```
1 #!/usr/bin/env lua
   -- vim : filetype=lua ts=2 sw=2 et :
10 -- (c) 2021 Tim Menzies. Permission is hereby granted, free of charge,
11 -- to any person obtaining a copy of this software and associated
  -- documentation files (the "Software"), to deal in the Software without
  -- restriction, including without limitation the rights to use, copy,
14 -- modify, merge, publish, distribute, sublicense, and/or sell copies
15 -- of the Software, and to permit persons to whom the Software is
16 -- furnished to do so, subject to the following conditions:
8 -- The above copyright notice and this permission notice shall be included in all
18 -- copies or substantial portions of the Software.
1 -- THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR 2 -- IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, 2 -- FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
24 -- AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
25 -- LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
26 -- OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
28 local help = [[
29 muse [OPTIONS]
31 Tree learner (binary splits on numerics using Gaussian approximation)
   (c) 2021 Tim Menzies <timm@ieee.org> MIT license.
   OPTIONS.
                 X Best examples are in 1..best*size(all)
      -best
                 X run one test, show stackdumps on fail
      -debug
                                                                       = pass
      -epsilon X
                     ignore differences under epsilon*stdev
                                                                       = .35
                 X How far to look for remove items
                                                                       = .9
      -Far
                 X Where to read data
                                                                       = ../../data/auto93.csv
      -file
                 X optimize, monitor, explore, challenge
                                                                       = optimize
      -goal
                      Show help
                      size of subset of a list
      -little
                 Х
                                                                       = 1024
                      Use more*#best for rest
                                                                        = 3.5
      -more
                     distance calc coefficient
                      Control for rounding numbers
      -round
      -seed
                     Random number seed;
                                                                       = 10019
                      Create subtrees while at least 2*stop egs = 4
      -Stop
                 X Min range size = size(eqs)^tiny
      -Tiny
                 X Pass/fail tests to run at start time
      -t.odo
                                                                       = pass
                      If "X=all", then run all.
                      If "X=ls" then list all.
                      Show low-level traces.
                                                                       = false
      -verbose
   Data read from "-file" is a csv file whose first row contains column
55 names (and the other row contain data. If a name contains ":",
    that column will get ignored. Otherwise, names starting with upper
   case denote numerics (and the other columns are symbolic). Names
   containing "!" are class columns and names containing "+" or "-"
59 are goals to be maximized or minimized. ]] --[[
61 Internally, columns names are read by a COLS object where numeric,
   symbolic, and ignored columns generate NUM, SYM, and SKIP instances
    (respectively). After row1, all the other rows are examples ('EG')
    which are stored in a SAMPLE. As each example is added to a sample,
   they are summarized in the COLS' objects.
   Note that SAMPLEs can be created from disk data, or at runtimes from
   lists of examples (see SAMPLE:clone()). --]]
   local b4={}; for k,_ in pairs(_ENV) do b4[k]=k end
   local THE = {} -- The THE global stores the global config for this software.
-- any line of help text startling with " -" has flag, default as first, last word
   help:gsub("\n [-]([^{\%}s]+)[^{\n}]*^{\%}s([^{\%}s]+)",
     function(flag,x)
        for n,word in ipairs(arg) do -- check for any updated to "flag" on command line -- use any command line "word" that matches the start of "flag"

if flag:match("^"...word:sub(2)...".*") then
             -- command line "word"s for booleans flip the default value x=(x=="false" and "true") or (x=="true" and "false") or arg[n+1] end end
