be correct]

- ☐ This method makes an assumption that the data is globally linear.
- ☑ This method makes an assumption that the data is locally linear.
- ☑ This method has lower bias compared to kNN
- ☐ This method has lower variance compared to kNN

No, the answer is incorrect.
Score: 0

Accepted Answers:
This method makes an assumption that the data is globally linear.
This method has lower variance compared to kNN

8) Which of the following shrinkage methods leads to sparse solution? ***1 point***

- ◉ Lasso regression
- ○ Ridge regression
- ○ Lasso and ridge regression both return sparse solutions

Yes, the answer is correct.
Score: 1

Accepted Answers:
Lasso regression

9) Consider the design matrix X of dimension $N \times (p+1)$. Which of the following statements are true? ***1 point***
[more than one option can be correct]

- ☑ The rowspace of X is the same as the columnspace of X^T
- ☐ The rowspace of X is the same as the rowspace of X^T
- ☐ The eigenvectors of XX^T are the same as the eigenvectors of X^TX
- ☑ The eigenvalues of XX^T are the same as the eigenvalues of X^TX

Yes, the answer is correct.
Score: 1

Accepted Answers:
The rowspace of X is the same as the columnspace of X^T
The eigenvalues of XX^T are the same as the eigenvalues of X^TX

10)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) ***1 point***

- ○ X
- ○

S

X_c

V

U

Yes, the answer is correct.
Score: 1

Accepted Answers:
V

Previous Page

End

Unit 5 - Week 3

Assignment 3

The due date for submitting this assignment has passed. **Due on 2018-02-14, 23:59 IST.**

Assignment submitted on 2018-02-14, 22:49 IST

1) Select the valid reasons for doing dimensionality reduction. ***1 point***
[more than one option may be correct]

- [x] Variance can be controlled by controlling the number of features used in the model.
- [x] Lower number of features render the model to better interpretability.
- [x] Upon reducing the feature pool, we can sometimes increase the prediction accuracy, though not always.
- [x] Upon reducing the feature pool, we can reduce the time taken to run inference on the model at the cost of increased model building time.

Yes, the answer is correct.
Score: 1

Accepted Answers:
Variance can be controlled by controlling the number of features used in the model.
Lower number of features render the model to better interpretability.
Upon reducing the feature pool, we can sometimes increase the prediction accuracy, though not always.
Upon reducing the feature pool, we can reduce the time taken to run inference on the model at the cost of increased model building time.

2) You work as a data analyst and your job is to analyse the growth of sale of a product. Suppose you decide to fit a linear regression model on a multivariate dataset. You find that a particular feature has a very high negative coefficient. What can you infer from this? ***1 point***

- (•) We can't comment on the importance of this feature without any additional information
- ( ) Since the feature has a large negative coefficient, so it is not an important feature. It is better to discard it.
- ( ) Since the magnitude of the coefficient is very high, we should never discard that feature.

Yes, the answer is correct.
Score: 1

Accepted Answers:
We can't comment on the importance of this feature without any additional information

3) \item Which of the following dimensionality reduction methods are **hard** thresholding methods? ***1 point***
(A method is **soft**-thresholding if it contains parameters that vary in a continuous fashion)

- [x] Forward Stepwise Regression
- [x] Backward Stepwise Regression
- [x] Forward Stagewise Regression
- [ ] Ridge Regression
- [ ] Least Absolute Shrinkage and Selection Operator (LASSO)
- [x] Principal Component Regression
- [x] Partial Least Squares Method

Yes, the answer is correct.
Score: 1

Accepted Answers:
Forward Stepwise Regression
Backward Stepwise Regression
Forward Stagewise Regression
Principal Component Regression
Partial Least Squares Method

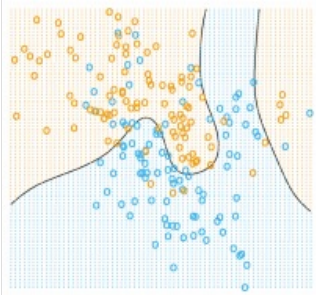
4) Suppose that in applying linear regression, we are working with a data set where there are a large number of features, many of which we suspect to be redundant. We have discussed how using regularization we can constrain the magnitude of the weights associated with each feature. In fact, using regularization we can eliminate certain redundant features where the magnitude of the weights are found to be zero. Suppose we have a choice between two regularization schemes, Ridge regularization, where the additional penalty term is the sum of squares of the weights and LASSO regularization, where the penalty term is the sum of the absolute values of the weights. Which method do you think will result in eliminating more features (by reducing corresponding weights to zero)? ***1 point***

- (•) LASSO
- ( ) Ridge
- ( ) either of LASSO or Ridge regression can be used

Yes, the answer is correct.
Score: 1

Accepted Answers:
LASSO

5) Which of the following algorithm could have generated this decision boundary? (consider the situation where we do not allow for basis expansion) ***1 point***



- ( ) Linear regression with indicator variable
- ( ) Logistic Regression
- ( ) 1-NN
- (•) None of the above

Yes, the answer is correct.
Score: 1

Accepted Answers:
None of the above

6) Which of the following is true about a logistic regression based classifier? ***1 point***

- [x] The logistic function is non-linear in the weights
- [ ] The logistic function is linear in the weights
- [x] The decision boundary is linear in the weights
- [ ] The decision boundary is non-linear in the weights

Yes, the answer is correct.
Score: 1

Accepted Answers:
The logistic function is non-linear in the weights
The decision boundary is linear in the weights

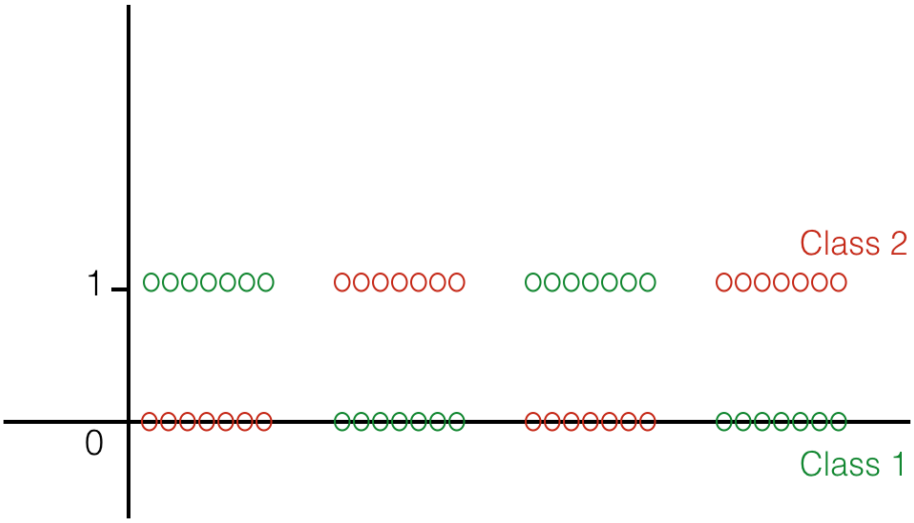
7) For a two class classification problem, which among the following are true? ***1 point***

- [ ] If both the covariance matrices are spherical and equal, the within class variance term has an effect on the LDA derived direction.
- [x] If both the covariance matrices are spherical and equal, the within class variance term has no effect on the LDA derived direction.
- [x] If both the covariance matrices are spherical but unequal, the within class variance term has an effect on the LDA derived direction.
- [ ] If both the covariance matrices are spherical but unequal, the within class variance term has no effect on the LDA derived direction.

