Exam Data Mining
Date: 5-11-2013
Time: 17.00-20.00

Question 1 Short Questions

(a) We speak of overfitting when a model is adjusted too much to the peculiarities of the training sample, causing bad generalization to unseen data. To prevent overfitting in classification trees, one can use cost-complexity pruning. We compute a nested sequence of trees where each tree in the sequence is the smallest minimizing subtree for an interval of alpha (the complexity penalty) values. We select the tree from the sequence that has the smallest error on a test set.

(b) Two disadvantages:

1. We may have to throw away a large part of the records just because one attribute value in the record is missing. The remaining data set may be quite small, which tends to have a negative effect on the quality of the model.
2. There may be a systematic difference between records with missing values and complete records, so if we only consider the complete records, we may have a biased sample.

(c)
1. Compute the mutual information between each pair of variables.
2. Make a complete (undirected) graph where each node corresponds to a variable.
3. Put the mutual information between a pair of variables as a weight on the edge between that pair.
4. Compute a maximum weight spanning tree.
5. Pick a root node, and let all the edges point away from the root.

(d) A-close is more efficient then Apriori on dense, highly correlated data sets.

(e) To predict the class label of one object, we need the class labels of the linked objects, because they determine the value of the link-attributes. But the class labels of these linked objects are also unknown. To bootstrap the whole process, we can start with

an initial classification that only uses the node-attributes and disregards the link-attributes. Starting from this initial classification we can iteratively improve the solution.

Question 2: Classification Trees

(a) The split between 12 and 14, and the split between 15 and 17.

(b) Split between 12 and 14:

$$\pi(\ell)i(\ell) + \pi(r)i(r) = \frac{1}{3} \times \frac{3}{3} \times \frac{0}{3} + \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} = \frac{4}{27} \approx 0.148$$

Split between 15 and 17:

$$\pi(\ell)i(\ell) + \pi(r)i(r) = \frac{7}{9} \times \frac{5}{7} \times \frac{2}{7} + \frac{2}{9} \times \frac{0}{2} \times \frac{2}{2} = \frac{10}{63} \approx 0.159$$

The split between 12 and 14 wins. It has an impurity reduction of

$$i(t) - (\pi(\ell)i(\ell) + \pi(r)i(r)) = \frac{5}{9} \times \frac{4}{9} - \frac{4}{27} = \frac{8}{81}$$

(c) No. Suppose minleaf=4: then neither of the splits listed under (a) would be acceptable. But we can make a split between 14 and 15 that satisfies the minleaf constraint. Hence, it is not correct to only consider splits on the border between segments in the presence of a minleaf constraint.

Question 3: Frequent Pattern Mining

(a) Level 1:

candidate	support	frequent?
A	5	✓
B	3	✓
C	3	✓
D	4	✓
E	1	✗
F	1	✗

Level 2:

candidate	support	frequent?
AB	2	✓
AC	2	✓
AD	3	✓
BC	1	✗
BD	1	✗
CD	3	✓

Level 3:

candidate	support	frequent?
ACD	2	✓

The "pre-candidate" ABC is generated from level 2 frequent item sets AB and AC, but it is pruned because level 2 subset BC is not frequent. The same reasoning applies to ABD. There are no level 4 candidates.

(b) Two times. The RMO-list is: (v_6, v_8).

(c) Six times.

	w_1	w_2	w_3
ϕ_1	v_1	v_3	v_6
ϕ_2	v_1	v_3	v_8
ϕ_3	v_1	v_5	v_6
ϕ_4	v_1	v_5	v_8
ϕ_5	v_1	v_7	v_8
ϕ_6	v_6	v_7	v_8

(d) Nine candidates. There are 3 nodes on the rightmost path of the tree, and to each we can add a node labeled a,b, or c as the rightmost child.

Question 4: Iterative Proportional Fitting

(a) Which margin constraints have to be satisfied by the fitted counts?

$$\hat{n}_1(x_1) = n_1(x_1)$$
$$\hat{n}_2(x_2) = n_2(x_2)$$

(b)

$$\hat{n}^{(1)} = \begin{array}{c|cc|c} & 0 & 1 & \\ \hline 0 & 40 & 40 & 80 \\ 1 & 10 & 10 & 20 \\ \hline & 50 & 50 & \end{array}$$

$$\hat{n}^{(2)} = \begin{array}{c|cc|c} & 0 & 1 & \\ \hline 0 & 72 & 8 & 80 \\ 1 & 18 & 2 & 20 \\ \hline & 90 & 10 & \end{array}$$

The algorithm has converged since all margin constraints have been satisfied simultaneously.

(c)

$$\hat{n}^{(1)} = \begin{array}{c|cc|c} & 0 & 1 & \\ \hline 0 & 75 & 5 & 80 \\ 1 & 15 & 5 & 20 \\ \hline & 90 & 10 & \end{array}$$

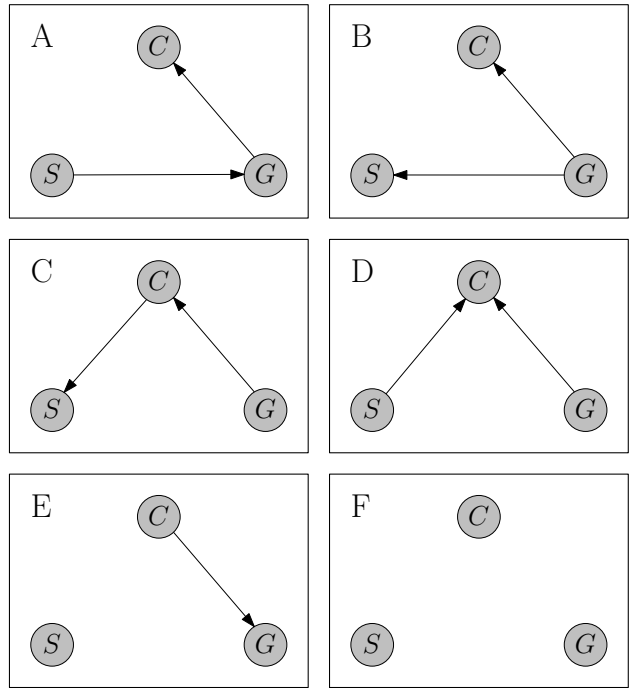
The algorithm has converged since all margin constraints have been satisfied simultaneously.

(d) Solution (b) is correct. Solution (c) doesn't satisfy the independence constraint. The initial solution of (c) didn't satisfy the independence constraint, in which case the algorithm may fail to converge to the correct solution. Of course, you can also check which solution corresponds to the one obtained by using the formula for the maximum likelihood fitted counts:

$$\hat{n}_{12}(x_1, x_2) = \frac{n_1(x_1)n_2(x_2)}{N}$$

Question 5: Bayesian Networks

(a) A and B are equivalent. E is equivalent to the current model.



(b) Note that we subtract 1 from the log-likelihood score, because the node G costs 1 parameter.

$$\mathrm{AIC}(G) = 213 \ln \frac{213}{474} + 261 \ln \frac{261}{474} - 1 = -327.12$$

(c) Recompute the contribution of node G to the score:

$$\mathrm{AIC}(G) = 77 \ln \frac{77}{150} + 73 \ln \frac{73}{150} + 136 \ln \frac{136}{324} + 188 \ln \frac{188}{324} - 2 = -326.31$$

The score of node G has increased. Since this is the only node whose score (parent set) has changed, the overall score of the model increases by this change.

(d) Survival and Center are independent given Grade.