Name and Student ID: Signature:

Machine Learning BLG527E, Jan 18, 2012, Final Exam.

1	2	3	4	Total

Duration: 120 minutes.

Write your answers neatly in the space provided for them. Write your name on each sheet. Books and notes are closed. Good Luck!

QUESTIONS

QUESTION1) [40 points, 10 points each] (use at most 10 sentences per question, use of formulas, drawings etc. to better express yourself is encouraged.)

a) How do you control an SVM's model complexity?

b) How do you train (i.e. decide on the best weights) a multilayer perceptron? How can you decide on the best complexity? How can you speed up the learning process?

c) How do you use Parzen windows for density estimation, classification and regression?

d) What are the differences and similarities between logistic regression and multilayer perceptron classifier?

QUESTION2) [20points] Consider the 2-means algorithm on a set S consisting of the following 4 points in the plane: b=(1,3), c=(2,3), e=(3,2), g=(4,2). The algorithm uses the Euclidean distance metric to assign each point to its nearest centroid; ties are broken in favor of the centroid to the left/down. A starting configuration is a subset of 2 starting points from S that form the initial centroids. A 2-partition is a partition of S into 2 subsets; thus {b,c}, {e,g} is a 2-partition; clearly any 2-partition induces a set of two centroids in the natural manner. A 2-partition is **stable** if repetition of the 2-means iteration with the induced centroids leaves it unchanged.

a. How many starting configurations are there?

b. What are the stable 2-partitions?

c. What is the number of starting configurations leading to each of the stable 2-partitions in (b) above?

QUESTION3) [20points] You need to write a machine learning algorithm to help people who are learning to write Turkish.

i) The algorithm needs to identify whether words in a sentence are in the order they are supposed to be in Turkish. For example, saying "Yaz Ali." is less common than "Ali yaz." in written language.

ii)If they are not, you need to suggest a better ordering.

What would you ask your customer to provide you with as data?
What machine learning method(s) would you use to solve tasks i) and ii).
How would you measure the success of your methods?

QUESTION4) [20points] You need to come up with a decision tree that separates four data points labeled as follows U = {(x,r(x))} = {((1,2),+), ((2,3),+), ((2,1),-), ((3,2),-)}.

Use entropy impurity:

$$I_m = -\sum_{i=1}^{K} p_m^i \log_2 p_m^i$$

where p_m^1 is the ratio of the instances in node m which are in class i.