Recognition January 16, 2017

Name:

1. (5 points) Decision tree classifiers.
   (a) (2 points) Describe the simplest top-down tree-building algorithm that selects the decision rule at a node by maximisation of information gain.
   (b) (3 points) Discuss advantages and disadvantages of decision trees.

2. (5 points) Describe the SVM learning algorithm. Discuss the time complexity of the algorithm. Compare its properties with AdaBoost, Logistic regression, the nearest neighbour and neural network learning and classification methods.

3. (5 points) The K-means algorithm.
   (a) (1 point) Describe the basic algorithm.
   (b) (1 point) Describe the generalisation to L_1 minimisation, i.e. to the minimisation of the sum of absolute differences to the cluster "centroid".
   (c) (1 point) Describe the generalisation to the minimisation of the sum of the Euclidean distances to the cluster "centroid" (the Weiszfeld's algorithm).
   (d) (1 point) Discuss the time complexity of the algorithm and its generalisations.
   (e) (1 point) Describe the initialisation scheme used in K-means++.

4. (5 points) Maximum likelihood estimation.
   (a) (4 points) The reliability of hard drives is defined by the probability density function $p(t) = \theta e^{-\theta t}, t \in (0, \infty)$, where $p(t)\delta$ is the probability that a hard drive fails in a short interval δ centred at time t. What is the maximum likelihood estimate of the reliability parameter θ if in an experiment with three hard drives the following lifetimes have been observed: $t_1 = 56, t_2 = 120$ and $t_3 = 424$?
   (b) (1 point + 1 bonus point) If the test described above finished at time $t = 300$ one hard drive would still be running, i.e. its time of failure would be unknown. What would an ML estimate for θ be in this case?