No, the answer is incorrect.
Score: 0

Accepted Answers:
If both the covariance matrices are spherical and equal, the within class variance term has no effect on the LDA derived direction.
If both the covariance matrices are spherical but unequal, the within class variance term has no effect on the LDA derived direction.

8) In the following dataset, there are two classes arranged in the following manner ***1 point***



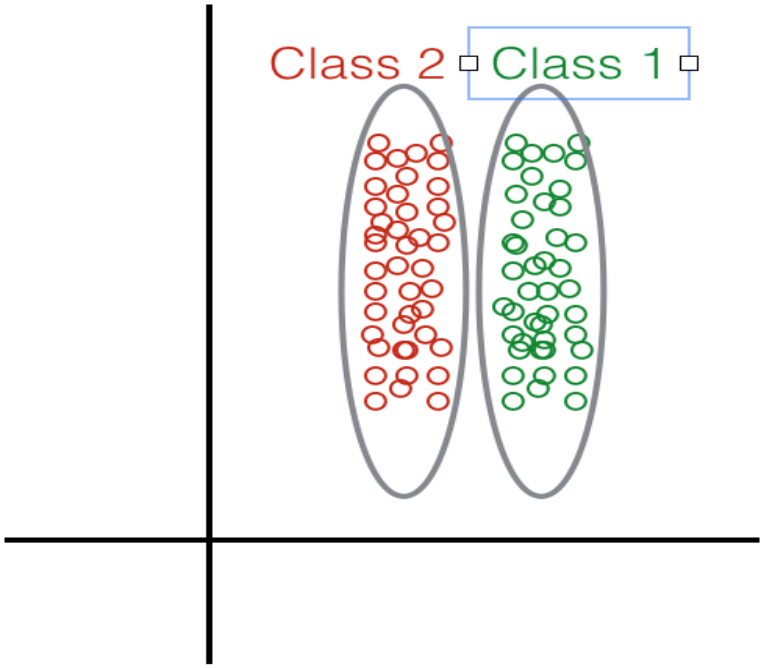
Which of the following bases functions would you use to prevent any masking?
Select all that apply.

- [x] $1, x, x^2, x^3$
- [x] $1, x, x^3, x^4$
- [ ] $1, x, x^2$
- [ ] $1, x, sin(x)$

Partially Correct.
Score: 0.66

Accepted Answers:
$1, x, x^2, x^3$
$1, x, x^3, x^4$
$1, x, sin(x)$

9) Given the following distribution of data points *1 point*



What method would you choose to perform dimensionality reduction?

- (•) Linear Discriminant Analysis
- ( ) Principal Component Analysis

Yes, the answer is correct.
Score: 1

Accepted Answers:
Linear Discriminant Analysis

10) In general, which of the following classification methods is the most resistant to gross outliers? ***1 point***

- ( ) Quadratic Discriminant Analysis (QDA)
- ( ) Linear Regression
- (•) Logistic regression
- ( ) Linear Discriminant Analysis (LDA)

Yes, the answer is correct.
Score: 1

Accepted Answers:
Logistic regression

Previous Page

End

Unit 6 - Week 4

Assignment 4

The due date for submitting this assignment has passed. **Due on 2018-02-23, 23:59 IST.**

Assignment submitted on 2018-02-21, 10:43 IST

1) Which the following loss functions are convex? *1 point*
[more than one options may be correct]

- [ ] 0-1 loss (sometimes referred as mis-classification loss)
- [x] Hinge loss
- [ ] Logistic loss
- [x] Squared error loss

Partially Correct.
Score: 0.67

Accepted Answers:
Hinge loss
Logistic loss
Squared error loss

2) When using SVMs, what effect, in general, can you expect on the size of the margins when the C parameter is decreased? *1 point*

- ( ) the margins may become wider
- ( ) the margins may become narrower
- ( ) no relation between C and margin sizes

No, the answer is incorrect.
Score: 0

Accepted Answers:
the margins may become wider

3) For Q3,4: Kindly download the synthetic dataset from this link. The dataset contains 1000 points and each input point contains 3 features. *2 points*
Train a linear regression model (without regularization) on the above dataset. Report the coefficients of the best fit model. Report the coefficients in the following format:
$\beta_0, \beta_1, \beta_2, \beta_3$

- ( ) -1.2, 2.1, 2.2, 1
- ( ) 1, 1.2, 2.1, 2.2
- ( ) -1, 1.2, 2.1, 2.2
- (•) 1, -1.2, 2.1, 2.2
- ( ) 1, 1.2, -2.1, -2.2

Yes, the answer is correct.
Score: 2

Accepted Answers:
1, -1.2, 2.1, 2.2

4) Train a l2 regularized linear regression model on the above dataset. Vary the regularization parameter from 1 to 10. As you increase the regularization parameter, absolute value of the coefficients (excluding the intercept) of the model: *2 points*

- ( ) increase
- ( ) first increase then decrease
- (•) decrease
- ( ) first decrease then increase
- ( ) does not change

Yes, the answer is correct.
Score: 2

Accepted Answers:
decrease

5) For Q5,6: Kindly download the modified version of Iris dataset from this link. The dataset contains 150 points and each input point contains 4 features and belongs to one among three classes. Use the first 100 points as the training data and the remaining 50 as test data. ***2 points***

\item Train a L2 regularized logistic regression classifier on the modified iris dataset. We recommend using sklearn. Use only the first two features for your model. We encourage you to explore the impact of varying different hyperparameters of the model. Kindly note that the C parameter mentioned below is the inverse of the regularization parameter λ. As part of the assignment, train a model with the following hyperparameters:
Model: logistic regression with one-vs-rest classifier, C = 1e4

For the above set of hyperparameters, report the best classification accuracy

- 0.88
- 0.86
- 0.92
- 0.68

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.86

6) Train an SVM classifier on the modified iris dataset. We recommend using sklearn. Use only the first two features for your model. We encourage you to explore the impact of varying different hyperparameters of the model. Specifically try different kernels and the associated hyperparameters. As part of the assignment train models with the following set of hyperparameters: ***2 points***
RBF-kernel, $gamma = 0.5$, one-vs-rest classifier, no-feature-normalization.
Try C = [0.01, 1, 10]. For the above set of hyperparameters, report the best classification accuracy along with total number of support vectors on the test data.

