# CIS501 Final Exam, Fall 2011

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

1. One way of handling missing data is by deleting incomplete records. However, this method is not very effective except in very simple situations. From the following list please select **two** reasons for this:
    a. Removing the records in this way can introduce noise into the system
    b. Data is often valuable/limited – deleting records is wasteful
    c. It is not possible to reduce the dimensionality of the dataset in this way
    d. Doing so may skew the data in cases where a particular field only applies for a specific subset of the data
    e. A better approach is to remove the missing fields
2. Which of the following sets of feature ↔ label pairings has the highest Gini impurity index?
    a. [(0,1),(0,0),(1,1),(0,1),(1,0)]
    b. [(0,0),(0,1),(1,0),(0,0),(1,1)]
    c. [(0,1),(0,0),(0,1),(1,0),(1,0)]
3. The C4.5 algorithm improves upon ID3 by incorporating which of the following elements into its learning process? (select **two**)
    a. Regularization
    b. Information Gain
    c. Smoothing
    d. Model selection
    e. Continuous variables
4. The following are two multinomial distributions over the following 5-word vocabulary {password,credit,investment,slimming,supplements}:

    $P_{nospam}$=[0.2,0.3,0.3,0.1,0.1]

    $P_{spam}$=[0.2,0.3,0.1,0.2,0.2]

    Using the multinomial event model, what are the probabilities P(*nospam*|D) and P(*spam*|D) for the following passage:

    D → "*.. for a month's free supply of XYZ supplements please provide your credit card details now. It will be the best investment .. all that is needed is a valid credit card..*"

    a. 0.009,0.006
    b. 0.0027,0.006

c. 0.0027,0.0018

d. 0.021,0.055

e. None of the above

5. Consider the following sequence of octave commands:

```
M1=randn(5,5);
M2=randn(3,6);
M = [M1(:);M2(:)]'
```

Which of the following correctly recovers M1?

a. reshape(M,5,8)(1:5,1:5)

b. reshape(M',5,8)(1:5,1:5)

c. reshape(M(1:25),5,5)

d. reshape(M(1:25),5,5)'

e. None of the above

6. Which of the following is a symptom of *overfitting*

a. Low test errors but poor generalization

b. Poor noise resistance despite reasonable test error

c. Low training errors but high test errors

d. Deterioration in performance when input dimensionality is increased.

e. None of the above

7. You are building a classification system for screening ultrasound images for incidences of prostate cancer. Based on statistics obtained from the department of health, approximately 95% of images screened will be “cancer free” while 5% of the images will depict cancerous growths.

   Your classifier can return either 0 (cancer free) or 1 (cancerous). Which **two** of the following statements are true?

a. If your classifier is permanently set to “0”, it will have precision of 100%, but recall of 0%

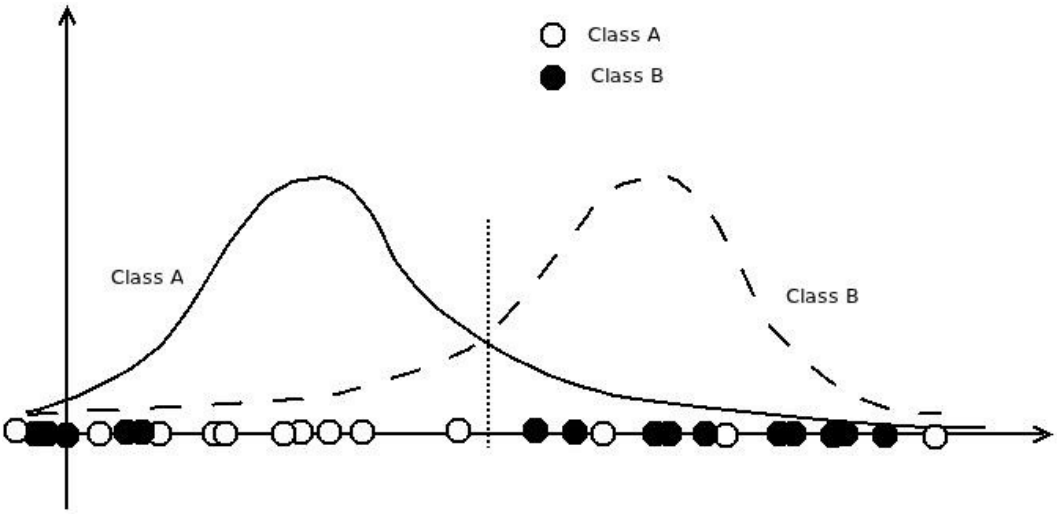
b. If your classifier is permanently set to “0”, it will have recall of 100% but precision of 0%

c. If your classifier is permanently set to “1”, it will have precision of 5% and recall of 100%

d. If your classifier is permanently set to “1”, it will have accuracy of 5% and precision of 100%

