CS176–Fall 2015 — Solutions to Homework 3

Manohar Jois

November 5, 2015

Problem 1

An idea here is to notice that if the UPGMA-built tree corresponding to δ isn't consistent with δ, then δ is not an ultrametric. Let N_v be the set of leaves that are descendants of node v in the tree. Let L_v and R_v be the same but separated into the left and right children of v, such that $N_v = L_v + R_v$. For leaf nodes x, N_x is just the set containing x itself. Also let h_v be the height of a node, where leaves have zero height.

We first build the tree, then we process the internal nodes from the bottom up. At each node, we set $N_v = L_v + R_v$. If we take the cross-product of L_v and R_v, we get the set of leaf pairs whose lowest common ancestor is node v. For each pair i, j, simply check if $\delta(i, j) = 2h_v$. If so, continue, but if not, then δ is not an ultrametric. If after the bottom-up pass all pairs are verified, then δ is an ultrametric.

This works because if δ is an ultrametric, then the tree distances will be consistent with delta. So the contrapositive is true as well. The distance between two leaves must simply be twice the height of their lowest common ancestor, since UPGMA trees obey the molecular clock property. Checking this against their dissimilarity is enough to check consistency.

There is a way to construct the UPGMA tree in quadratic time, and during the verification process each leaf pair gets constructed and verified exactly once (by the lowest common ancestor node of the two leaves). This makes the algorithm $O(n^2)$.

Problem 2

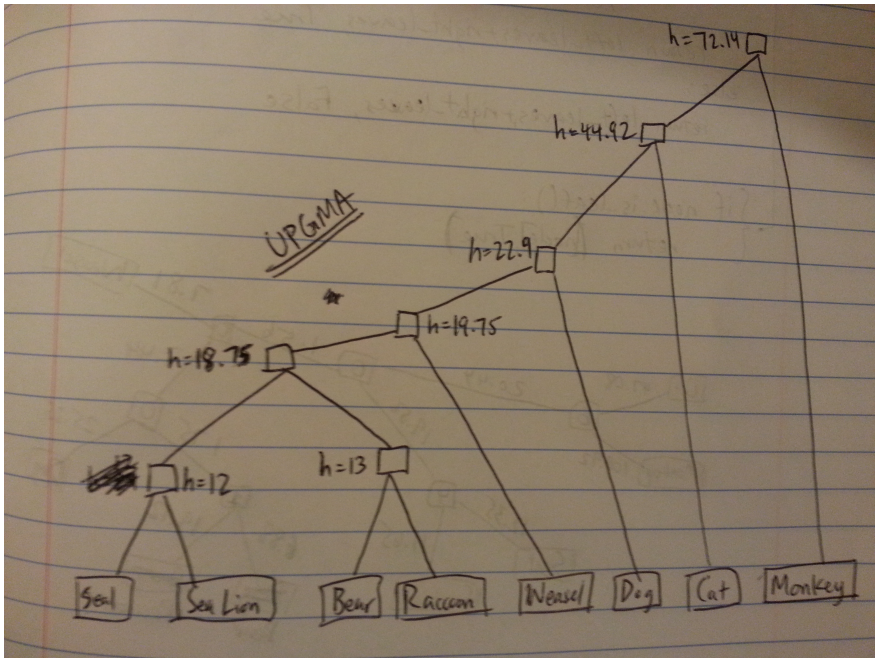
Let *internal edges* be those connecting internal nodes in the tree and *external edges* be those connecting leaves to internal nodes. Let $r(i) = \sum_{j \in L} d(i,j)$. For each external edge, we can safely set its weight w to zero without affecting the relative ordering of the Q-criterion because each every r-value will decrease by w, which decreases all Q-values by $\frac{2w}{|L|-2}$.

Consider an internal edge connecting two internal nodes u and v. Let N_u be the number of leaves closer to u and N_v be the same for v. From this it is easy to see that $r(u) = r(v) + N_v w_{uv} - N_u w_{uv}$. This implies that the node with greater r-value is the one closer to fewer leaves, which further implies that the internal node with greatest r-value is adjacent to only one internal edge. Call this node z. Also note that $r(z) = r(x)$ for all leaves x adjacent to z since external edge weights are zero.

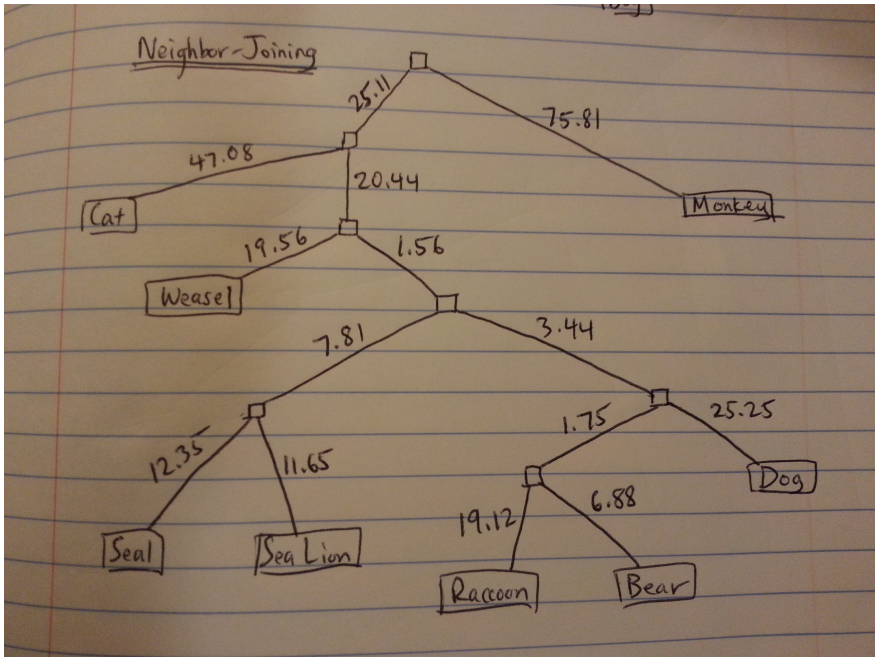
Now take arbitrary leaves a and b which are not neighbors, and consider arbitrary leaves x and y which are neighbors connected by z. Clearly $d(a,b) \geq 0 = d(x,y)$ because there is at least one internal edge between the former. Also since z has the maximal r-value, $r(a), r(b) \leq r(z) = r(x) = r(y)$. Using these two inequalities, it is clear that $Q(x,y) \leq Q(a,b)$, proving that cherry leaves minimize the Q-criterion.

Problem 3

(a)


h=72.14
h=44.92
UPGMA
h=22.9
h=19.75
h=18.75
h=12
h=13
Seal
Sea Lion
Bear
Raccoon
Weasel
Dog
Cat
Monkey

(b)


Neighbor-Joining
25.11
75.81
47.08
20.44
Cat
Monkey
19.56
1.56
Weasel
7.81
3.44
1.75
25.25
12.35
11.65
Dog
6.88
19.12
Seal
Sea Lion
Raccoon
Bear

(c) The main difference in the two rooted trees is the depth of the leaves. The UPGMA tree is constructed to show depth as analogous to time, so each leaf (species) is at the same depth. The neighbor-join tree edges are more analogous to the actual dissimilarities in the matrix δ. Also, interestingly, the leaf for Dog is much closer to Seal, Sea Lion, Bear and Raccoon in the NJ tree than in the UPGMA tree.

Problem 4

Problem 5