ASSIGNMENT 12: (Spr 2020)

1. For the posted FURNITURE data, run Extree to obtain both an additive tree and an extended tree solution.
[NOTE: if you can't get EXTREE to run on your machine, I have posted the EXTREE output for the Sports data, you can just use that output to do the rest of the assignment.]
<EXTREE output is appended below>

3. Discuss the interpretation of the tree and marked features in the EXTREE solution.
<See annotated EXTREE output below>

2. Using an F test, test if the extended tree solution offers improvement in fit (adjusted for # parameters) over the additive tree (see test below). Report your conclusion.

Alternative formula for F_{obs} to compare the fit of two nested regression models, where Model A has one or more predictors in addition to the predictors used in Model C:

$$F_{obs} = F^* = \frac{PRE/(PA-PC)}{(1-PRE)/(N-PA)} = \frac{\Delta R^2/(PA-PC)}{(1-R_A^2)/(N-PA)}$$

where $\Delta R^2 = R_A^2 - R_C^2$ = difference in fit (RSQ) between the two regression models

PA is the number of parameters (#predictors + 1) for Model A (the more complex or "augmented" model)
PC is the number of parameters (#predictors +1) for Model C (the "compact" model)
F_{obs} is tested against F_{crit} with (PA-PC) numerator d.f. and (N-PA) denominator d.f.
N is the number of data points (proximities).

For the Furniture data, EXTREE gives an RSQ of .8345 for the additive tree ("Model C"), and RSQ = .9058 for the extended tree ("Model A") due to the addition of 10 marked features. There are n(n-1)/2 proximities, which equals (20)(19)/2 = 190. The number of parameters for the additive tree = 2n-2 = 38, and ten more for the EXTREE. So,

$$F_{obs} = \frac{\Delta R^2/(PA-PC)}{(1-R_A^2)/(N-PA)} = \frac{(.9058-.8345/(48-38)}{(1-.9058)/(190-48)} = \frac{.0713/(10)}{(.0942)/(142)} = 10.748$$

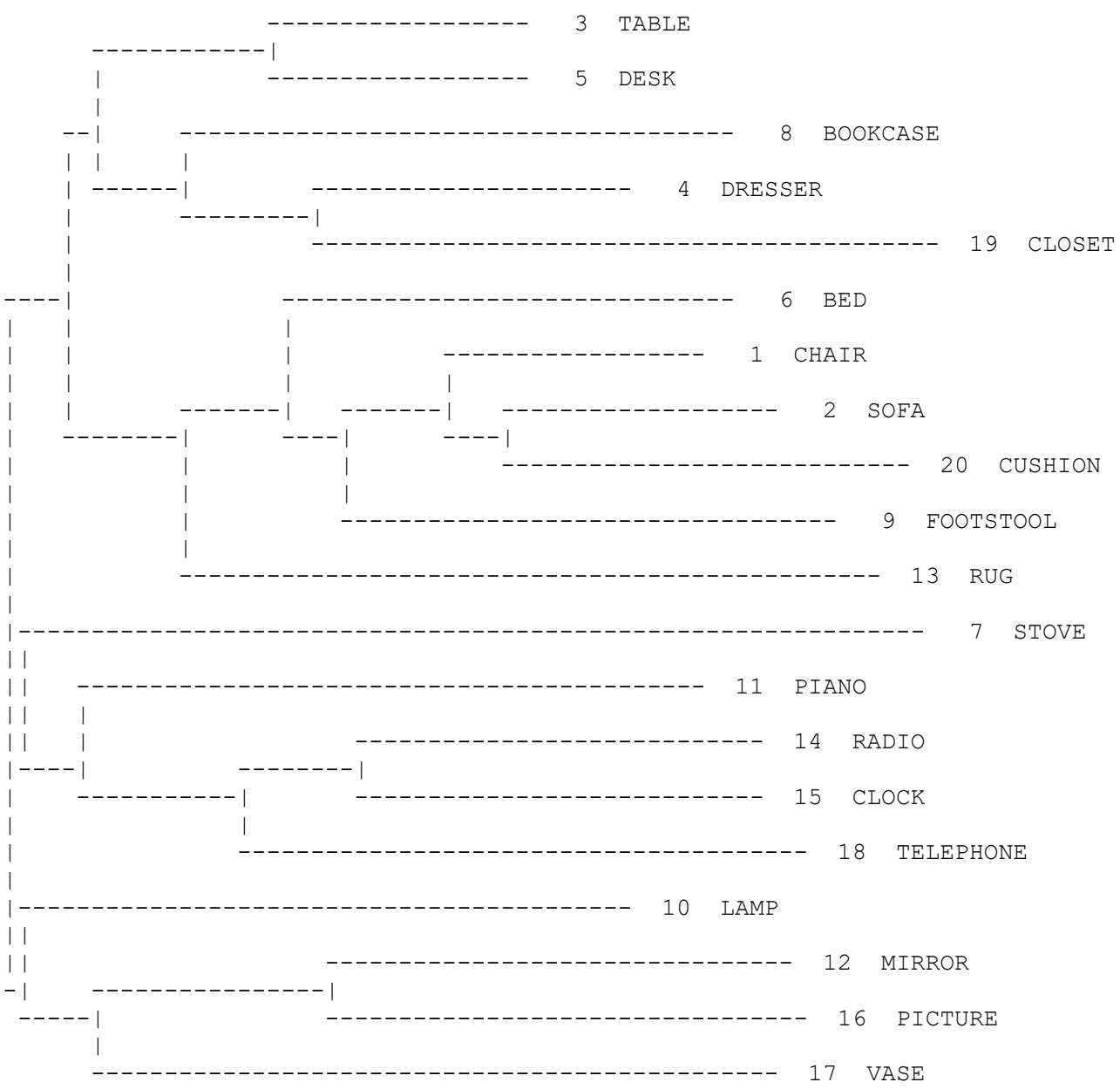
d.f. = (10,142), thus F_{crit} = F(.05; 10, 142) = 1.91, so the test is significant, reject H0 that the two models fit equally well. In other words, the extended tree significantly improves the fit.