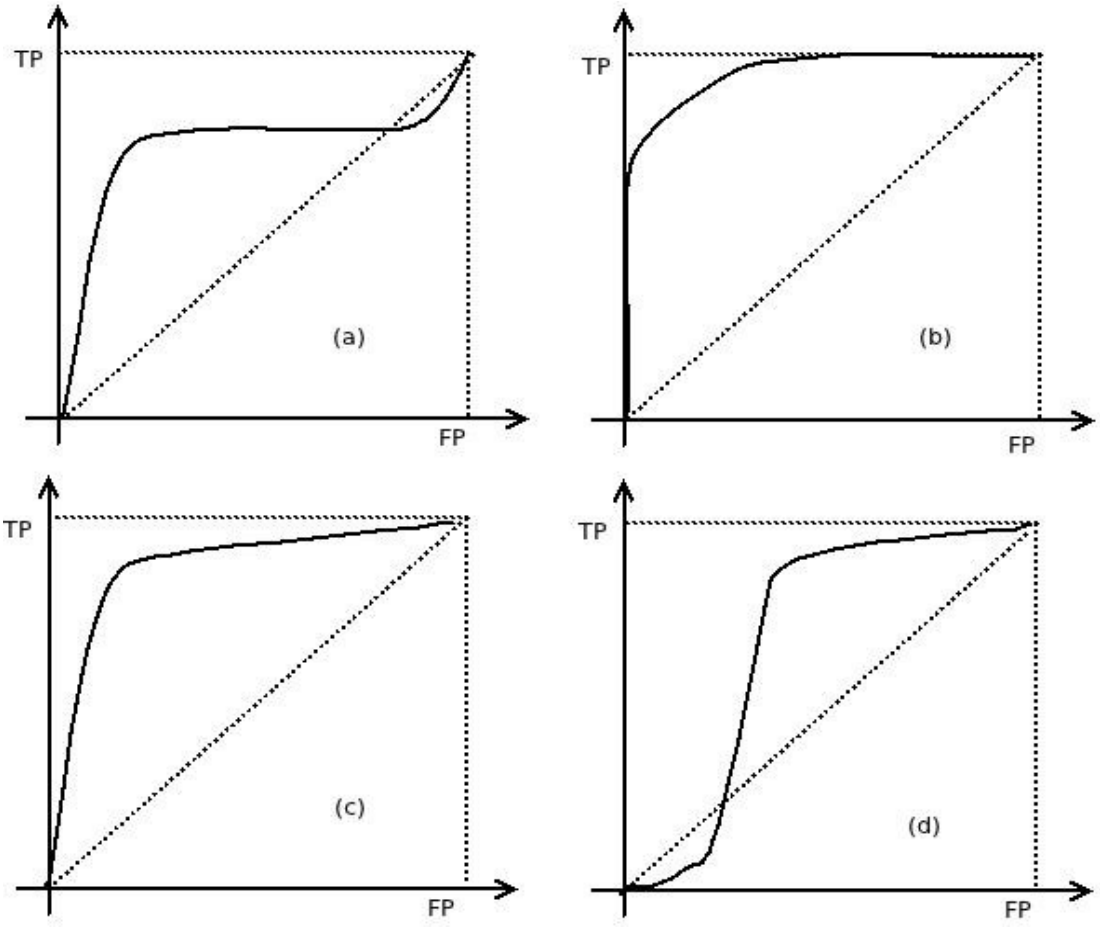
e. If your classifier is permanently set to “1”, it will have a 0.05 false positive rate.

8. The following figure depicts a 1-D data set, overlaid by the theoretical distributions of the two classes:




Suppose that the prior probability of both classes are equal, and that a *Bayes Optimal* classifier has been built.

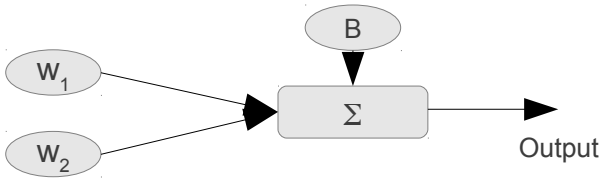
From the perspective of Class B, what would the resulting ROC curve look like? (Answer: A)



9. Look at the following truth table:

| | $x_2=0$ | $x_2=1$ |
|---|---|---|
| $x_1=0$ | 1 | 1 |
| $x_1=1$ | 0 | 1 |

You would like to model the above logical relationship using a perceptron network:




What do you think are appropriate values for the network parameters $w_1$, $w_2$ and B respectively?