- 0.88, 69
- 0.44, 69
- 0.68, 44
- 0.34, 44

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.88, 69

Previous Page

End

Unit 7 - Week 5

Assignment 5

The due date for submitting this assignment has passed. **Due on 2018-02-28, 23:59 IST.**

Assignment submitted on 2018-02-28, 23:56 IST

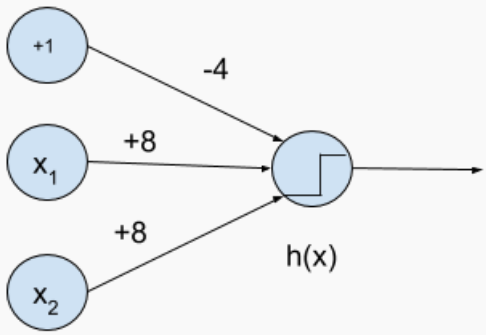
1) Which of the following is/are true about the Perceptron classifier? *1 point*

- [x] It can learn a OR function
- [x] It can learn a AND function
- [x] The obtained separating hyperplane depends on the order in which the points are presented in the training process.
- [ ] For a linearly separable problem, there exists some initialization of the weights which might lead to non-convergent cases.

Yes, the answer is correct.
Score: 1

Accepted Answers:
It can learn a OR function
It can learn a AND function
The obtained separating hyperplane depends on the order in which the points are presented in the training process.

2) 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? *1 point*



- (•) OR
- ( ) AND
- ( ) NAND
- ( ) None of the above

Yes, the answer is correct.
Score: 1

Accepted Answers:
OR

3) We have a function which takes a two-dimensional input $x = (x_1, x_2)$ and has two parameters $w = (w_1, w_2)$ given by $f(x, w) = \sigma(\sigma(x_1 w_1) w_2 + x_2)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$. *2 points*
We use backpropagation to estimate the right parameter values. We start by setting both the parameters to 0.
Assume that we are given a training point $x_1 = 0, x_2 = 1, y = 5$.
Given this information **answer this and next question**.
What is the value of $\frac{\partial f}{\partial w_2}$?

- ( ) 0.693
- (•) 0.098
- ( ) 0.125

DOWNLOAD VIDEOS

- -0.531

Yes, the answer is correct.
Score: 2

Accepted Answers:
0.098

4) If the learning rate is 0.5, what will be the value of w_2 after one update using backpropagation algorithm? ***2 points***

- -0.5625
- -0.4423
- 0.5625
- 0.4423

Yes, the answer is correct.
Score: 2

Accepted Answers:
0.4423

5) We are given a 2-class classification problem with [0/1] output labels. We plan to use a neural network to implement the classifier. Which of the following functions is a suitable choice for the output neurons? ***1 point***

- Hyperbolic Tangent Neuron - tanh(.)
- Linear Neuron
- Arctangent Neuron - arctan(.)
- Logistic Sigmoid Neuron

Yes, the answer is correct.
Score: 1

Accepted Answers:
Logistic Sigmoid Neuron

6) Given N samples $x_1, x_2, \ldots, x_N$ drawn independently from a Gaussian distribution with variance σ_2 and unknown mean μ, find the MLE of the mean. ***1 point***

- $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{\sigma^2}$
- $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{2\sigma^2 N}$
- $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$
- $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N-1}$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

7) Which of the following statements is false: ***1 point***

- The chances of overfitting decrease with Increasing the number of hidden nodes and increasing the number of hidden layers.
- A neural network with one hidden layer can represent any Boolean function given sufficient number of hidden units and appropriate activation functions.
- 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 are used.

Yes, the answer is correct.
Score: 1

Accepted Answers:
The chances of overfitting decrease with Increasing the number of hidden nodes and increasing the number of hidden layers.

8) Which of the following are true when comparing ANNs and SVMs? ***1 point***

[multiple options may be correct]

- [x] ANN error surface has multiple local minima while SVM error surface has only one minima
- [x] After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.
- [x] In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.
- [ ] 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.

Yes, the answer is correct.
Score: 1

Accepted Answers:
ANN error surface has multiple local minima while SVM error surface has only one minima
After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.
In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.

Previous Page End

Unit 8 - Week 6

Assignment 6

The due date for submitting this assignment has passed. **Due on 2018-03-07, 23:59 IST.**

Assignment submitted on 2018-03-07, 23:42 IST

1) What is specified at any non-leaf node in a decision tree? *1 point*

- (•) Class of instance
- ( ) Data value description
- ( ) Test specification
- ( ) Data process description

No, the answer is incorrect.
Score: 0

Accepted Answers:
Test specification

2) Which of the following statements are true with respect to the application of Cost-Complexity Pruning and Reduced Error Pruning with Cross-Validation? *1 point*
(Multiple options may be correct)

- [x] In Reduced Error Pruning, the pruned tree error can never be less than the original tree on the training dataset.
- [x] In Cost Complexity Pruning, the pruned tree error can never be less than the original tree on the training dataset.
- [ ] In Cost Complexity Pruning, the pruned tree error can never be less than the original tree on the validation dataset.
- [ ] In Reduced Error Pruning, the pruned tree error can never be less than the original tree on the validation dataset.

Yes, the answer is correct.
Score: 1

Accepted Answers:
In Reduced Error Pruning, the pruned tree error can never be less than the original tree on the training dataset.
In Cost Complexity Pruning, the pruned tree error can never be less than the original tree on the training dataset.

3) Which of these classifiers do not require any additional modifications to their original descriptions (as seen in the lectures) to use them when we have more than 2 classes? *1 point*
(Multiple options may be correct)

- [ ] decision trees
- [x] logistic regression
- [x] support vector machines
- [x] k nearest neighbors

No, the answer is incorrect.
Score: 0

Accepted Answers:
decision trees
k nearest neighbors

4) Consider the following data set. *1 point*

DOWNLOAD VIDEOS

price	maintenance	capacity	airbag	profitable
low	low	2	no	yes
low	med	4	yes	yes
low	low	4	no	yes
low	med	4	no	no
low	high	4	no	no
med	med	4	no	no
med	med	4	yes	yes
med	high	2	yes	no
med	high	5	no	yes
high	med	4	yes	yes
high	med	2	yes	yes
high	high	2	yes	no
high	high	5	yes	yes

Considering 'profitable' as the binary valued attribute we are trying to predict, which of the attributes would you select as the root in a decision tree with multi-way splits using the cross-entropy impurity measure?

- price
- (•) maintenance
- capacity
- airbag

Yes, the answer is correct.
Score: 1

Accepted Answers:
maintenance

5) In the above data set, what is the value of cross entropy when we consider capacity as the attribute to split on (multi-way splits)? ***1 point***

- 0.7973
- (•) 0.8684
- 0.8382
- 0.7688

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.8382

6) For the same data set, suppose we decide to construct a decision tree using binary splits and the gini index impurity measure. Which among the following feature and split point combinations would be the best to use as the root node assuming that we consider each of the input features to be unordered? ***1 point***

- (•) price {med,low}|{high}
- capacity {2,4}|{5}
- maintenance {high}|{med,low}
- maintenance {high,med}|{low}

No, the answer is incorrect.
Score: 0

Accepted Answers:
maintenance {high}|{med,low}

7) In the above question, what is the gini index value when we decide to split on the attribute price according to the following split: {med}|{low, high}? ***1 point***

- 0.4505
- (•) 0.4573
- 0.4196
- 0.4615

No, the answer is incorrect.