        if x=="true" then x=true elseif x=="false" then x=false else x=tonumber(x) or x end
        THE[flag] = x end)
83 THE.seed = THE.seed or 10019
84 if THE.h then return print (help) end
```

```
90 local function same(x,...) return x end
 local function upto(x,y) return x < y end
local function over(x,y) return not(upto(x,y)) end
95 local function push(t,x) table.insert(t,x); return x end
96 local function sort(t,f) table.sort(t,f); return t end
97 local function ones(a,b) return a[1] < b[1] end</pre>
100 local copy, keys, map, sum
function copy(t, u) u={};for k,v in pairs(t) do u[k]=v end; return u function map(t,f, u) u={};for _,v in pairs(t) do u[1+#u] =f(v) end; return u
                                                                                                        end
                                                                                                        end
103 function sum(t,f, n) n=0 ;for _,v in pairs(t) do n=n+(f or same)(v) end;return n
                                                                                                        end
104 function keys(t, u)
105  u={}; for k,_ in pairs(t) do

106  if tostring(k):sub(1,1) ~= "_" then u[1+#u]=k end end;
107 return sort (u) end
109 -- printing utils
110 local fmt = string.format
local function say(...) if THE.verbose then print(fmt(...)) end end
local function btw(...) io.stderr:write(fmt(...).."\n") end
local function hue(n,s) return string.format("\27[1m\27[%sm\%s\27[0m",n,s) end
115 local o
116 local function out(x) print(o(x)) end
117 function o(t, u,f) -- convert nested tables to a string
     local function f(k) return fmt(":%s %s", hue(33,k), o(t[k])) end
if type(t) ~= "table" then return tostring(t) end
     u = #t>0 and map(t, o) or map(keys(t), f)
return hue(32,(t._is or "")).."{"..table.concat(u,"").."}" end
123 -- reading from file
124 local function coerce(x)
if x=="true" then return true elseif x=="false" then return false end
     return tonumber(x) or x end
128 local function csv(file, x,line)
function line(x, t)
       t=\{\}; for y in x:gsub("[\tau]*",""):gmatch"([\(^1,)\))" do push(t,coerce(y)) end
130
        return t end
      file = io.input(file)
132
      return function( x)
133
        if x then return line(x) else io.close(file) end end end
138 local log = math.log
139 local sqrt= math.sqrt
   local function rnd(x,d, n) n=10^{\circ}(d or THE.round); return math.floor(x*n+0.5) / n end
141 local function rnds(t,d)
     return map(t, function(x) return type(x) == "number" and rnd(x,d) or x end) end
144 -- random stuff (LUA's built-in randoms give different results on different platfors)
145 local rand
146 local function randi(lo,hi) return math.floor(0.5 + rand(lo,hi)) end
147 function rand(lo,hi)
    lo, hi = lo or 0, hi or 1

THE.seed = (16807 * THE.seed) % 2147483647

return lo + (hi-lo) * THE.seed / 2147483647 end
152 local function any(t) return t[randi(1,#t)] end
153 local function shuffle(t, j)
154 for i=#t,2,-1 do j=randi(1,i); t[i],t[j]=t[j],t[i] end; return t end
155 local function some(t,n, u)
156 if n >= #t then return shuffle(copy(t)) end
     u={}; for i=1, n do push(u, any(t)) end; return u end
159 -- objects
160 local function is(x) return getmetatable(x) end
161 local function as (mt, x) return setmetatable (x, mt) end
162 local function of (s. obi)
      obj = {_is=s, __tostring=o}
164
     obi. index = obi
     return as({ call=function(,...) return obj.new(...) end},obj) end
```

```
170 local SYM=of"SYM"
function SYM.new(inits,at,txt, i)
    i= as(SYM, {n=0, at=at or 0, txt=txt or "",
                  has={}, mode=nil, most=0})
     for _,x in pairs(inits or {}) do i:add(x) end
    return i end
177 -- Summarizing
178 function SYM.merge(i,j,
     k = SYM({},i.at, i.txt)
     for x,n in pairs(i.has) do k:add(x,n) end
     for x,n in pairs(j.has) do k:add(x,n) end
     return k end
184 function SYM.mid(i) return i.mode end
185 function SYM.spread(i)
    return sum(i.has, function(n) return -n/i.n*log(n/i.n,2) end) end
189 function SYM.add(i,x,n)
   if x ~= "?" then
       n = n \text{ or } 1
       in = n + in
       i.has[x] = (i.has[x] or 0) + n
       if i.has[x] > i.most then i.mode, i.most = x, i.has[x] end
194
       return x end end
197 -- querving
198 function SYM.dist(i,x,y) return x==y and 0 or 1 end
200 -- discretization
201 function SYM.splits(i, j,_,
                                  cut, tmp)
  function cut(x) return
      {val=x, at=i.at, txt=fmt("%s==%s",i.txt,x),
        when = function(z) return z==x end end
    return map (keys (i:merge (j).has), cut) end
206
207 --
208 --
         209 --
211 -- Columns for values we want to ignore.
