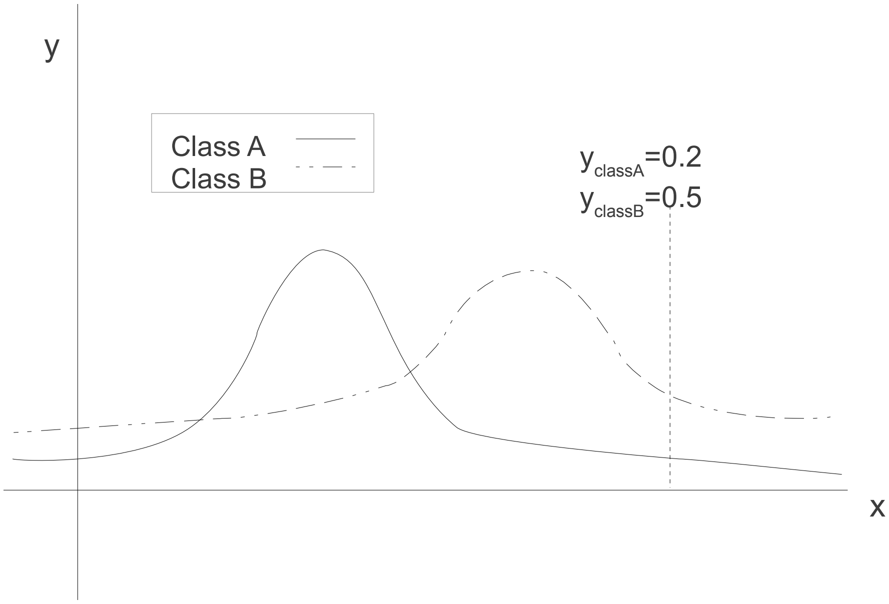
CIS501 Final exam, Fall 2012

*Answer all questions. Unless stated otherwise, select a single **best** answer to each question. Questions carry one mark each unless marked otherwise.*

1. Broadly speaking there are two ways in which complexity can be controlled - by constraining the data, and by constraining the model. For the second option, which amongst these is NOT a valid approach:
   a. Reducing the number of parameters in the model
   b. Limiting the size of the model
   c. Limiting the magnitude of model parameter values
   d. Using second order information when configuring the model parameters
   e. Removing elements of the model which do not contribute to performance
2. In terms of “constraining the data”, one approach is to reduce the dimensionality of the input space. The following are benefits of doing this EXCEPT for:
   a. Reducing the dimensionality simply reduces the amount of data to be processed, thus saving computational resources
   b. Removing redundant channels can help to remove noise from the data
   c. By reducing the parameter space, reducing the dimensionality can help to avoid overfitting
   d. Reducing the dimensionality can help to improve the validity and reliability of distance measurements
   e. All are valid reasons

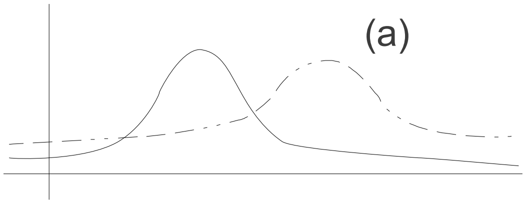


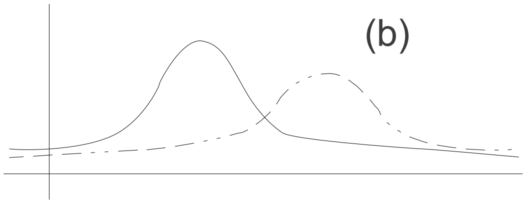
The curves depict the *likelihoods* for classes A and B respectively. Also, you are given that $p(A)=3\times p(B)$

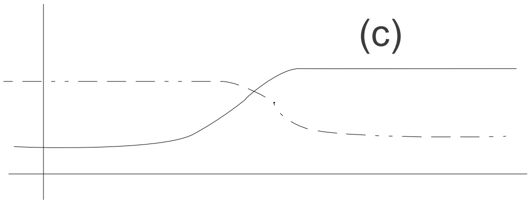
Based on this figure, please answer the following *two* questions