Note that the individual proximity values usually cannot be assumed to be independent observations, so the validity of the F test is justified as a permutation test (Freedman & Lane, 1983).

EXTREE output for Furniture (edited):

```
extree analysis (EXTREE version 1.5):
furniture similarity ratings, converted to dissims

(   0.0  needed for positivity of distances )
  -90.0  added to exactly satisfy triangle inequality

transformed data distances:
 118.0
 174.0 234.0
 249.0 290.0 177.0
etc.

warning: threshold must be >= 0.0 (resetting it)

 node     length    children    label

   1      54.7                  CHAIR
   2      56.7                  SOFA
   3      56.4                  TABLE
   4      67.6                  DRESSER
   5      55.6                  DESK
   6      95.3                  BED
   7     193.8                  STOVE
   8     116.0                  BOOKCASE
   9     104.2                  FOOTSTOOL
  10     130.5                  LAMP
  11     135.8                  PIANO
  12      97.3                  MIRROR
  13     150.8                  RUG
  14      85.3                  RADIO
  15      84.8                  CLOCK
  16      99.7                  PICTURE
  17     146.9                  VASE
  18     120.9                  TELEPHONE
  19     136.4                  CLOSET
  20      84.3                  CUSHION
  21      12.8      2  20
  22      36.3      3   5
  23      26.5      4  19
  24      53.6     12  16
  25      27.6     14  15
  26      21.4      1  21
  27      19.5      8  23
  28      16.1     24  17
  29      35.9     25  18
  30      14.3     26   9
  31      22.0      6  30
  32      10.1     11  29
  33      25.4     31  13
  34       6.7     22  27
  35       5.9      7  32
  36       3.4     10  28
```

```
  37      14.8     34  33
  38       0.0     35  36
  39       0.0     37  38
```




```
stress formula 1 = 0.0609
stress formula 2 = 0.3540
r(monotonic) squared=0.8747
r-squared (p.v.a.f.)=0.8345
```

extree analysis:

  10 marked features will be tried
features smaller than   0.0 will be eliminated

i	j	estimate	[set(i)]	[set(j)]
20	13	51.8	[20]	[13]
14	11	51.3	[14]	[11]
1	22	42.6	[1]	[3 5]
12	23	42.2	[12]	[4 19]
4	12	39.1	[4]	[12]
5	1	41.9	[5]	[1]
21	13	37.8	[2 20]	[13]
14	18	36.8	[14]	[18]
21	6	36.7	[2 20]	[6]
19	12	34.0	[19]	[12]
26	6	30.7	[1 2 20]	[6]
12	27	30.6	[12]	[4 8 19]
3	1	30.5	[3]	[1]
16	17	28.8	[16]	[17]
1	11	28.4	[1]	[11]
30	11	26.9	[1 2 9 20]	[11]
15	10	26.6	[15]	[10]
2	6	25.1	[2]	[6]
4	22	24.8	[4]	[3 5]
26	13	24.1	[1 2 20]	[13]
15	7	24.1	[15]	[7]
4	6	23.6	[4]	[6]
25	10	23.2	[14 15]	[10]
23	24	23.0	[4 19]	[12 16]
20	6	22.4	[20]	[6]
5	8	22.7	[5]	[8]
9	11	22.3	[9]	[11]
29	7	22.3	[14 15 18]	[7]
29	10	21.6	[14 15 18]	[10]
24	27	22.3	[12 16]	[4 8 19]
3	4	21.5	[3]	[4]
26	11	20.9	[1 2 20]	[11]
15	6	20.5	[15]	[6]
16	8	20.2	[16]	[8]
11	33	19.9	[11]	[1 2 6 9 13 20]
4	24	19.7	[4]	[12 16]
19	24	19.7	[19]	[12 16]
30	13	19.4	[1 2 9 20]	[13]
5	10	19.2	[5]	[10]
27	28	19.2	[4 8 19]	[12 16 17]
19	7	19.1	[19]	[7]
25	7	19.0	[14 15]	[7]
23	6	18.0	[4 19]	[6]
8	22	17.7	[8]	[3 5]
1	34	17.5	[1]	[3 4 5 8 19]
25	6	17.3	[14 15]	[6]