212 local SKIP=of"SKIP"
213 function SKIP.new(inits,at,txt)
  return as (SKIP, {at=at or 0, txt=txt or ""}) end
216 function SKIP.mid(i)
                            return "?" end
217 function SKIP.spread(i) return 0 end
218 function SKIP.add(i,x)
                            return x end
219 function SKIP.splits(i,_) return {} end
220
221 --
225 local NUM=of"NUM"
  _has={},
                    mu=0, m2=0, lo=math.huge, hi=-math.huge})
    for _,x in pairs(inits or {}) do i:add(x) end
    return i end
234 -- summarizing
235 function NUM.mid(i) return i.mu end

236 function NUM.spread(i) return (i.m2/(i.n-1))^0.5 end
239 function NUM.add(i,x, d)
240 if x ~= "?" then
      push(i._has, x)
241
       i.n = i.n + 1
243
       d
             = x
                         - i.mu
      i.mu = i.mu + d/i.n

i.m2 = i.m2 + d*(x-i.mu)
      i.lo = math.min(x, i.lo)
      i.hi = math.max(x, i.hi) end
     return x end
250 function NUM.merge(i,j,
     k = NUM({}), i.at, i.txt)
     for _,v in pairs(i._has) do k:add(v) end
     for _,v in pairs(j._has) do k:add(v) end
     return k end
255
```

```
256 -- querying
257 function NUM.norm(i,x)
     return math.abs(i.hi - i.lo) < 1E-9 and 0 or (x-i.lo)/(i.hi-i.lo) end
260 function NUM.dist(i,x,y)
    if x=="?" then y=i:norm(y); x=y>0.5 and 0 or 1
elseif y=="?" then x=i:norm(x); y=x>0.5 and 0 or 1
      else x, y = i:norm(x), i:norm(y) end
      return math.abs(x-y) end
266 -- discretization
267 local spans
    function NUM.splits(i,j,
                                         xys, cuts)
      function cuts(x,s,at) return {
        \{val=x, at=at, txt=fmt("%s <= %s", s, rnd(x)), when=function(z) return z <= x end\}, \{val=x, at=at, txt=fmt("%s > %s", s, rnd(x)), when=function(z) return z > x end}\}
      xys={}
273
      for _,x in pairs(i._has) do push(xys, {x=x, y="best"}) end
274
      for _,x in pairs(j._has) do push(xys, {x=x, y="rest"}) end
      return spans(sort(xys, function(a,b) return a.x < b.x end),</pre>
                        (#xys) ^THE.Tiny,
                       THE.epsilon*(i.n*i:spread() + j.n*j:spread())/(i.n + j.n),
270
                       i,
SYM) end
280
```

page 4

```
283 --
286 -- Return a list of 'spans' {lo=,hi=,col=col}.
287 -- Sort the list of pairs 'xys' then split it into 'spans' of cardinally at
   -- least 'tiny'. Ensure that the max-min of each span is more that 'trivial'.
289 function spans(xys, tiny, trivial,col,yklass)
      local function mergeable(a,b, new,b4)
         new = a:merge(b)
         b4 = (a.n*a:spread() + b.n*b:spread()) / new.n
if new:spread() <= b4 then return new end
295
       local function merge(b4)
         local j, tmp = 0, {}
296
         while j < #b4 do
j = j + 1
297
298
299
            local now, after = b4[j], b4[j+1]
300
            if after then
301
               local simpler = mergeable(now.has, after.has)
302
               if simpler then
                 now = \{col=col, lo=now.lo, hi= after.hi, has=simpler\}

j = j + 1 end end
303
304
305
            push(tmp,now) end
306
         return #tmp==#b4 and b4 or merge(tmp)
         coal function div( spans,span)
span = {col=col,lo=xys[1].x, hi=xys[1].x, has=yklass()}
spans = {span}
for div
307
       end
       local function div(
308
309
310
         for j, xy in pairs (xys) do
  local x, y = xy.x, xy.y
  if j < #xys - tiny and
        x ~= xys[j+1].x and</pre>
311
312
                                                    -- enough items remaining after split
-- next item is different (so can split here)
313
314
                  span.has.n > tiny and -- span has enough items span.hi - span.lo > trivial -- span is not trivially small
315
316
            then span = push(spans, {col=col, lo=span.hi, hi=x, has=yklass()}) -- then new span
317
            end
318
            span.hi = x
319
            span.has:add(y)
320
321
          spans[1].lo = -math.huge
322
          spans[#spans].hi = math.huge
323
         return spans end