Score: 0

Accepted Answers:
0.4615

8) An important factor that influences the variance of decision trees is the average height of the tree. For the same dataset, if we limited the height of the trees to some H, how would the variance of the decision tree algorithm be affected? ***1 point***

- (•) Variance may increase with tree length H.
- ( ) Variance may decrease with tree length H.
- ( ) Variance is unaffected by tree length H.

Yes, the answer is correct.
Score: 1

Accepted Answers:
Variance may increase with tree length H.

9) In which of the following situations is it appropriate to introduce a new category 'Missing' for missing values? ***1 point***
(Multiple answers may be correct)

- [x] When values are missing because the 108 emergency operator is sometimes attending a very urgent distress call.
- [ ] When values are missing because the attendant spilled coffee on the papers from which the data was extracted.
- [ ] When values are missing because the warehouse storing the paper records went up in flames and burnt parts of it.
- [x] When values are missing because the nurse/doctor finds the patient's situation too urgent.

Yes, the answer is correct.
Score: 1

Accepted Answers:
When values are missing because the 108 emergency operator is sometimes attending a very urgent distress call.
When values are missing because the nurse/doctor finds the patient's situation too urgent.

10) Which of the following properties are true in the context of decision trees? ***1 point***
(Multiple answers may be correct)

- [x] High bias
- [ ] High variance
- [x] Lack of smoothness of prediction surfaces
- [x] Unbounded parameter set

No, the answer is incorrect.
Score: 0

Accepted Answers:
High variance
Lack of smoothness of prediction surfaces
Unbounded parameter set

Previous Page End

Unit 9 - Week 7

Assignment 7

The due date for submitting this assignment has passed. **Due on 2018-03-14, 23:59 IST.**

Assignment submitted on 2018-03-14, 23:51 IST

1) Which of the following factors need to be taken into account while setting up an experiment? (More than one answer may be correct) *1 point*

- [x] Floor/Ceiling Effects
- [ ] Order Effects
- [x] Sampling Bias

Partially Correct.
Score: 0.67

Accepted Answers:
Floor/Ceiling Effects
Order Effects
Sampling Bias

2) Select the correct equations. *1 point*
TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative
(More than one answer may be correct)

- [x] Precision = $\frac{TP}{TP+FP}$
- [ ] Recall = $\frac{FP}{TP+FP}$
- [x] Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- [x] Recall = $\frac{TP}{TP+FN}$

Yes, the answer is correct.
Score: 1

Accepted Answers:
Precision = $\frac{TP}{TP+FP}$
Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
Recall = $\frac{TP}{TP+FN}$

3) Which of the following measure best analyze the performance of a classifier? *1 point*

- ( ) Precision
- ( ) Recall
- ( ) Accuracy
- ( ) Time complexity
- (•) Depends on the application

Yes, the answer is correct.
Score: 1

Accepted Answers:
Depends on the application

4) For the ROC curve of True positive rate vs False positive rate, which of the following are true? *1 point*

- ○ The curve is always concave (negative convex).
- ○ The curve is never concave.
- ◉ The curve may or may not be concave.

Yes, the answer is correct.
Score: 1

Accepted Answers:
The curve may or may not be concave.

5) What are the quantities in the **Receiver Operating Characteristics** (ROC) curve along the x and y axes? ***1 point***

- ○ x - Precision, y - Recall
- ○ x - True Positive, y - True Negative
- ○ x - Specificity, y - Sensitivity
- ◉ x - False Positive Rate, y - True Positive Rate

Yes, the answer is correct.
Score: 1

Accepted Answers:
x - False Positive Rate, y - True Positive Rate

6) In case of limited training data, which technique, bagging or stacking, would be preferred, and why? ***1 point***

- ◉ Bagging, because we can combine as many classifier as we want by training each on a different sample of the training data
- ○ Bagging, because we use the same classification algorithms on all samples of the training data
- ○ Stacking, because each classifier is trained on all of the available data
- ○ Stacking, because we can use different classification algorithms on the training data

No, the answer is incorrect.
Score: 0

Accepted Answers:
Stacking, because each classifier is trained on all of the available data

7) How does bagging help in improving the classification performance? (Multiple answers may be correct) ***1 point***

- ☐ If the parameters of the resultant classifiers are fully uncorrelated (independent), then bagging is inefficient.
- ☐ It helps reduce bias
- ☑ If the parameters of the resultant classifiers are fully correlated, then bagging is inefficient.
- ☑ It helps reduce variance

Yes, the answer is correct.
Score: 1

Accepted Answers:
If the parameters of the resultant classifiers are fully correlated, then bagging is inefficient.
It helps reduce variance

8) Which among the following prevents over-fitting when we perform bagging? ***1 point***

- ○ The use of sampling with replacement as the sampling technique
- ○ The use of weak classifiers
- ○ The use of classification algorithms which are not prone to overfitting
- ◉ The practice of validation performed on every classifier trained

No, the answer is incorrect.
Score: 0

Accepted Answers:
The use of sampling with replacement as the sampling technique

9) Which of the following statements are TRUE when comparing Committee Machines and Stacking? (Multiple answers may be correct) ***1 point***

- ☑ Committee Machines are, in general, special cases of 2-layer stacking where the second-layer classifier provides uniform weightage.
- ☐ Both Committee Machines and Stacking have similar mechanisms, but Stacking uses different classifiers while Committee Machines use similar classifiers.
- ☑ Committee Machines are more powerful than Stacking
- ☐ Committee Machines are less powerful than Stacking

No, the answer is incorrect.
Score: 0

Accepted Answers:
Committee Machines are, in general, special cases of 2-layer stacking where the second-layer classifier provides uniform weightage.
Committee Machines are less powerful than Stacking

10) Which of the following are true about using 5-fold cross validation with a data set of size n = 100 to select the value of k in the kNN algorithm? *1 point*
(More than one option may be correct)

- [ ] Will always result in the same k since it does not involve any randomness.
- [x] Might give different answers depending on the splitting in 5 fold cross validation.
- [ ] Does not make sense since n is larger than the number of folds.

Yes, the answer is correct.
Score: 1

Accepted Answers:
Might give different answers depending on the splitting in 5 fold cross validation.