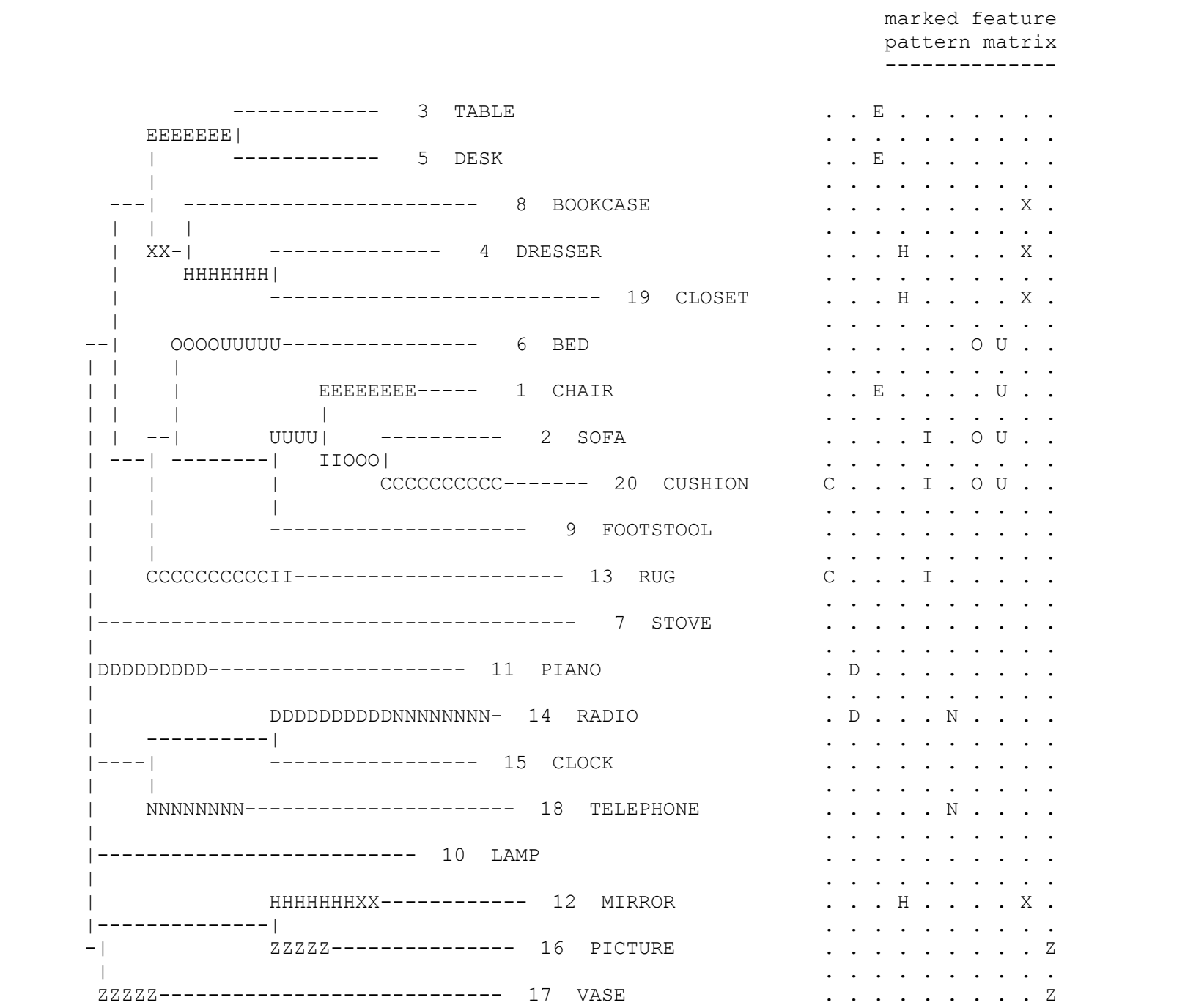
```
  3 30       17.0       [  3 ]       [  1  2  9 20 ]
 23 28       17.1       [  4 19 ]       [ 12 16 17 ]
  6 29       16.3       [  6 ]       [ 14 15 18 ]
  3  9       16.2       [  3 ]       [  9 ]
  3 31       16.1       [  3 ]       [  1  2  6  9 20 ]
  4 28       15.7       [  4 ]       [ 12 16 17 ]
 20 11       15.5       [ 20 ]       [ 11 ]
 17 11       15.1       [ 17 ]       [ 11 ]
 30 22       15.0       [  1  2  9 20 ]       [  3  5 ]
  2  8       14.8       [  2 ]       [  8 ]
 31 22       14.8       [  1  2  6  9 20 ]       [  3  5 ]
 11 31       14.7       [ 11 ]       [  1  2  6  9 20 ]
 22 10       14.1       [  3  5 ]       [ 10 ]
 11 13       14.2       [ 11 ]       [ 13 ]

checking for cliques & redundant patterns of marked features

feature C
  20   ( 20 )
  13   ( 13 )
feature D
  14   ( 14 )
  11   ( 11 )
feature E
   1   (  1 )
  22   (  3  5 )
feature H
  12   ( 12 )
  23   (  4 19 )
feature I
  21   (  2 20 )
  13   ( 13 )
feature N
  14   ( 14 )
  18   ( 18 )
feature O
  21   (  2 20 )
   6   (  6 )
feature U
  26   (  1  2 20 )
   6   (  6 )
feature X
  12   ( 12 )
  27   (  4  8 19 )
feature Z
  16   ( 16 )
  17   ( 17 )


iteration:   1

spillover of marked features on arcs:   21  22  23  32
features smaller than threshold:
maximum leaf spillover=    1.7
```