a. -10,10,5

b. 10,-10,5

c. 10,10,5

d. -10,-10,5

e. -10,-10,10

10. For the network in the previous question, which of the following truth tables would be a "problem"?

a.

| | $x_2=0$ | $x_2=1$ |
|---|---|---|
| $x_1=0$ | 0 | 1 |
| $x_1=1$ | 0 | 1 |

b. √

| | $x_2=0$ | $x_2=1$ |
|---|---|---|
| $x_1=0$ | 1 | 0 |
| $x_1=1$ | 0 | 1 |

c.

| | $x_2=0$ | $x_2=1$ |
|---|---|---|
| $x_1=0$ | 1 | 0 |
| $x_1=1$ | 1 | 0 |

d.

| | $x_2=0$ | $x_2=1$ |
|---|---|---|
| $x_1=0$ | 0 | 1 |
| $x_1=1$ | 1 | 1 |

e.

| | $x_2=0$ | $x_2=1$ |
|---|---|---|
| $x_1=0$ | 1 | 1 |
| $x_1=1$ | 0 | 0 |

11. For an MLP, which of the following actions are NOT likely to help reduce the test error:
    a. Halting the training algorithm when the drop in training error starts to level off.
    b. Using as large a training set as possible
    c. Using the “optimal brain damage” algorithm to trim network weights
    d. By penalizing large weight values to keep the predicted output “smooth”
    e. They could all potentially be helpful
12. For the the following error function:

    $$E(x, y)=x^2+y^2$$

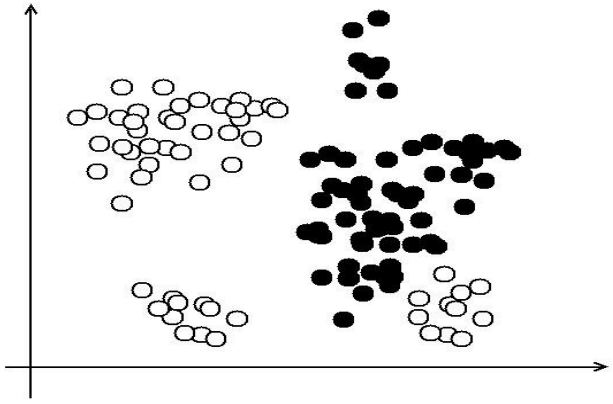
    At a certain point in time $t$, $x$=1,$y$=0.5.

    Assuming that gradient descent optimization is used, and taking the learning rate $\eta$=0.1, what is the value of $x$ and $y$ at time $t$+1?
    a. $x$=3,$y$=1.5
    b. $x$=0.8,$y$=0.4
    c. $x$=1.2,$y$=0.6
    d. $x$=-1,$y$=-0.5
    e. None of the above
13. Newton's method provides an iterative method for finding the zero crossing in a function *f(x)*, where:

    $$x_{t+1}=x_t-\frac{f(x_t)}{f'(x_t)}$$

    For the problem above, apply this technique to obtain the values of $x$ and $y$ at time $t$+1.
    a. $x$=0.8,$y$=0.4
    b. $x$=2,$y$=1
    c. $x$=-1,$y$=0
    d. $x$=0,$y$=0
    e. None of the above

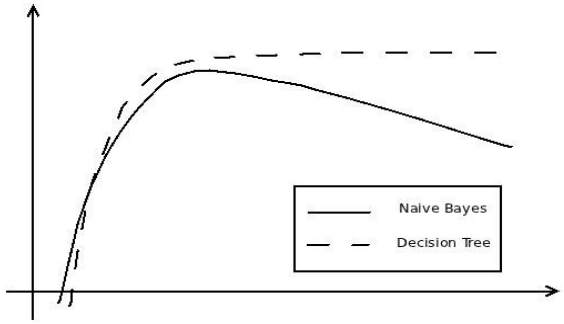
14. Why does Newton's method often provide better performance compared to basic gradient descent?
    a. By giving greater emphasis to directions in parameter space which have been stretched
    b. By giving greater emphasis to directions in parameter space that have been compressed
    c. By avoiding local minima problems
    d. By providing faster convergence
    e. None of the above
15. In the Levenberg-Marquadt algorithm, what role does $\lambda$, the damping factor, play?
    a. It determines the proportion of the previous update which is added to the current update size.
    b. It determines the extent to which the update term follows Newton's method
    c. It is a form of regularization
    d. It prevents oscillatory behaviour in the iterations, which result in slow convergence
    e. None of the above
16. Consider the following scatterplot:



    If using an MLP to distinguish between the two classes, how many hidden and output units would be required?
    a. 1,1
    b. 2,1
    c. 2,2
    d. 1,2
    e. None of the above
17. Which of the following tasks can be described as “unsupervised learning”:

a. Learning to detect when a child needs to see a doctor
b. Learning when to buy/sell a stock
c. Learning to swim
d. Learning which oranges are ripe in a supermarket
e. Learning the colours of the rainbow

