Data Mining Final Exam Spring 2016, May 11, 2016

Instructions: Please fill out your name and signature in the spaces provided below. The exam contains 24 questions. Each numbered question is weighted the same. Read the questions carefully. Some of the questions ask for multiple responses. After the exam has begun do not ask to leave for any reason except if you plan to submit your paper. Illegible answers will result in lost points. Good luck.

Name:

Signature:

1. Suppose we have an unlabeled dataset with 100 instances and 5 features. What is the dimensionality of the covariance matrix that we will calculate as the first step of performing Principal Components Analysis?

   The covariance matrix is 5×5.

2. PCA consists of performing and eigenvector decomposition of the covariance matrix $\mathbf{R}\mathbf{v}_i = \lambda_i \mathbf{v}_i$. What can you say about the value of $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$? In principle, what is the greatest lower bound for the λ_i?

   For eigenvectors corresponding to different eigenvalues $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ will be 0. Because of the symmetry of the covariance matrix the greatest lower bound of the eigenvalues is 0.

3. Briefly describe an algorithm that finds the best rank 2 approximation to a given dataset based on PCA?

   To find the best rank 2 approximation to our dataset, in terms of the Frobenius norm, we project all of the instances onto the first two principal components. Components corresponding to the two largest eigenvalues.

4. (Refer to Fig. 1). This figure shows a supervised learning problem (distinguishing plusses from squares). What is the number of features? What is the number of instances?

   2 features x_1 and x_2, and 22 instances

5. (Refer to Fig. 1). Recall that, in this case, a linear classifier takes the form $h_1 x_1 + h_2 x_2 \geq \theta$. Which of the following classifier/s is/are the most accurate? (Assume that the classifier predicts the 'plus' class when the inequality is true). Provide the *confusion matrix* of the most accurate classifier.

$$x_1 + x_2 \geq 0 \tag{1}$$

$$x_1 - x_2 \geq 0 \tag{2}$$

$$x_2 - x_1 \geq 0 \tag{3}$$

$$x_1 + x_2 \geq 7 \tag{4}$$

Accuracy of classifier is the number of instances that are predicted correctly. (1) is the most accurate. The accuracies are as follows: (1) 11/22, (2) 8/22, (3) 8/22, (4) 8/22

The confusion matrix for the first classifier is

$$\begin{bmatrix} 3 & 5 \\ 6 & 8 \end{bmatrix}$$

Because there are 3 correct squares, 6 incorrect plusses, 5 incorrect plusses, and 8 correct plusses

6. (Refer to Fig. 1). Which, if any, of the classifiers in the previous problem is incorrect for 100 percent of the given instances?

   None of the classifiers are incorrect 100 percent of the time.

7. (Refer to Fig. 1). In principle, what is the best accuracy that can be achieved using a linear model for these particular data instances?

   The data is not linearly separable. The best we can do is 15/22 with the classifier $x_2 - x_1 \geq 1$.

8. (Refer to Fig. 1). Of the following new features we could create from the given data $\{f_a, f_b, f_c, f_d\}$ which is most useful for our classification problem? What is the best accuracy we can achieve with a classifier of the form $h_1 x_1 + h_2 x_2 + h_a f_a + h_b f_b + h_c f_c + h_d f_d \geq \theta$?

$$f_a = x_1^2 \tag{5}$$

$$f_b = x_2^2 \tag{6}$$

$$f_c = |x_1 + x_2| \tag{7}$$

$$f_d = (x_1 - x_2)^2 \tag{8}$$

   f_d is the most useful feature because it makes the data linearly separable. Using this feature alone we can predict all of the instances correctly.