       return merge(div()) end
```

```
353 local EG=of"EG"
function EG.new(cells) return as(EG, {cells=cells}) end
356 -- Sumamrizing
357 function EG.cols(i,all)
return map(all, function(c) return i.cells[c.at] end) end
361 function EG.dist(i,j,cols, a,b,d,n,inc)
     d, n = 0, 0
     for _,col in pairs(cols) do
      a,b = i.cells[col.at], j.cells[col.at]
inc = a=="?" and b=="?" and 1 or col:dist(a,b)
      d = d + inc^THE.p
     n = n + 1 end
return (d/n)^(1/THE.p) end
370 -- Sorting
371 function EG.better(i,j,cols, e,
372 n,s1,s2,e = #cols, 0, 0, 2.71828
                                        e,n,a,b,s1,s2)
     for _,col in pairs(cols) do
      a = col:norm(i.cells[col.at])
374
       b = col:norm(j.cells[col.at])
375
      s1 = s1 - e^{(col.w * (a-b)/n)}

s2 = s2 - e^{(col.w * (b-a)/n)} end
     return s1/n < s2/n end
380 --
           381 --
382 --
383 --
384 -- SAMPLEs hold many examples
385 local SAMPLE=of"SAMPLE"
386 function SAMPLE.new(inits,
     i = as(SAMPLE, {cols=nil, egs={}})
389 if type(inits) =="string" then for eg in csv(inits) do i:add(eg) end end
389 if type(inits) =="stable" then for eg in pairs(inits) do i:add(eg) end end
392 -- Create a new sample with the same structure as this one
393 function SAMPLE.clone(i,inits, tmp)
     tmp = SAMPLE.new()
      tmp:add(i.cols.names)
     for _,eg in pairs(inits or {}) do tmp:add(eg) end
     return tmp end
400 function SAMPLE.add(i,eg)
     eg = eg.cells and eg.cells or eg
     if i.cols
     then push (i.egs, EG(eg)); i.cols:add(eg)
     else i.cols = COLS(eg) end end
407 function SAMPLE.neighbors(i,egl,egs,cols,
     dist_eg2 = function(eg2) return {eg1:dist(eg2,cols or i.cols.xs),eg2} end
      return sort (map (egs or i.egs, dist_eg2), ones) end
411 function SAMPLE.distance_farEg(i,eg1,egs,cols, tmp)
tmp = i:neighbors(eg1, egs, cols)
     tmp = tmp[#tmp*THE.Far//1]
     return tmp[2], tmp[1] end
416 -- Unsupervised discretization
417 function SAMPLE.best(i)
      local rest, div = {}
      function div(egs, lvl, one,
                                             tmp,a,b,c,two,want,low,good)
        tmp = i:clone(egs)
say("%s%s\t%s",
        string.rep("|. ",lv1), #egs,o(rnds(tmp:mid(tmp.cols.ys),1)))
if #egs < 2*(#i.egs)^THE.epsilon then
          return i:clone(egs), i:clone(some(rest,THE.more*#egs)) end
        one = one or i:distance_farEg(any(egs), egs, i.cols.xs)
        two,c = i:distance_farEg(one,
                                                       eas, i.cols.xs)
        for _,eg in pairs(egs) do
          a = eg:dist(one, i.cols.xs)
          b = eg:dist(two, i.cols.xs)
eg.x = (a^2 + c^2 - b^2)/(2*c) end
        low = one:better(two,i.cols.ys)
        good = {}
        for n,eg in pairs(sort(egs,function(a,b) return a.x < b.x end)) do
  if n < .5*#egs then push(low and good or rest, eg)</pre>
                           else push (low and rest or good, eg) end end
        return div(good, lvl+1,two) end
      return div(same(i.egs,THE.little), 0) end
```

page 8

## 

```
function SAMPLE:splits(other,both, place,score)
function place(eg,cuts, x)

for _,cut in pairs(cuts) do
cut.has = cut.has or self:clone()
x = eg.cells[cut.at]
if x = "?" and cut.when(x) then return cut.has:add(eg) end end end
function score(cut, m,n)
m,n = #(cut.has.egs), #both.egs; print(m,n); return -m/n*log(m/n,2) end
local best, cutsx, cuts, tmp = math.huge
for pos,col in pairs(both.cols.xs) do
print("eps", col.at, col:spread()*THE.epsilon)
cutsx = col:splits(other.cols.xs[pos], col:spread()*THE.epsilon)
for _,eg in pairs(both.egs) do place(eg, cutsx) end
tmp = sum(cutsx, score)
if tmp < best then best,cuts = tmp,cutsx end end
return cuts end</pre>
```

```
471 --
             472 --
474 --
475 local go={}
476 function go.pass() return true end
function go.base() return true end () function go.base() seq(THE); say("%s",o(s)) end function go.bad( s) assert(false) end
480 function go.sort( u,t)