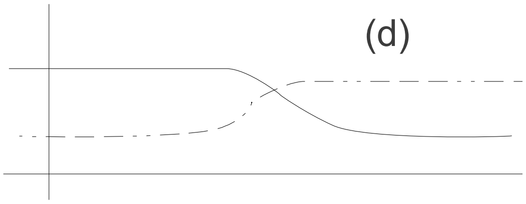
3. At the point depicted by the vertical dotted line, what are the values of the *posterior distribution* for classes *A* and *B*?
    a. 0.15,0.125
    b. 0.20,0.50
    c. 0.55,0.45
    d. 0.75,0.25
    e. 0.90,0.10

4. Approximately, what would the curves of the posterior distributions looks like?

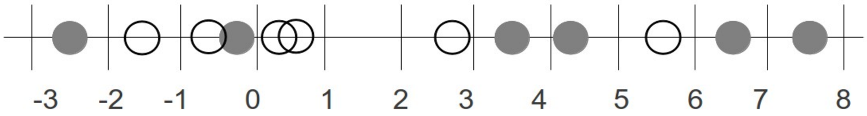








5. Consider the following three applications: (i) Fraud detection (ii) Disease diagnosis (iii) Web search – for each one, which is *more* important, Precision or Recall?
    a. (i) Precision (ii) Recall (iii) Recall
    b. (i) Precision (ii) Recall (iii) Precision
    c. (i) Precision (ii) Precision (iii) Recall
    d. (i) Recall (ii) Precision (iii) Recall
    e. (i) Recall (ii) Recall (iii) Precision
6. The non-disjoint and lazy discretization methods are significant improvements over equal width discretization (i.e. the "histogram method"). Why is this?
    a. They place the point of interest at (or near) the center of the bin used to perform density estimation
    b. They give better density estimates in regions with low data density
    c. They have lower computational requirements
    d. They allow contributions from all points in the data set to be taken into account
    e. They work well for sparse data sets
7. Consider the following distribution of values taken from the training set with a single continuous feature:



    When designing a decision node in a tree, which of the following suggested thresholds would be a reasonable choice:
    a. -2
    b. -1
    c. +1
    d. +2
    e. +3
8. The following procedures are valid methods of pruning a decision tree EXCEPT for:
    a. Stopping training when the number of training instances in a particular node drops below a threshold
    b. Stopping training when the depth of a tree exceeds a particular threshold

c. Deleting nodes where the information gain fails to exceed a particular threshold.

d. Deleting nodes that do not improve (or only result in small improvements) to the test error

e. None of the above

The following three questions address the following **(feature:label)** *pairs:*

[(0:'A'),(0:'A'),(0:'A'),(0:'B'),(1:'B'),(1:'B'),(1:'B'),(1:'B')]

9. calculate the information gain (IG), using base-2 logarithms:

a. 0.14

b. 0.23

c. 0.31

d. 0.45

e. 0.55

10. For the same set of pairings, please calculate the Gini impurity index:

a. 1/16

b. 2/16

c. 3/16

d. 4/16

e. 5/16

11. Let's say that this is one feature (let's call it “F1”) in a Naive Bayes Classifier. What is P(F1|”B”)?

a. 1/5

b. 2/5

c. 3/5

d. 4/5

e. 1

The following description applies to the next two questions

The following two are expressions designed for evaluating cluster quality. For simplicity, they are written for the two cluster case:

$$\frac{\sum_{i\in c_1}\sum_{j\in c_2} d_{ij}}{\sum_{k,l\in c_1} d_{kl} + \sum_{m,n\in c_2} d_{mn}} \quad , \quad \frac{\min_{i\in c_{1,}\, j\in c_2} d_{ij}}{max\left\{\max_{k,l\in c_1} d_{kl},\ \max_{m,n\in c_2} d_{mn}\right\}}$$

Given the following:

d_{ij} → Distance between points *i* and *j*

c_h → set of points in cluster *h* (for simplicity, h∈{1,2})

