T-61.3050 PROBLEMS 9/2007

In T1 on 16 November 2007 at 10 o'clock.

You should solve the problems before the problem session and give the solved problems to the assistant. Please write clearly and leave a wide (left or right) margin. The solutions should be stapled together **with a cover sheet** containing your name, student number and the numbers of problems you have solved.

For the problems where a "correct" solution exists (math and algorithm questions) the assistant will present one possible solution during the session. In some cases the questions do not have a single correct answer, but the idea is that you think about the problem and are prepared to discuss it with the assistant and other students during the session.

See `http://www.cis.hut.fi/Opinnot/T-61.3050/2007/problems` for up-to-date information of the problem sessions.

This problem sheet has two pages.

1. You want to travel on foot from Town A to Town B, but there is a jungle between the towns. The distance is too long to be covered in one day. Also, due to potential life-threatening hazards you might encounter in the jungle during the night, you are only prepared to proceed during daytime. You have a map that shows a path through the jungle, and you have marked a number of safe locations on the path to set up a camp for the night. As you are in a hurry (or maybe just afraid of sleeping in the jungle, who knows), you want to minimize the number of times you actually have to set up the camp.

   Consider the following greedy strategy: You start from Town A and walk on the path always stopping at the safe locations. Whenever you arrive at a safe location, you check your map and estimate whether or not you can reach the next safe location before nightfall. If yes, you proceed, otherwise you set up your camp at the current location and continue on the next day until you reach Town B.

   Does this strategy minimize the number of nights you have to spend alone in the jungle? If yes, give a proof, if not, give a counterexample. (You might want to check out the `http://www.cs.cornell.edu/Courses/cs482/2007su/ahead.pdf`)

Algorithm 1 Algorithm for Problem 2.

```
1: CLUSTER($X$,$d$,$k$) {Input: $X$, set of points; $d$, the Euclidean distances
   between the points in $X$; $k$, number of clusters. Output: Clustering of
   points in $X$ to $k$ clusters.}
2: $Q \leftarrow \{\{u, v\} : u, v \in X\}$
3: $T \leftarrow \emptyset$
4: while $Q \neq \emptyset$ do
5:     $\{u^*, v^*\} \leftarrow \arg\min_{\{u,v\} \in Q} d(u, v)$ {$d(u, v)$ is the Euclidean distance be-
       tween $u$ and $v$}
6:     Remove $\{u^*, v^*\}$ from $Q$
7:     if pairs in $T$ do not form an undirected path between $u^*$ and $v^*$ then
8:         Add $\{u^*, v^*\}$ to $T$
9:     end if
10: end while
11: Remove those $k - 1$ pairs from $T$ for which $d(u, v)$ is highest
12: return clustering where each cluster is formed by a connected component
    in $T$
```

2. Consider the algorithm 1 for computing a clustering of the set X of points to k disjoint clusters.

   (a) Let $G = (X, E)$ be an undirected graph, where X is the set of vertices and E is equal to Q after it has been initialized in line 2 of the algorithm. What property of G does the algorithm compute?

   (b) The *spacing* of a clustering is defined as the shortest distance between any two points that belong to different clusters. Show that the above algorithm constructs a clustering with maximum spacing.

3. Derive the EM update formulae for Gaussian mixture model when the data is a set of real numbers (that is, forget multivariate distributions, $d = 1$), variance and the distribution of classes is fixed to σ^2 and $P(G_i) = \pi_i = 1/k$, respectively.

4. Take the example data of Mitchell's (see the lecture slides) and with pen and paper, simulate the running of the ID3 algorithm. What is the resulting decision tree? Would the use of various impurity measures make a difference?

Problems 1 and 2 are taken from the book *Algorithm Design* by Jon Kleinberg and Eva Tardos.