iteration: 2

spillover of marked features on arcs:
features smaller than threshold:
maximum leaf spillover= 0.0

node	length	children		label
1	59.9			CHAIR
2	45.8			SOFA
3	54.9			TABLE
4	65.1			DRESSER
5	54.1			DESK
6	118.3			BED
7	195.2			STOVE
8	115.1			BOOKCASE
9	99.2			FOOTSTOOL
10	128.7			LAMP
11	149.2			PIANO
12	100.8			MIRROR
13	161.9			RUG
14	90.1			RADIO
15	79.9			CLOCK
16	96.2			PICTURE
17	156.2			VASE
18	139.3			TELEPHONE
19	133.8			CLOSET
20	78.7			CUSHION
21	24.1	2	20	
22	38.8	3	5	
23	32.5	4	19	
24	68.0	12	16	
25	48.8	14	15	
26	22.6	1	21	
27	17.0	8	23	
28	5.2	24	17	
29	28.3	25	18	
30	39.5	26	9	
31	8.6	6	30	
32	0.0	11	29	
33	13.1	31	13	
34	11.1	22	27	
35	1.4	7	32	
36	4.5	10	28	
37	14.3	34	33	
38	0.0	35	36	
39	0.0	37	38	
40	47.2	20	13	"C"
41	47.7	14	11	"D"
42	38.8	1	22	"E"
43	32.5	12	23	"H"
44	7.0	21	13	"I"

45	39.6	14	18	"N"
46	17.1	21	6	"O"
47	22.4	26	6	"U"
48	7.1	12	27	"X"
49	23.7	16	17	"Z"



```
stress formula 1 = 0.0464
stress formula 2 = 0.2593
r(monotonic) squared=0.9328
r-squared (p.v.a.f.)=0.9058


final set of marked features:

feature  objects sharing feature
-------  ----------------------

   C     [ CUSHION, RUG, ]

   D     [ RADIO, PIANO, ]

   E     [ CHAIR, TABLE, DESK, ]

   H     [ MIRROR, DRESSER, CLOSET, ]

   I     [ SOFA, CUSHION, RUG, ]

   N     [ RADIO, TELEPHONE, ]

   O     [ SOFA, CUSHION, BED, ]

   U     [ CHAIR, SOFA, CUSHION, BED, ]

   X     [ MIRROR, DRESSER, BOOKCASE, CLOSET, ]

   Z     [ PICTURE, VASE, ]


model distances:
 129.8
 170.9 258.4
 269.4 279.4 208.3
 170.1 257.7 109.0 207.5
 195.5 171.3 244.8 265.8 244.1
 etc.
```

INTERPRETATION:
The structure of the tree makes sense. The tightest clusters seem due to either conceptual similarity (e.g. table-desk, mirror-picture), part-whole relationships (sofa-cushion), or real-world associations (radio-clock). Ditto for the marked features: some are due to conceptual similarity (cushion-rug), some to similarity of function (radio-telephone, cushion-rug), or real-world associations (mirror-dresser-closet).
Some of the marked features seem partly redundant. We could use backwards stepwise regression to prune the set of retained features.