12. For *better* clustering performance, we would want the two measures to be:
    a. Lower, Lower
    b. Lower, Higher
    c. Higher, Lower
    d. Higher, Higher
    e. Unable to tell
13. For the second expression, what is the purpose of applying the *max* operations (in the denominator)?
    a. So as to minimize the *worst case distances* between points in the same clusters.
    b. So as to maximize the *worst case distances* between points in the same clusters.
    c. So as to minimize the *best case distances* between points in the same clusters.
    d. So as to maximize the *best case distances* between points in the same clusters.
    e. None of the above
14. The "C-index" for evaluating clusterings is defined as:

$$C=\frac{S-S_{min}}{S_{max}-S_{min}}$$

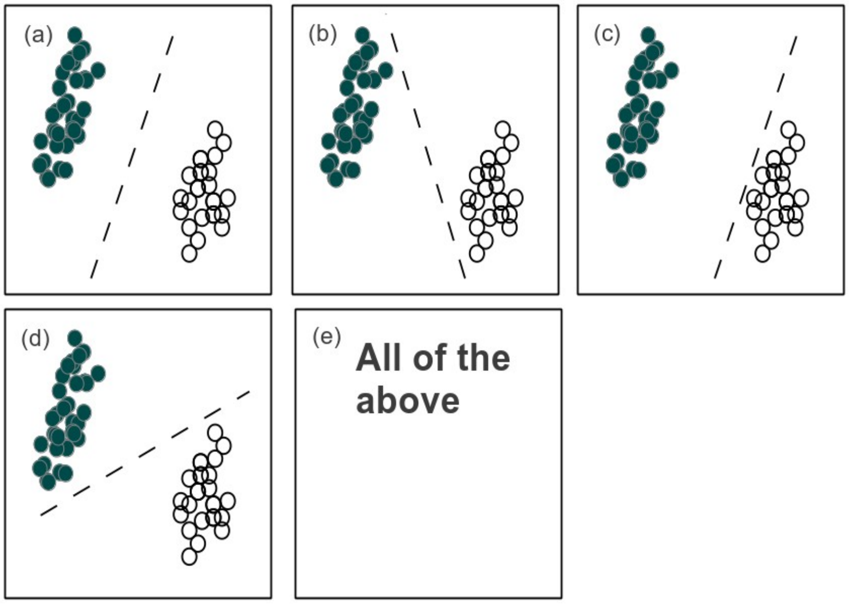
where a smaller value of *C* is "better". For the following two clusters of points, what is the value of the C-index:

(1,2,3),(6,7,8)

   a. 0
   b. 0.2
   c. 0.4
   d. 0.6
   e. 0.8
   f. 1

15. The following figures depict decision boundaries induced by a single perceptron. The task is a two class classification problem, where the training data points have been clearly labeled using filled and empty circles.

Which of the figures denote a “valid” boundary – i.e. one corresponding to a converged perceptron, assuming that the perceptron learning rule was used.



16. If you wanted to use gradient descent to train the standard perceptron, what modification is required?
    a. Use a “smooth” activation function
    b. Addition of a hidden layer
    c. Addition of hidden units
    d. Removal of bias term
    e. No modification is required

The passage below applies to the following three questions

The “error” term for a particular optimization problem takes the following expression:

$$E(w_{1,}w_2)=w_1^2+2\mathrm{w}_2^2-w_1$$

where, w_1 and w_2 are the parameters to be optimized.

17. What is the minimum value of the error function *E*, and at what values of w_1 and w_2 does this occur?
    a. -0.22,1/3,0
    b. -0.22,2/3,0
    c. -0.22,0,1/3
    d. -0.25,0,0.5

e. -0.25,0.5,0

18. Iterative techniques are frequently used to solve optimization problems. You decide to use gradient descent with the initial parameter values of

(w_1,w_2) =(1,1)

and a learning rate of 0.1. What is the initial update term?

a. (0,-0.2)

b. (-0.1,-0.2)

c. (-0.1,-0.4)

d. (-0.2,-0.4)

e. (-0.3,-0.6)