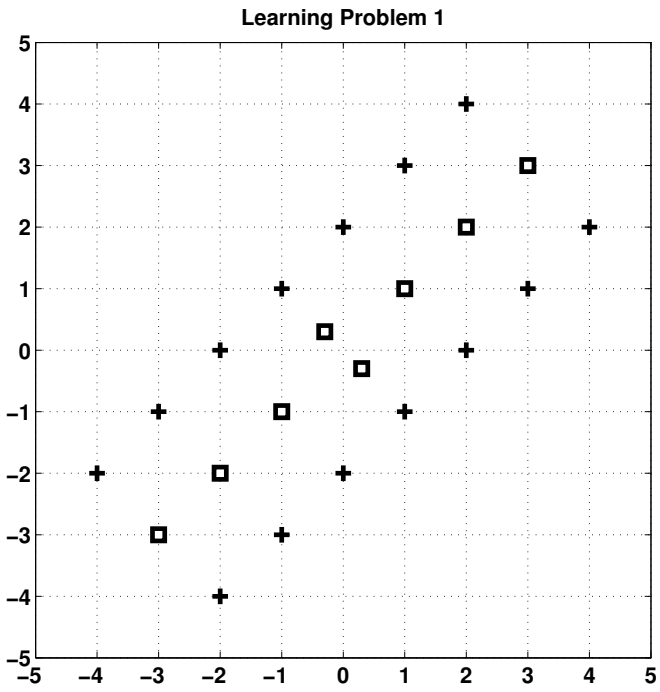

Figure 1: Let the horizontal axis correspond to x_1 and the vertical axis correspond to x_2.

9. In the context of support vector machines, what does *Mercer's condition* tell us?

   Mercer's Condition being true for the kernel function tells us that the kernel can be expressed as an inner product in an implicit space.

10. Provide a precise definition of the term *connected component* in the context of graph theory.

    A connected component of a graph is a subset of the nodes such that, for any two of the nodes in the subset there is a path, in the graph, between them.

11. Calculate the *degree distribution* of the graph shown in Fig. 2.

    The degree distribution counts the number of nodes for each degree. The degree distribution of the graph in Fig. 2 is $[0, 4, 3, 2, 1]$

12. How does a *scale-free* network differ from a network generated according to the Erdos-Renyi Model?

    The Erdos-Renyi Model is a generative model where each possible edge between any two nodes exists with some probability. The Erdos-Renyi model has a degree distribution that is binomial. A scale free network has a degree distribution that is power law.

13. What do we mean when we say that a node or edge has a high *centrality*? What is at least one way to quantify the centrality of a node or edge?

    Generally, high centrality means that the node is important to providing connectivity in the network.

14. What is meant by saying that a *norm*, denoted for instance by $||u||$, satisfies the triangle inequality?

    The triangle inequality says that the norm satisfies the familiar property of euclidean distances. That the sum of the lengths of two sides of a triangle is greater than or equal to the length of the third side. $||x|| + ||y|| \geq ||x + y||$

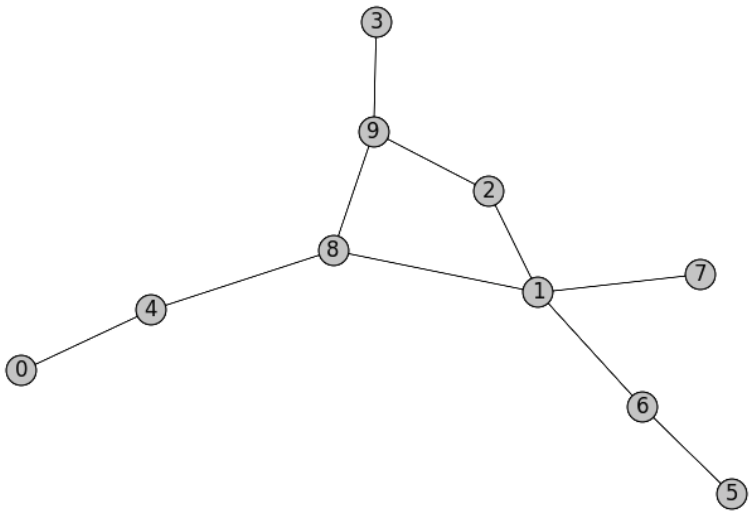

Figure 2: Example graph

15. Describe some of the information that is not considered when we analyze a textual corpus based upon it's *term-document matrix*.

    The term-document matrix only considers frequencies or relative frequencies of words. It does not consider, for instance, concepts that depend on the order of words such as grammar or context.

16. Explain how you would perform 3-fold cross validation on a dataset containing 100 instances.

    In 3-fold cross validation you will first partition the data into three groups. Say of sizes 33,33,34. Two of the partitions will be used as training for a particular type of model and the remaining can be used for the test data set. This process can be repeated three times and the results averaged together.