481 t={}; for i=100,1,-1 do push(t,i) end
      t=sort(t, function(x,y)
           if x+y<20 then return x>y else return x<y end end)</pre>
       assert(sum(t, function(x) return x*100 end) == 505000)
       assert (t[1] == 10)
       assert (t[#t]==100)
       u=copy(t)
      t[1] = 99
488
      assert(u[1] ~= 99) end
491 function go.file(n)
      for _,t in pairs{{"true",true,"boolean"}, {"false",false,"boolean"}, {"42.1",42.1,"number"}, {"32zz","32zz","string"}, {"nil","nil","string"}} do
         assert (coerce (t[1]) ==t[2])
405
         assert (type (coerce (t[1])) == t[3]) end
406
497
      n = 0
       for row in csv(THE.file) do
498
499
         n = n + 1
         assert (#row==8)
500
         assert (n==1 or type (row[1]) == "number")
         assert (n==1 or type (row[8]) == "number") end end
504 function go.rand(t,u)
      t,u={},{}; for i=1,20 do push(u,push(t,100*rand())) end
t= sort(rnds(t,0))
      assert(t[1]==3 and t[#t]==88)
      t= sort(some(t, 4))
508
      assert (#t==4)
      assert (t[1]==7)
      assert (79.5 == rnds(shuffle(u))[1])
511
function go.num( cut,min, z,r1,r2,x,y)

z = NUM(9,2,5,4,12,7,8,11,9,3,7,4,12,5,4,10,9,6,9,4)

assert(7 == z:mid(), 3.06 == rnd(z:spread(),2))
      x, y = NUM(), NUM()

for i=1,20 do x:add(rand(1,5)) end
      for i=1,20 do y:add(randi(20,30)) end end
function go.sym( cut,min,w,z)

w = SYM{"m","m","m","m","b","b","c"}

z = SYM{"a","a","a","a","b","b","c"}
      assert (1.38 == rnd(z:spread(),2))
       for _, cut in pairs(w:splits(z)) do say("%s",o(cut)) end
528 function go.sample( s,egs,xs,ys,scopy)
      s=SAMPLE(THE.file)
      scopy=s:clone(s.egs)
say("%s %s",s.cols.all[1]:spread(), scopy.cols.all[1]:spread())
      xs,ys= s.cols.xs, s.cols.ys
      assert (4 == #xs)
assert (3 == #ys)
       egs=s:sorted()
       say(o(rnds(s:mid(ys),1)))
537
       say(o(rnds(map(s:spread(ys), function(x) return .35*x end), 1)));say("")
       for i=1,10 do say("%s", o(rnds(egs[i]:cols(ys),1))) end;
538
      for i=#eqs, #eqs-10,-1 do say(o(rnds(eqs[i]:cols(ys),1))) end
539
542 function go.dist( s,xs,sorted, show)
      s=SAMPLE(THE.file)
       xs= s.cols.xs
      sorted = s:neighbors(s.egs[1], s.egs,xs)
show=function(i) say("%s%s",rnd(sorted[i][1],2),
                                      o(sorted[i][2]:cols(xs))) end
do show(i) end; say("")
       for i = 1.10
      for i=#sorted-10, #sorted do show(i) end end
function go.far( s,xs,d,eg2)

s = SAMPLE(THE.file)

s = s.cols.xs
       for k,eg1 in pairs(shuffle(s.egs)) do
  if k > 10 then break end
555
          eq2,d = s:distance_farEq(eq1, s.eqs, xs)
556
         say("%s %s %s", rnd(d), o(eg1:cols(xs)), o(eg2:cols(xs))) end end
557
559 local goals
```

page 10

```
560 function goals.optimize(b,r) if b+r>1E-2 and b>r then return b^2/(b+r+1E-31) end end
561 function goals.monitor(b,r) if b+r>1E-2 and r>b then return r^2/(b+r+1E-31) end
562 function goals.explore(b,r) if b+r>1E-2 then return 1/(b+r+1E-31) end
563 function goals.challenge(b,r) if b+r>1E-2
                                                                     then return 1/(math.abs(b-r)+1E-31) end
565 function go.best( all,best,rest)
       all = SAMPLE(THE.file)
       best, rest = all:best()
       say(o(best.cols.all[1]))
       say("%s%s", #best.egs, #rest.egs)
       for n, coll in pairs (best.cols.xs) do
         if is(col1) == NUM then
    say""
573
            for n, cut in pairs(col1:splits(rest.cols.xs[n])) do
574
              say(o{n,col1.txt,cut.lo,cut.hi}) end end end
575
576
       end
577 --
578 --
579 --
580 --
581 local fails, defaults, todos, ok, msg
582 fails, defaults = 0, copy(THE)
583 go[ THE.debug ]()
ok,msg = pcall( go[todo])

if ok then btw("%s%s",hue(32,"—PASS"),todo)

else btw("%s%s %s",hue(31,"—FAIL"),todo,msg); fails=fails+1 end end
btw (hue (33, "-- %s error(s)"), fails)
for k, v in pairs (_ENV) do

if not b4[k] then btw(hue (31, "-- rogue? %s %s"), k, type (v)) end end
595 os.exit(fails)
```