In The Name of God, The Compassionate, The Merciful

Name: .....

Student ID#: .....

Statistical Pattern Recognition (CE-725)
Department of Computer Engineering
Sharif University of Technology

Final Exam - Spring 2010
(120 minutes – 150+10 points)

1. SVM and kernel methods (50 points)

(a) True/False - In this section you must explain your answers.

a1. ..... (5 points) In a two-class classification problem, any decision boundary that we get from a generative model with class-conditional Gaussian distributions could in principle be reproduced with an SVM and a polynomial kernel of degree less than or equal to three.

a2. ..... (5 points) The values of the margins obtained by two different kernels $K_1(x,x_0)$ and $k_2(x,x_0)$ on the same training set do not tell us which classifier will perform better on the test set.

a3. ..... (5 points) After mapped into higher dimensional feature space, through a radial basis kernel function, 1-NN using unweighted Euclidean distance may be able to achieve better classification performance than in original space.

(b) (10 points) Assume we use radial basis kernel function $K(x_i, x_j) = exp(-1/2\ |x_i-x_j|^2)$. Show that for any two input instances x_i and x_j the squared Euclidean distance of their corresponding points in the higher dimensional feature space is less than 2.

(c) (10 points) Suppose that the embedding of a kernel k is Φ. Find the embedding of the kernel k^2.

(d) (15 points) Prove that if $k_1(x, y)$ is a valid kernel, then the following one is valid, too:

$$k(x,y) = \frac{k_1(x,y)}{\sqrt{k_1(x,x)}\sqrt{k_1(y,y)}}$$

2. Clustering (50 points)

(a) (10 points) You are given a Gaussian mixture model, and all its class probabilities and Gaussian mean locations are learned using EM, but the covariance matrices are forced to be the identity matrices for each class. Rather than using a mixture of Gaussians, you evaluate the probability of each of the K classes given the datapoint and take the the cluster with the highest probability to be the cluster that produced the point. Is this equivalent to doing using a K-means model?

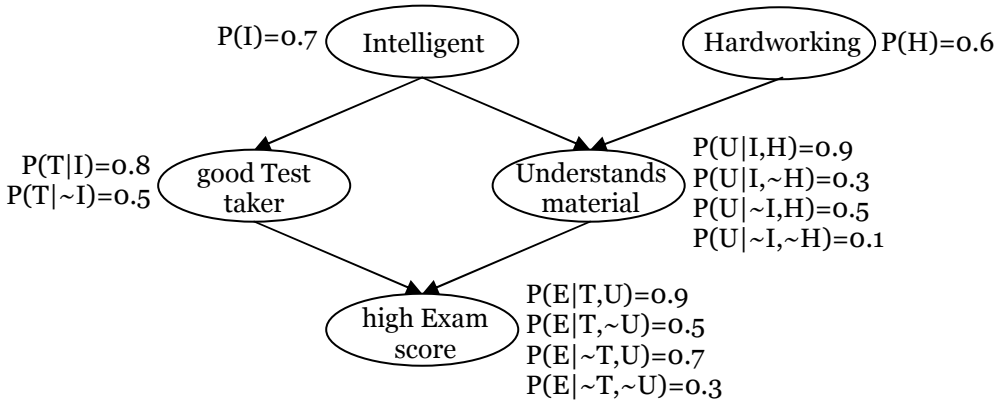
(b) (10 points) Suppose you've done K-means and your K is equal to your number of data points with each cluster defined by a single datapoint. Say that you classify test data points as part of the cluster that they would belong to according to your distance metric. Is this equivalent to 1-NN?

(c) (10 points) Using hierarchical clustering (with the minimum distance criteria), what is the maximum height of the hierarchy tree required to cluster a set of N points? What is the minimum height?

(d) (20 points) Show that the clustering algorithm based on minimizing the determinant of the within class scatter is invariant against the linear non-singular transformation of the data.

3. Graphical Methods (30 points)

We are going to take the perspective of an instructor who wants to determine whether a student has understood the material, based on the exam score. The following figure gives a Bayes net for this. As you can see, whether the student scores high on the exam is influenced both by whether she is a good test taker, and whether she understood the material. Both of those, in turn, are influenced by whether she is intelligent; whether she understood the material is also influenced by whether she is a hard worker:



Compute the probability that a student who did well on the test, actually understood the material.

4. Expectation Maximization (30 points)

Assume that a set of points (X, Y, Z) is generated according to the generative model (X, Y, and Z are Boolean variables):

- The variable X is set to 1 with probability α.
- The variable Y is set to 1 with probability β.
- If (X, Y) = (1, 1) then Z = 1 with probability λ_{11} .
- If (X, Y) = (0, 1) then Z = 1 with probability λ_{01} .
- If (X, Y) = (1, 0) then Z = 1 with probability λ_{10} .
- If (X, Y) = (0, 0) then Z = 1 with probability λ_{00} .

We need to estimate the parameters of this model. However, one of the variables, Y, is not observed. We are given a set of m data points $D = \{(x_i, z_i) ; 1 \le i \le m\}$.

Derive the E-step and M-step, and give explicit expressions for the parameter updates in the EM process.

Good Luck!