```
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,
  -- to any person obtaining a copy of this software and associated
  -- documentation files (the "Software"), to deal in the Software without
13 -- 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.
27 -- THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR 22 -- IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, 25 -- 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
27 -- SOFTWARE
28 local help = [[
29 muse [OPTIONS]
   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
                     run one test, show stackdumps on fail
      -debug
                                                                        = pass
      -epsilon X
                      ignore differences under epsilon*stdev
                                                                        = .35
                      How far to look for remove items
                                                                        = .9
      -Far
                  X Where to read data
                                                                        = ../../data/auto93.csv
      -file
                      Show help
      -h
      -little
                X size of subset of a list
                                                                        = 1024
                      distance calc coefficient
                                                                        = 2
      -p
      -round
                      Control for rounding numbers
      -seed
                      Random number seed;
                                                                        = 10019
                  X Create subtrees while at least 2*stop egs = 4
      -Tiny
                  X Min range size = size(eqs)^tiny
                                                                       = .5
                  X Pass/fail tests to run at start time
                                                                        = pass
      -todo
                      If "X=all", then run all.
                       If "X=ls" then list all.
   Data read from "-file" is a csv file whose first row contains column
   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
55 containing "!" are class columns and names containing "+" or "-"
   are goals to be maximized or minimized. ]] --[[
58 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)
an THE seed = THE seed or 10019
81 if THE.h then return print (help) end
```

```
87 local same
88 function same(x,...) return x end
91 local push, sort, ones
   function push(t,x) table.insert(t,x); return x end
function sort(t,f) table.sort(t,f); return t end
function ones(a,b) return a[1] < b[1] end</pre>
97 local copy, keys, map, sum
98 function copy(t, u) u={}; for k, v in pairs(t) do u[k]=v
                                                                                   end: return u
99 function keys(t, u) u={};for k,_ in pairs(t) do u[1+#u]=k
100 function map(t,f, u) u={};for k,v in pairs(t) do u[1+#u] =f(k,v)
                                                                                   end: return sort(u) end
                                                                                   end: return 11
                                                                                                           end
function sum(t,f, n) n=0; for _,v in pairs(t) do n=n+(f or same)(v) end; return n
103 -- printing utils
104 local hue, shout, out, say, fmt, btw
105 fmt = string.format
function say(...) print(string.format(...)) end
or function btw(...) io.stderr:write(fmt(...)."\n") end

function hue(n,s) return string.format("\27[1m\27[\sim \8]\27[0m",n,s) end
function shout (x) print (out (x)) end
function out(t, u,key,val) -- convert nested tables to a string
function key(_,k) return string.format(":% %s", k, out(t[k])) end
      function val(_,v) return out(v) end
if type(t) ~= "table" then return tostring(t) end
     u = #t>0 and map(t, val) or map(keys(t), key)
return (t._is or "").."{"..table.concat(u, "").."}" end
114
117 -- reading from file
118 local coerce, csv
119 function coerce(x)
    if x=="true" then return true elseif x=="false" then return false end
      return tonumber(x) or x end
123 function csv(file,
    file = io.input(file)
124
      return function( t,tmp)
       x = io.read()
126
        if x then -- kill space, split on ",", return non-empty lines
t={};for y in x:gsub("[t|*",""):gmatch"([^,]+)" do push(t,coerce(y)) end
127
128
           if #t>0 then return t end
129
        else io.close(file) end end end
132 -- maths
133 local log, sqrt, rnd, rnds
134 log = math.log
135 sqrt= math.sqrt
function rnd(x,d, n) n=10^(d or THE.round); return math.floor(x*n+0.5) / n end