18. The following is a graph of classification accuracy vs. number of input features for two popular classifiers (c4.5 decision tree, and Naive Bayes):




Can you suggest a reason for the observed difference in performance?

a. Decision trees are able to provide better coverage of high-dimensional spaces
b. The Naive Bayes classifier is susceptible to high dimensional inputs, caused by the additional input features.
c. C4.5 has built-in feature selection mechanisms
d. The Naive Bayes classifier does not provide any weighting of the input features
e. None of the above

19. You are trying to create a custom designed metric for cluster quality, where a higher value indicates better clustering. Given the following:

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

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

The following expressions would make reasonable measures of cluster "quality" **except** for?

a. $$\frac{\sum_{i \in c_1} \sum_{j \in c_2} d_{ij}}{\sum_{k,l \in c_1} d_{kl} + \sum_{m,n \in c_2} d_{mn}}$$

b. $$\frac{\sum_{i \in c_1} \sum_{j \in c_2} d_{ij}}{\max_{k,l \in c_1} d_{kl} \cdot \max_{m,n \in c_2} d_{mn}}$$

c. $$\frac{\min_{i \in c_1,\, j \in c_2} d_{ij}}{\max\left\{\max_{k,l \in c_1} d_{kl},\ \max_{m,n \in c_2} d_{mn}\right\}}$$

d. $$\frac{\sum_{i \in c_1} \sum_{j \in c_2} d_{ij}}{\min_{k \in c_1,\, l \in c_2} d_{kl}}$$ √

e. They all look alright

20. Which of the following statements about PCA is NOT correct:

a. It works by finding the directions in which most of the variance (power) of the data set is concentrated

b. It can be used to reduce the dimensionality of a data set by identifying the informative directions in the input space

c. The “principle components” are given by the eigenvectors of the covariance matrix of the data.

d. The corresponding eigenvalues can be used to determine the number of significant “dimensions”

e. It is plagued by local minima problems, and would benefit from a second-order learning algorithm

21. The following are all methods which can potentially be used for determining *n*, the number of principle components to retain, *except for*:

a. If used as a method of feature selection, perform classification using a range of different values for *n*, then choose the *n* which results in the highest accuracy.

b. Choose *n* such that 80% of the variance in the data set is retained

c. Create a plot of the variance captured by each of the principle components (sorted in descending order), and locate the “kink” in the graph

d. Choose *n* such that the eigenvectors are situated as closely to the points as possible.

e. They are all valid

22. In hierarchical clustering, what is the difference between agglomerative and divisive methods?

a. Divisive clustering algorithms start with the entire data set and try to determine the class and subclasses within.

b. Agglomerative clustering methods build an explicit statistical model of the clusters present

c. Agglomerative clustering algorithms tend to be iterative in nature

d. Divisive clustering algorithms tend to be iterative in nature, and hence are susceptible to local minima and slow convergence issues

e. Agglomerative clustering algorithms are faster

23. Here is a set of data points: [0,1,5,7,8,11].

    Using UPGMA, what are clusters detected?

    a. (0,1),(5,((7,8),11))

    b. (0,1),((5,(7,8)),11)

    c. ((0,1),5),((7,8),11)

    d. ((0,1),5),((7,8),11)

    e. (0,1),(((5,7),8),11)

24. For the data set above, and starting with the following two centroids: (0,1), and using the k-means algorithm, what are the two final clusters detected?

    a. (0),(1,5,7,8,11)

    b. (0,1),(5,7,8,11)

    c. (0,1,5),(7,8,11)

    d. (0,1,5,7),(8,11)

    e. (0,1,5,7,8),(11)

25. Which of the following is a valid reason for selecting the Self-Organizing-Map (SOM) over PCA:

    a. When the input data is extremely high dimensional, and as such a very large number of principle components are required

    b. PCA does not allow for different learning rates, making it inflexible

    c. When the input data has been “warped” and does not fall on a line or plane in the feature space

    d. SOM takes into account neighborhood relationships and thus preserves the topology of the input space.

    e. None of the above are correct

26. You are using a SOM to map a high-dimensional, highly complex data set. Which of the following statements is **NOT** true:

    a. In the initial phase of training, use a large neighborhood function, to allow the nodes to “unfold”

    b. In the latter phases of training, use a smaller neighborhood function, so that individual nodes can be fine-tuned

    c. The number of nodes should be just large enough to capture the shape of the data, but not so large as to result in overfitting

d. The learning rate parameter should be decreased throughout the learning process so that the final distribution of nodes matches the distribution of data

e. These are all valid comments