19. Does this point to the minimum of the error function? Why is this?

a. Yes, the vector gradient term indicates the direction of steepest descent, which will lead to the minimum

b. Yes, the error surface is convex and has no local minima, and so gradient descent leads right to the minimum of the error function

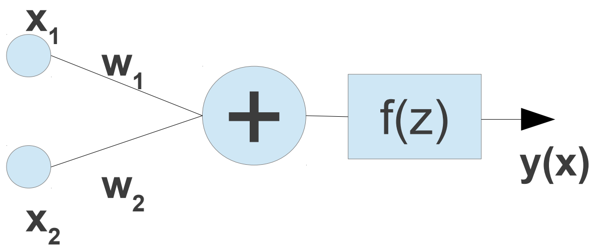
c. No, the path to the minimum is obstructed by numerous local minima, hence gradient descent does not provide an optimum path

d. No, the error surface has been “stretched”, resulting in w_2 being over-emphasized by the gradient function.

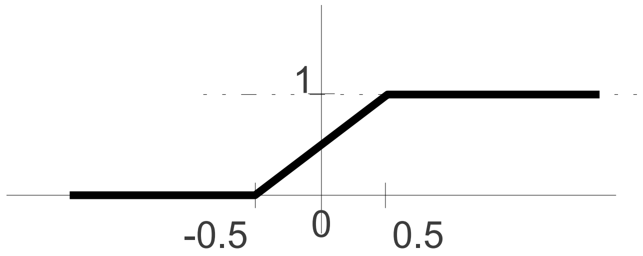
e. None of the above

<u>The description below applies to the following three questions</u>

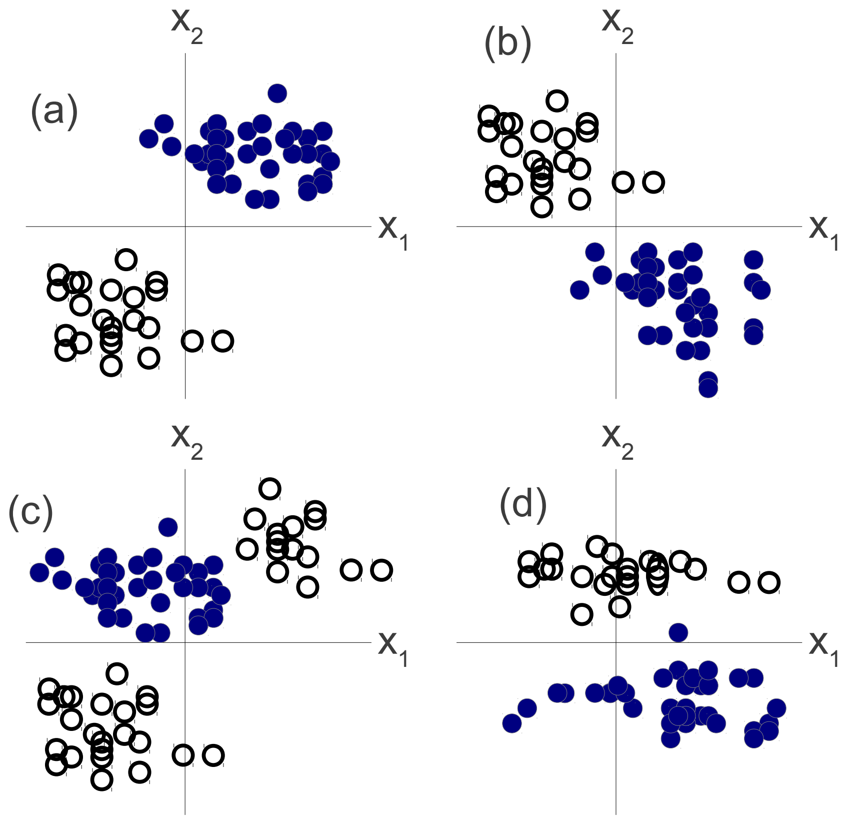
Consider the following network architecture:



Where, the activation function f(z) is depicted by:



20. If x_1=1, x_2=2, w_1=w_2=0.5, what is the network's output?
    a. 0.5
    b. 1
    c. 1.5
    d. 2
    e. 2.5
21. Assuming that the target output is 0, what is the gradient of the error function w.r.t. w_2?
    a. 0
    b. 0.5
    c. 1
    d. 1.5
    e. 2
22. Which of the following classification problems can be appropriately handled by this particularly network?