   function rnds(t,d)
   return map(t, function(_,x) return type(x) == "number" and rnd(x,d) or x end) end
140 -- random stuff (LUA's built-in randoms give different results on different platfors)
141 local randi, rand, any, some, shuffle
   function randi(lo, hi) return math.floor(0.5 + rand(lo, hi)) end
143 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
148 function any(t)
                             return t[randi(1, #t)] end
149 function some (t, n,
                            u)
if n >= #t then return shuffle(copy(t)) end
     u={}; for i=1,n do push(u,any(t)) end; return u end
finction shuffle(t, j)
for i=#t,2,-1 do j=randi(1,i); t[i],t[j]=t[j],t[i] end; return t end
156 -- objects
157 local ako, has, obj
158 ako= getmetatable
159 function has (mt, x) return setmetatable (x, mt) end
function obj(s, o,new)
o = {_is=s, __tostring=out}
      o. index = o
      return setmetatable(o, {__call=function(_,...) return o.new(...) end}) end
```

```
168 local NUM=obj"NUM"
   function NUM.new(inits,at,txt,
     self = has(NUM, {n=0, at=at or 0, txt=txt or "",
                        w=(txt or ""):find"-" and -1 or 1,
                        mu=0, m2=0, lo=math.huge, hi=-math.huge})
     for _,x in pairs(inits or {}) do self:add(x) end
      return self end
176 -- summarizing
function NUM:mid() return self.mu end function NUM:spread() return (self.m2/(self.n-1))^0.5 end
180 -- updating
181 function NUM:add(x, d)
    if x ~= "?" then
        self.n = self.n + 1
        d = x - self.mu

self.mu = self.mu + d/self.n
        self.m2 = self.m2 + d*(x-self.mu)
        self.lo = math.min(x, self.lo)
        self.hi = math.max(x, self.hi) end
     return x end
191 -- querving
192 function NUM:norm(x,
                                lo,hi)
     lo, hi = self.lo, self.hi
      return math.abs(hi - lo) < 1E-9 and 0 or (x-lo)/(hi-lo) end
196 function NUM:dist(x,y)
     if x=="?" then y=self:norm(y); x=y>0.5 and 0 or 1
elseif y=="?" then x=self:norm(x); y=x>0.5 and 0 or 1
      else x, y = self:norm(x), self:norm(y) end
      return math.abs(x-y) end
202 -- discretization
   local roots
203
   function NUM:splits(other,
                                         cuts, cut)
      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}}
208
      cut = roots(self:mid(), other:mid(), self.n, other.n, self:spread(), other:spread())
209
     return cuts (cut, self.txt, self.at) end
212
    function _roots(m1, m2, n1, n2, s1, s2,
                                                  default, a,b,c,root1,root2)
      if m1 > m2 then -- make sure our comparison goes the right way
         return _roots(m2, m1, n2, n1, s2, s1) end
     if s1 ~=s2 then -- try the roots

a = 1/(2*s1^2) - 1/(2*s2^2)

b = m2/(s2^2) - m1/(s1^2)
        c = \frac{m1}{2} / (2*s1^2) - \frac{m2}{2} / (2*s2^2) - \log(s2/s1)
root1 = \frac{(-b - sqrt(b*b - 4*a*c))}{(2*a)}
        root2 = (-b + sqrt(b*b - 4*a*c))/(2*a)
          - return something in between the means
        if m1<= root1 and root1<=m2 then return root1 end</pre>
        if m1<= root2 and root2<=m2 then return root2 end end
      -- else return the weighted sum
      return (n1*m1+n2*m2)/(n1+n2) end
```

```
231 local SYM=obj"SYM"
232 function SYM.new(inits,at,txt, self)
    self= has(SYM, {n=0, at=at or 0, txt=txt or "",
                     seen={}, mode=nil, most=0})
    seen={}, mode=nil, most=0})
for _,x in pairs(inits or {}) do self:add(x) end
236 return self end
238 -- Summarizing
239 function SYM:mid() return self.mode end
240 function SYM:spread()
    return sum(self.seen, function(n) return -n/self.n*log(n/self.n,2) end) end
243 -- update
244 function SYM:add(x)
245 if x ~= "?" then
       self.n = 1 + self.n
       self.seen[x] = (self.seen[x] or 0) + 1
       if self.seen[x] > self.most then self.mode, self.most = x, self.seen[x] end
       return x end end
251 -- querying
252 function SYM:dist(x,y) return x==y and 0 or 1 end
253
254 -- discretization
255 function SYM:splits(other,
                                   cut, out)
    function cut( ,x) return
      {val=x, at=self.at, txt=fmt("%s==%s", self.txt,x),
        when = function(z) return z==x end end
258
250
     for k,_ in pairs(self.seen) do out[k]=k end
     for k, in pairs (other. seen) do out [k]=k end
     return map (sort (out), cut) end
264 --
265 --
268 -- Columns for values we want to ignore.
269 local SKIP=obj"SKIP"
270 function SKIP.new(inits,at,txt)
return has (SKIP, {n=0, at=at or 0, txt=txt or ""}) end
                           return "?" end
274 function SKIP:spread() return 0 end
275 function SKIP:add(x) return x
   function SKIP:splits(_) return {}
282 -- Convert column headers into NUMs and SYMs, etc.