Previous Page End

Unit 10 - Week 8

Assignment 8

The due date for submitting this assignment has passed. **Due on 2018-03-21, 23:59 IST.**

Assignment submitted on 2018-03-21, 23:48 IST

1) Which of the following properties is false in the case of a Bayesian Network? *1 point*

- The edges are directed
- Contains cycles
- Represents conditional independence relations among random variables
- All of the above

Yes, the answer is correct.
Score: 1

Accepted Answers:
Contains cycles

2) A and B are Boolean random variables. *1 point*

Given:
$P(A = True) = 0.3, P(A = False) = 0.7, P(B = True|A = True) = 0.4, P(B = False|A = True) = 0.6, P(B = True|A = False) = 0.6, P(B = False|A$
Calculate $P(A = True|B = False)$ by Bayes rule.

- 0.49
- 0.39
- 0.37
- 0.28

Yes, the answer is correct.
Score: 1

Accepted Answers:
0.39

3) If you have a bad classifier, which of the following ensemble methods will give the worst performance when including the given classifier? *1 point*

- Gradient Boosting
- AdaBoost
- Bagging
- Committee Machine

No, the answer is incorrect.
Score: 0

Accepted Answers:
Bagging

4) Which among Gradient Boosting and AdaBoost is less susceptible to outliers considering their respective loss functions? *1 point*

- AdaBoost
- Gradient Boost
- On average, both are equally susceptible.

Yes, the answer is correct.
Score: 1

Accepted Answers:
Gradient Boost

5) Which of the following method(s) is/are not inherently sequential? *1 point*
[Note: Multiple options may be correct]

- [ ] Gradient Boosting
- [x] Committee machines
- [ ] AdaBoost

Yes, the answer is correct.
Score: 1

Accepted Answers:
Committee machines

6) Boosting techniques typically give very high accuracy classifiers by sequentially training a collection of similar low-accuracy classifiers. *1 point*
Which of the following statements are true with respect to Boosting?
[multiple options may be correct]

- [x] LogitBoost (like AdaBoost, but with Logistic Loss instead of Exponential Loss) is less susceptible to overfitting than AdaBoost.
- [x] Boosting techniques tend to have low bias and high variance
- [ ] Boosting techniques tend to have low variance and high bias
- [ ] For basic linear regression classifiers, there is no effect of using Gradient Boosting.

Partially Correct.
Score: 0.66

Accepted Answers:
LogitBoost (like AdaBoost, but with Logistic Loss instead of Exponential Loss) is less susceptible to overfitting than AdaBoost.
Boosting techniques tend to have low bias and high variance
For basic linear regression classifiers, there is no effect of using Gradient Boosting.

7) While using Random Forests, if the input data is such that it contains a large number (>80%) of irrelevant features (the target variable is independent of the these features), which of the following statement is TRUE? *1 point*

- Random Forests have reduced performance as the fraction of irrelevant features increases.
- Random forests have increased performance as the fraction of irrelevant features increases.
- The fraction of irrelevant features doesn't impact the performance of random forest.

Yes, the answer is correct.
Score: 1

Accepted Answers:
Random Forests have reduced performance as the fraction of irrelevant features increases.

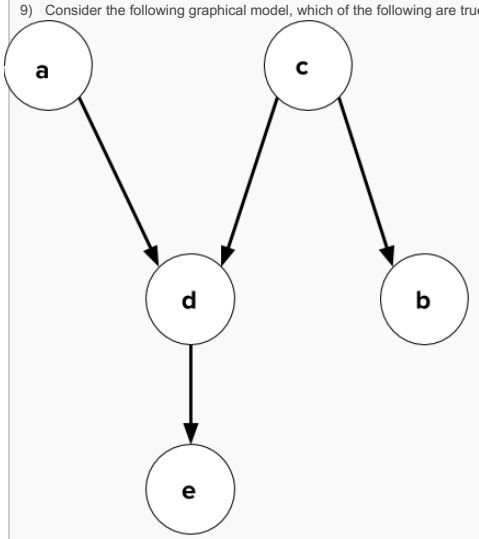
8) Which of the following statements are true about ensemble classifiers? ***1 point***
[Note: Multiple options may be correct]

- [x] The different learners in boosting based ensembles can be trained in parallel
- [x] The different learners in bagging based ensembles can be trained in parallel
- [x] Boosting based algorithms which iteratively re-weight training points, such as AdaBoost, are more sensitive to noise than bagging based methods.
- [ ] Boosting methods generally use strong learners as individual classifiers
- [x] Boosting methods generally use weak learners as individual classifiers.
- [ ] An individual classifier in a bagging based ensemble is trained with every point in the training set
- [x] An individual classifier in a boosting based ensemble is trained with every point in the training set.

No, the answer is incorrect.
Score: 0

Accepted Answers:
The different learners in bagging based ensembles can be trained in parallel
Boosting based algorithms which iteratively re-weight training points, such as AdaBoost, are more sensitive to noise than bagging based methods.
Boosting methods generally use weak learners as individual classifiers.
An individual classifier in a boosting based ensemble is trained with every point in the training set.

9) Consider the following graphical model, which of the following are true about the model? ***1 point***



[Note: multiple options may be correct]

- [x] d is independent of b when c is known
- [ ] a is independent of c when e is known
- [ ] a is independent of b when e is known
- [x] a is independent of b when c is known

Yes, the answer is correct.
Score: 1

Accepted Answers:
d is independent of b when c is known
a is independent of b when c is known

10) You are faced with a five class classification problem, with one class being the class of interest, i.e., you care more about correctly classifying data points belonging to that one class than the others. You are given a data set to use as training data. You analyze the data and observe the following properties. Which among these properties do you think are unfavorable from the point of view of applying general machine learning techniques? ***1 point***

- ( ) All data points come from the same distribution
- (•) There is class imbalance with much more data points belonging to the class of interest than the other classes
- ( ) The given data set has some missing values
- ( ) Each data point in the data set is independent of the other data points

No, the answer is incorrect.
Score: 0

Accepted Answers:
The given data set has some missing values

Previous Page End

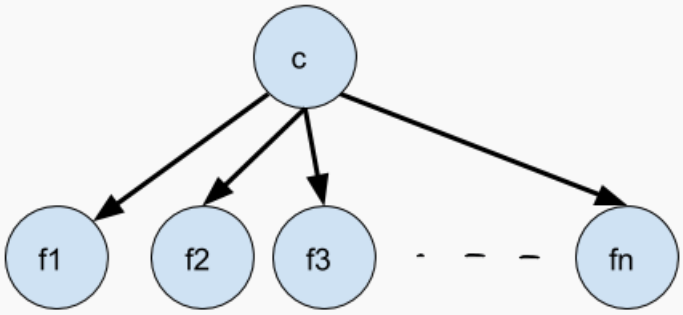
Unit 11 - Week 9

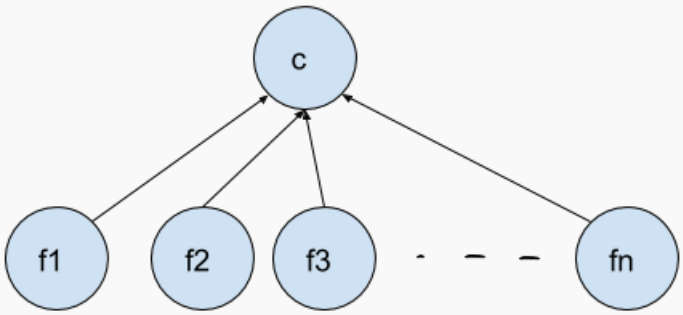
Assignment 9

The due date for submitting this assignment has passed. **Due on 2018-03-28, 23:59 IST.**