23. Consider the following three descriptions:

(i) "Finding clusters by combining individual elements into larger collections"

(ii) "A form of learning where only parameters related to large inputs and erroneous outputs are changed"

(iii) "A process by which input data points are matched with network weights"

Which three terms match the descriptions above:

a. Divisive clustering; Hebbian Learning; Projection

b. Agglomerative clustering; Gradient descent optimization; Distance Calculation

c. Agglomerative clustering; Hebbian Learning; Projection

d. Divisive clustering; Gradient descent optimization; Projection

e. Divisive clustering; Hebbian Learning; Distance Calculation

<u>This description applies to the following two questions:</u>

The following is the distance matrix between a set of five points:

	1	2	3	4	5
1	0	1	3.5	5	6
2	1	0	2.2	4	5
3	3.5	2.2	0	2	3.3
4	5	4	2	0	1
5	6	5	3.3	1	0

24. You decide to use the k-centers (also known as k-medoids) algorithm to cluster these points. Set points 1 and 2 to be the initial centroids. What will the initial clustering be?

    a. (1),(2,3,4,5)

    b. (1,3),(2,4,5)

    c. (1,3,4),(2,5)

    d. (1,3,4,5),(2)

    e. (1,3,5),(2,4)

25. At some point during a run of the k-centers algorithm, we are left with the following two clusters:

    $\mathbf{C_1}$:(1,2,3,4) $\mathbf{C_2}$:(5)

    For cluster $\mathbf{C_1}$, what should be the centroid be?

    a. 1

    b. 2

    c. 3

    d. 4

    e. 5

26. Which of the following is a valid reason to choose the k-centers algorithm over k-means:

    a. Works better with high dimensional data

    b. Lower computational complexity

    c. Lower memory requirements

    d. Better able to handle missing data points

    e. Better resistance to outliers

27. The self organizing map is similar to the *k*-means clustering algorithm in some ways, but is different in others. Name one similarity and one difference:

a. **S**: Nodes (or centroids) strive to be representative of a group of data points; **D**: The nodes of a SOM are topologically linked

b. **S**: Nodes (or centroids) strive to be representative of a group of data points; **D**: SOM is trained using an iterative rule

c. **S**: Both methods provide nonlinear visualization capabilities; **D**: The nodes of a SOM are topologically linked

d. **S**: Both methods provide nonlinear visualization capabilities; **D**: SOM is trained using an iterative rule

e. **S**: Both methods provide nonlinear visualization capabilities; **D**: SOM can work in multiple dimensions

28. When training the SOM, the learning rate and neighborhood size need to be gradually:

a. Increased; Increased

b. Increased; Decreased

c. Decreased; Increased

d. Decreased; Decreased

e. None of the above

29. With reference to the previous question, why do these parameters need to be changed?

a. Early in the training, the map may be overfitting the data, while later on this is no longer a problem

b. Early in the training, the map may be underfitting the data, while later on this is no longer a problem

c. Early in the training the parameters have very low variability and as such need a higher degree of tuning

d. Early in the training, the map needs to be “fine-tuned”, while later on it simply needs to be “folded”.

e. Early in the training, the map needs to be “unfolded”, while later on, it needs to be “fine-tuned”

30. Consider the following three statements:

(i) “Principle Component Analysis works by finding the subspace which contains the largest proportion of the variance in the data”

(ii) “Provided that there are no local minima, the direction of steepest descent gives the most direct path to the global minimum”

(iii) “The Levenberg-Marquardt algorithm gives better performance than backpropagation by using second order information about the error surface”

Indicate whether each statement is “true” or “False”:

a. True; True; True
b. False; True; True
c. True; False; True
d. True; True; False
e. False; False; False