17. In the sparse reconstruction/compressive sampling problem we are attempting to solve an equation $\mathbf{y} = \mathbf{A}\mathbf{x}$, for $\mathbf{A}$ an $m \times n$ matrix with $m << n$ and $\mathbf{x}$ and $\mathbf{y}$ appropriately sized vectors. Under what conditions does the equation not have any solutions?

    In general, the problem will not have a solution if the matrix A has rank less than m.

18. Continuing the previous problem and suppose $\mathbf{y}$ and $\mathbf{A}$ are given, what would be the objective function if we wanted to find $\mathbf{x}$ to minimize the squared error? How would we regularize the objective to improve our chances of finding a *sparse* solution?

    $J(x) = \sum(\mathbf{y} - \mathbf{A}\mathbf{x})^2$, for regularization we can use the sparsity or l_1 penalty. $J(x) = \sum(\mathbf{y} - \mathbf{A}\mathbf{x})^2 + \alpha \sum |\mathbf{x}|$

19. Describe three different techniques for *visualizing* a dataset and also the types of data for which the techniques are appropriate.

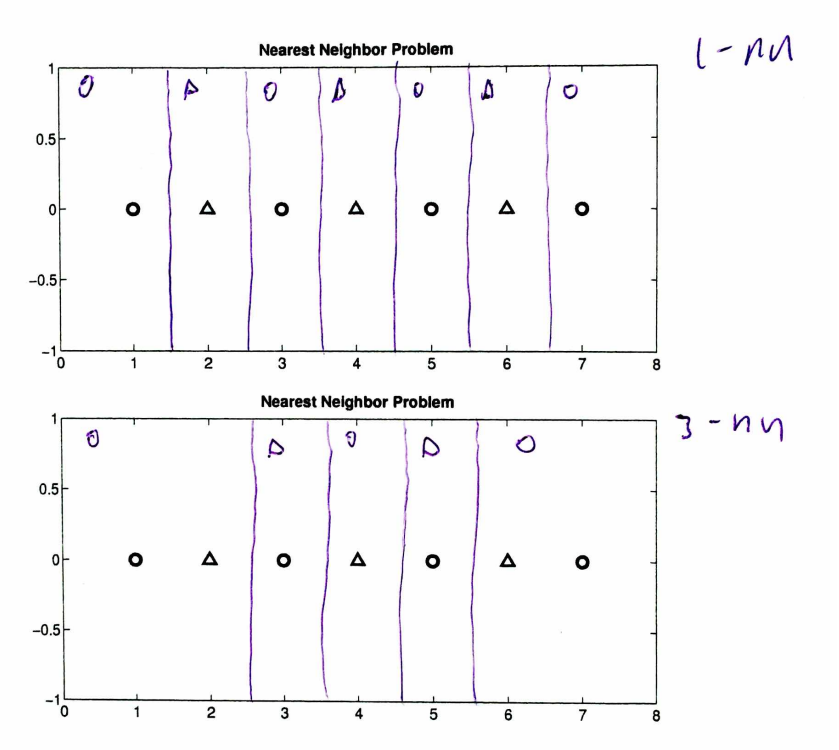

Figure 3: Sample data for nearest neighbor problem

...

20. What does the k-means algorithm do? Explain some of its weaknesses or limitations.

    k-means is an algorithm for clustering data. Appropriate for unsupervised learning. Some of its limitations are: dependence on initial conditions, have to select model order in advance.

21. When is a matrix *diagonalizable*?

    A matrix A is diagonalizable if we can find an invertible transformation P such that PAP^{-1} is a diagonal matrix.

22. What does a non-trivial linear subspace of $\mathbb{R}^2$ look like?

    A non-trivial subspace of the plane is a line that passes through the origin. Trivial subspaces would be the entire space or the empty space.

23. Contrast Principal Components Analysis with Locally Linear Embedding. How are they similar and how are they different?

    PCA and LLE are both unsupervised learning methods that look for structure in an unlabeled dataset. PCA characterizes the data in terms of linear subspaces while LLE is able to find non-linear manifold structure.

24. Consider the labeled (circles and triangles) dataset shown in Fig. 3. Sketch the decision boundaries produced by the nearest neighbor algorithm for this dataset. In particular, show the cases of 1-nearest neighbor and 3-nearest neighbor. Sketch the boundaries directly on the figures. An extra copy of the graphic is provided for your convenience.

    See figure