Assignment submitted on 2018-03-28, 23:52 IST

1) Which of the following graphical models capture the Naive Bayes assumption, where c represents the class label and f_i are the features? *1 point*

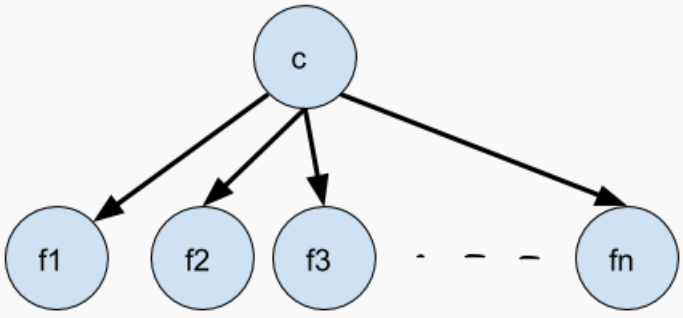




- It cannot be captured by a graphical model.
- Graphical model can capture the assumption, but the given models don't do it.

Yes, the answer is correct.
Score: 1

Accepted Answers:



2) Select the correct pair of graphs and their implied independence results. *2 points*
(Note 1: Arrows represent dependence as in normal Bayesian Networks.
Note 2: More than one options may be correct)

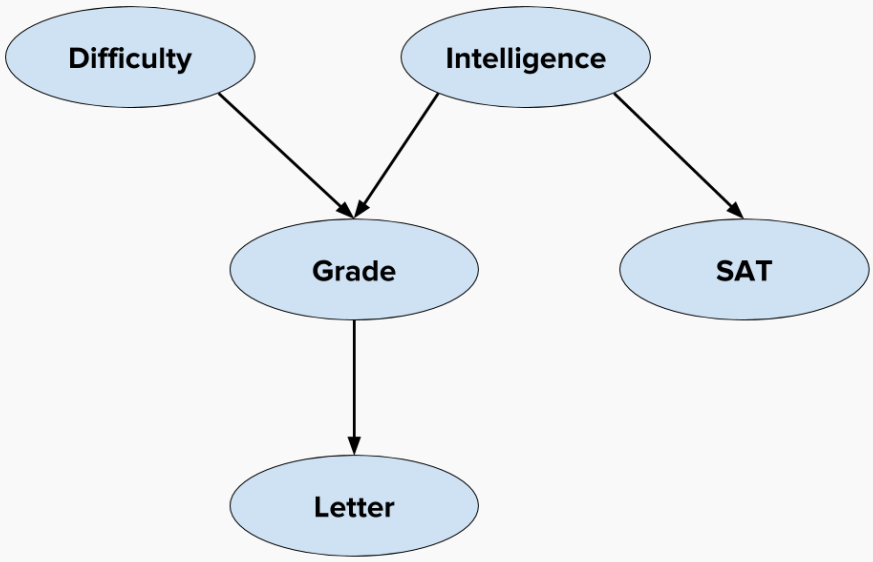
- $A \rightarrow C \leftarrow B$, implies A is independent of B given C

- [x] $A \rightarrow C \leftarrow B$, implies A depends on B if C is known.
- [x] $A \rightarrow C \rightarrow B$, implies B is independent of A if C is known.
- [x] $A \leftarrow C \rightarrow B$, implies A is independent of B given C.

Yes, the answer is correct.
Score: 2

Accepted Answers:
$A \rightarrow C \leftarrow B$, implies A depends on B if C is known.
$A \rightarrow C \rightarrow B$, implies B is independent of A if C is known.
$A \leftarrow C \rightarrow B$, implies A is independent of B given C.

3) Here is a popular toy graphical model. It models the grades obtained by a student in a course and it's implications. Difficulty represents the difficulty of the course and intelligence is an indicator of how intelligent the student is, SAT represents the SAT scores of the student and Letter presents the event of the student receiving a letter of recommendation from the faculty teaching the course. ***2 points***



Given this graphical model, which of the following statements are true?
(Note - More than one can be correct.)

- [x] Given the grade, difficulty and letter are independent variables.
- [ ] Given grade, difficulty and intelligence are independent
- [x] Without knowing any information, Difficulty and Intelligence are independent.
- [x] Given the intelligence, SAT and grades are independent.

Yes, the answer is correct.
Score: 2

Accepted Answers:
Given the grade, difficulty and letter are independent variables.
Without knowing any information, Difficulty and Intelligence are independent.
Given the intelligence, SAT and grades are independent.

4) The random variables given in the previous model are modeled as discrete variables and the corresponding CPDs are as below. ***1 point***

d^0	d^1
0.6	0.4

i^0	i^1
0.6	0.4

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.2	0.8
g^2	0.4	0.6
g^3	0.99	0.01

What is the probability of i^1, d^0, g^2, s^1, l^0 occurring?

- 0.004608
- 0.006144
- 0.003992
- 0.007309

Yes, the answer is correct.
Score: 1

Accepted Answers:
0.006144

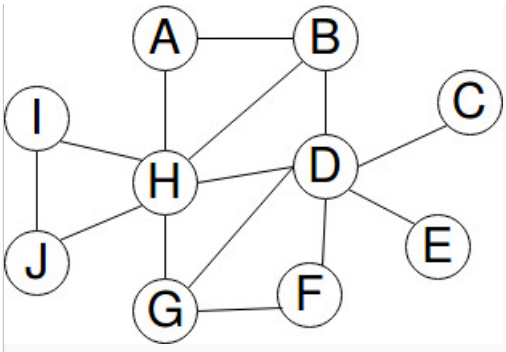
5) Using the given example and CPD's compute the probability of following assignment, i^1, g^1, s^1, l^1 irrespective of the difficulty of the course? ***1 point***

- 0.160
- 0.371
- 0.662
- 0.189

Yes, the answer is correct.
Score: 1

Accepted Answers:
0.189

6) Consider the following Markov Random Field. ***2 points***



Which of the following nodes will have no effect on D given the Markov Blanket of D?
(Note: more than one options may be correct)

- [x] A
- [ ] B
- [ ] C
- [ ] E
- [ ] F
- [ ] G
- [ ] H
- [x] I
- [x] J

Yes, the answer is correct.
Score: 2

Accepted Answers:
A
I
J

7) Select the correct pairs of (Graphical Model, Inference Algorithm) ***1 point***
(note: more than one option may be correct)

- [x] (Bayesian Networks, Variable Elimination)
- [ ] (Viterbi Algorithm, Markov Random Fields)
- [x] (Viterbi Algorithm, Hidden Markov Models)
- [x] (Belief Propagation, Markov Random Fields)
- [x] (Variable Elimination, Markov Random Fields)

Yes, the answer is correct.
Score: 1

Accepted Answers:
(Bayesian Networks, Variable Elimination)
(Viterbi Algorithm, Hidden Markov Models)
(Belief Propagation, Markov Random Fields)
(Variable Elimination, Markov Random Fields)

Previous Page

End

Unit 12 - Week 10

Assignment 10

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2018-04-04, 23:59 IST.

1) In the CURE clustering algorithm, representative points of a cluster are moved a fraction of the distance between their original location and the centroid of the cluster. Would it make more sense to move them all a fixed distance towards the centroid instead? Why or why not? ***1 point***

- ○ Yes, because this approach will ensure that the original cluster shape is preserved.
- ○ No, because this approach will not be as effective against outliers as the original approach.

No, the answer is incorrect.
Score: 0

Accepted Answers:
No, because this approach will not be as effective against outliers as the original approach.

2) If you are trying to find clusters in a dataset of roughly 5 billion data points (each data point taking roughly 128 bytes of data) and you have a machine with 2TB of RAM. Considering the large data size, which of the following algorithms would you use to efficiently find patterns in the data? ***1 point***
(Note: More than one option may be correct)

- ☐ BIRCH
- ☐ CURE
- ☐ K-means
- ☐ Logistic Regression

No, the answer is incorrect.
Score: 0

Accepted Answers:
BIRCH
CURE

3) In k-means clustering, globally minimizing the objective function for a known k is: ***1 point***

- ○ NP hard
- ○ impossible
- ○ possible in linear time
- ○ possible in polynomial time

No, the answer is incorrect.
Score: 0

Accepted Answers:
NP hard

4) What assumption does the CURE clustering algorithm make with regards to the shape of the clusters? ***1 point***

- ○ No assumption
- ○ Spherical
- ○ Elliptical

No, the answer is incorrect.
Score: 0

Accepted Answers:
No assumption

5) What would, in general, be the effect of increasing MinPts in DBSCAN while retaining the same Eps parameter? ***1 point***
(Note that more than one statement may be correct)

- ☐ Increase in the sizes of individual clusters
- ☐ Decrease in the sizes of individual clusters
- ☐ Increase in the number of clusters
- ☐ Decrease in the number of clusters

No, the answer is incorrect.

Score: 0

Accepted Answers:
Decrease in the sizes of individual clusters
Increase in the number of clusters

6) In the following three questions we would like to understand the utility of different clustering algorithms on particular dataset. Kindly download DS1 and DS2. The first two columns in the dataset correspond to the co-ordinates of each data point. The third column corresponds two the actual cluster label. ***1 point***

Visualize the dataset DS1. Which of the following algorithms will be able to recover the true clusters?

- K-means clustering
- Single link hierarchical clustering
- Complete link hierarchical clustering
- None of the above

No, the answer is incorrect.
Score: 0

Accepted Answers:
Single link hierarchical clustering

7) Visualize the dataset DS2. Which of the following algorithms will be able to recover the true clusters. ***1 point***

- K-means clustering
- Single link hierarchical clustering
- Complete link hierarchical clustering
- None of the above

No, the answer is incorrect.
Score: 0

Accepted Answers:
None of the above

8) For the dataset DS1, compute the Rand Index for the following methods: K-means, Single link hierarchical clustering and Complete link hierarchical clustering.
which of the method will return clustering with maximum Rand Index? ***1 point***

- K-means clustering
- Single link hierarchical clustering
- Complete link hierarchical clustering
- All the above will return the same clusters and hence have equal Rand Index

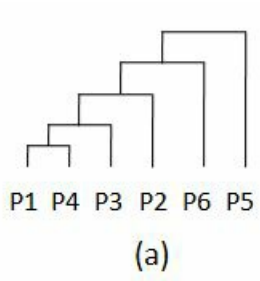
No, the answer is incorrect.
Score: 0

Accepted Answers:
Single link hierarchical clustering

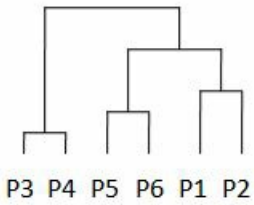
9) Consider the similarity matrix given below: ***1 point***

	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

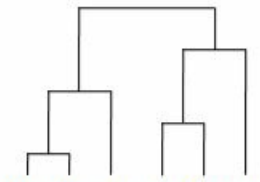
Which of the following shows the hierarchy of clusters created by the single link clustering algorithm?



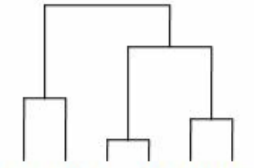
(a)



(b)



(c)



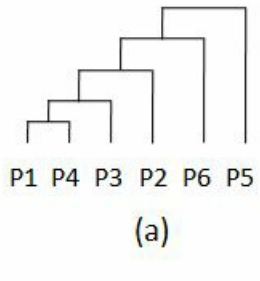
(d)

- (a)
- (b)
- (c)
- (d)

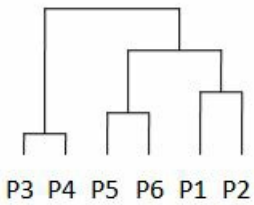
No, the answer is incorrect.
Score: 0

Accepted Answers:
(b)

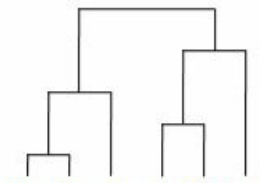
10) For the similarity matrix given in the previous question, which of the following shows the hierarchy of clusters created by the complete link clustering algorithm. ***1 point***



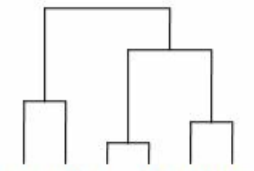
(a)



(b)



(c)



(d)

- (a)
- (b)
- (c)
- (d)

No, the answer is incorrect.
Score: 0

Accepted Answers:
(d)

Previous Page

End

Unit 13 - Week 11

Assignment 11

The due date for submitting this assignment has passed. **Due on 2018-04-11, 23:59 IST.**

Assignment submitted on 2018-04-11, 23:03 IST

1) Given N samples $x_1, x_2, \ldots, x_N$ drawn independently from a Gaussian distribution with variance σ and unknown mean μ, find the MLE of the mean. ***1 point***

- $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{\sigma^2}$
- $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{2\sigma^2 N}$
- $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$
- $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N-1}$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

2) Suppose we are trying to model a p dimensional Gaussian distribution. What is the actual number of independent parameters that need to be estimated? ***2 points***

- 2
- 2p
- p(p+1)
- p(p+3)/2

Yes, the answer is correct.
Score: 2

Accepted Answers:
p(p+3)/2

3) You are given n p-dimensional data points. The task is to learn a classifier to distinguish between k classes. You come to know that the dataset has missing values. Can you use EM algorithm to fill in the missing values ? (without making any further assumptions) ***2 points***

- Yes
- No

No, the answer is incorrect.
Score: 0

Accepted Answers:
No

4) During parameter estimation for a GMM model using data X, which of the following quantities are you minimizing (directly or indirectly)? ***1 point***

- Log-likelihood
- Negative Log-likelihood
- Cross-entropy
- Residual Sum of Squares (RSS)

Yes, the answer is correct.
Score: 1

Accepted Answers:
Negative Log-likelihood

5) In Gaussian Mixture Models, π_i are the mixing coefficients. Select the correct conditions that the mixing coefficients need to satisfy for a valid GMM model. (More than one option may be correct) ***2 points***

- [x] $0 \leq \pi_i \leq 1 \forall i$
- [ ] $-1 \leq \pi_i \leq 1 \forall i$
- [x] $\sum_i \pi_i = 1$
- [ ] $\sum_i \pi_i$ need not be bounded

Yes, the answer is correct.
Score: 2

Accepted Answers:
$0 \leq \pi_i \leq 1 \forall i$
$\sum_i \pi_i = 1$

6) Expectation-Maximization, or the EM algorithm, consists of two steps - E step and the M-step. Using the following notation, select the correct set of equations used at each step of the algorithm. (More than one option may be correct) ***2 points***

Notation:
$\mathbf{X}$− Known/Given variables/data
$\mathbf{Z}$− Hidden/Unknown variables
θ− Total set of parameters to be learned
$\theta_{\mathbf{k}}$− Values of all the parameters after stage k
$\mathbf{Q}(\cdot, \cdot)$− The Q-function as described in the lectures

- [x] $\mathbf{E} - \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_{m-1}}[log(Pr(\mathbf{X},\mathbf{Z}|\theta))]$
- [ ] $\mathbf{E} - \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta}[log(Pr(\mathbf{X},\mathbf{Z}|\theta_m))]$
- [x] $\mathbf{M} - argmax_\theta \sum_{\mathbf{Z}} Pr(\mathbf{Z}|\mathbf{X},\theta_{m-1}) \cdot log(Pr(\mathbf{X},\mathbf{Z}|\theta))$
- [x] $\mathbf{M} - argmax_\theta \mathbf{Q}(\theta, \theta_{m-1})$
- [ ] $\mathbf{M} - argmax_\theta \mathbf{Q}(\theta, \theta_{m-2})$

Yes, the answer is correct.
Score: 2

Accepted Answers:
$\boldsymbol{E} - E_{\boldsymbol{Z}|\boldsymbol{X},\theta_{m-1}}[log(Pr(\boldsymbol{X},\boldsymbol{Z}|\theta))]$
$\boldsymbol{M} - argmax_\theta \sum_{\boldsymbol{Z}} Pr(\boldsymbol{Z}|\boldsymbol{X},\theta_{m-1}) \cdot log(Pr(\boldsymbol{X},\boldsymbol{Z}|\theta))$
$\boldsymbol{M} - argmax_\theta \boldsymbol{Q}(\theta, \theta_{m-1})$

Previous Page

End

Unit 14 - Week 12

Assignment 12

The due date for submitting this assignment has passed. **Due on 2018-04-18, 23:59 IST.**

Assignment submitted on 2018-04-18, 12:05 IST

1) In a tournament classifier with N classes. What is the complexity of the number of classifiers we require? ***1 point***

- $\mathcal{O}(N)$
- $\mathcal{O}(N^2)$
- $\mathcal{O}(N.log(N))$
- $\mathcal{O}(log(N))$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\mathcal{O}(N)$

2) In the context of Reinforcement Learning algorithms, which of the following definitions constitutes a valid Markov State? (More than one option may be correct) ***2 points***

- [x] For Chess: Positions of yours and the opponent's remaining pieces
- [x] For Tic-Tac-Toe: A snapshot of the game board (all Xs, Os and empty spaces)
- [ ] For Chess: Positions of your pieces and the identities of the opponents defeated pieces.
- [x] For Tennis: Position and Velocity of the ball
- [ ] For Tennis: Position of the ball

Yes, the answer is correct.
Score: 2

Accepted Answers:
For Chess: Positions of yours and the opponent's remaining pieces
For Tic-Tac-Toe: A snapshot of the game board (all Xs, Os and empty spaces)
For Tennis: Position and Velocity of the ball

3) 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? ***1 point***

- +5 for reaching the finish line, -1 for going off the road
- (•) +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line
- +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line, +1 for going off the road
- -5 for reaching the finish line, +0.1 for every second that passes before the agent reaches the finish line.

Yes, the answer is correct.
Score: 1

Accepted Answers:
+5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line

4) Recall the tic-tac-toe example from the reinforcement learning lecture. For each board position, we maintain the probability of winning from that board position. These board positions are updated using temporal difference learning. Assume that the probability values are all initialized to 0.5 and the opponent is an imperfect player. Suppose we always select the greedy action, i.e., the action which leads to the next state with highest probability. What problem(s) can you expect to encounter following this strategy? ***2 points***

- It may happen that we never win even once.
- No problems, this is an optimal strategy.
- If a path exists in the game tree that always leads to victory, such a path can never be found using this strategy.
- It is possible that we never fully explore the entire game tree.

No, the answer is incorrect.

Score: 0

Accepted Answers:
It is possible that we never fully explore the entire game tree.

5) Suppose we want an RL agent to learn to play the game of golf. For training purposes, we make use of a golf simulator program. Assume that the original reward distribution gives a reward of +10 when the golf ball is hit into the hole and -1 for all other transitions. To aide the agent's learning process, we propose to give an additional reward of +3 whenever the ball is within a 1 metre radius of the hole. Is this additional reward a good idea or not? Why? ***1 point***

- Yes. The additional reward will help speed-up learning.
- Yes. Getting the ball to within a metre of the hole is like a sub-goal and hence, should be rewarded.
- No. The additional reward may actually hinder learning.
- No. It violates the idea that a goal must be outside the agent's direct control.

No, the answer is incorrect.
Score: 0

Accepted Answers:
No. The additional reward may actually hinder learning.

6) You want to toss a fair coin a number of times and obtain the probability of it falling on it's heads by taking a simple average. What is the estimated number of times you'll have to toss the coin to make sure that your estimated probability is within 10% of the actual probability, at least 90% of the time? (Please note that $ln(x)$ denotes $\log_e(x)$) ***2 points***

- 400*ln(20)
- 800*ln(20)
- 200*ln(20)
- 100*ln(40)

No, the answer is incorrect.
Score: 0

Accepted Answers:
*200*ln(20)*

7) You face a particularly challenging RL problem, where the reward distribution keeps changing with time. In order to gain maximum reward in this scenario, does it make sense to stop exploration or continue exploration? ***1 point***

- Stop exploration
- Continue exploration

No, the answer is incorrect.
Score: 0

Accepted Answers:
Continue exploration

Previous Page

End