283 local COLS=obj"COLS"
284 function COLS.new(names,
                             self, new, what)
   self = has(COLS, {names=names, xs={}}, all={}, ys={}})
     for n,x in pairs (names) do
      new = (x:find":" and SKIP or x:match"^[A-Z]" and NUM or SYM)({},n,x)
       push(self.all, new)
       if not x:find":" then
        if x:find"!" then self.klass = new end
         what = (x:find"-" or x:find"+") and "ys" or "xs"
         push(self[what], new) end end
    return self end
295 -- Updates
296 function COLS:add(eg)
     return map (eq, function (n, x) self.all[n]:add(x); return x end) end
```

```
304 local EG=obj"EG"
305 function EG.new(cells) return has(EG, {cells=cells}) end
308 function EG:cols(all)
return map(all, function(_,c) return self.cells[c.at] end) end
311 -- Queries
function EG:dist(other,cols, a,b,d,n,inc)
     d, n = 0, 0
     for _,col in pairs(cols) do
      a,b = self.cells[col.at], other.cells[col.at]
inc = a=="?" and b=="?" and 1 or col:dist(a,b)
      d = d + inc^THE.p
317
      n = n + 1 \text{ end}
     return (d/n)^(1/THE.p) end
319
321 -- Sorting
322 function EG:better(other,cols,
                                        e,n,a,b,s1,s2)
     n,s1,s2,e = #cols, 0, 0, 2.71828
     for _, num in pairs(cols) do
      a = num:norm(self.cells[ num.at])
       b = num:norm(other.cells[num.at])
326
      s1 = s1 - e^(num.w * (a-b)/n)
s2 = s2 - e^(num.w * (b-a)/n) end
327
     return s1/n < s2/n end
331 --
          (_ /\ |\/| |_) |
332 --
333 --
334 --
335 -- SAMPLEs hold many examples
336 local SAMPLE=obj"SAMPLE"
337 function SAMPLE.new(inits,
                                   self)
     self = has(SAMPLE, {cols=nil, eqs={}})
     if type (inits) == "string" then for eg in csv(inits) do self:add(eg) end end if type (inits) == "table" then for eg in pairs(inits) do self:add(eg) end end
343 -- Create a new sample with the same structure as this one
344 function SAMPLE:clone(inits, out)
     out = SAMPLE.new()
     out:add(self.cols.names)
     for _,eg in pairs(inits or {}) do out:add(eg) end
    function SAMPLE:add(eq)
     eg = eg.cells and eg.cells or eg
     if self.cols
     then push(self.egs, EG(eg)); self.cols:add(eg)
     else self.cols = COLS(eg) end end
   function SAMPLE: neighbors (eg1, egs, cols,
                                                      dist eq2)
     dist_eg2 = function(_,eg2) return {eg1:dist(eg2,cols or self.cols.xs),eg2} end
     return sort (map (egs or self.egs, dist_eg2), ones) end
362 function SAMPLE:distance_farExample(eg1,egs,cols,
     tmp = self:neighbors(eg1, egs, cols)
     return table.unpack(tmp[#tmp*THE.Far//1]) end
   -- Unsupervised discretization
   function SAMPLE: twain (egs, cols,
                                                  _, north, south, a, b, c, lo, hi)
     _,north = self:distance_farExample(any(egs), egs, cols)
      c, south = self:distance_farExample(north, egs, cols)
     for _,eg in pairs(egs) do
       a = eg:dist(north, cols)
       b = eg:dist(south, cols)
        eg.x = (a^2 + c^2 - b^2)/(2*c) end
     lo, hi = self:clone(), self:clone()
     for n,eg in pairs(sort(egs, function(a,b) return a.x < b.x end)) do
  if n < .5*#egs then lo:add(eg) else hi:add(eg) end end</pre>
     return lo, hi end
379 function SAMPLE:mid(cols)
     return map(cols or self.cols.all, function(_,col) return col:mid() end) end
    function SAMPLE:spread(cols)
382
     return map(cols or self.cols.all,function(,col) return col:spread() end) end
385 function SAMPLE sorted()
     self.egs= sort(self.egs, function(eg1,eg2) return eg1:better(eg2,self.cols.ys) end)
     return self.egs end
387
388
```



```
393 -- need to sort first
function SAMPLE:splits(other,both,
                                           guess, divide, cuts, unplaced, place, score)
    function guess (todos, cuts,
       for _,todo in pairs(todos) do
          f=function(_, cut)
                    return {Row(cut.has:mid()):dist(todo, both.cols.xs),cut} end
          sort(map(cuts, f), ones)[1][2].has:add(todo) end
       return cuts end
      function divide (cuts,
                               todos.placed)
        todos = {}
        for _,eg in pairs(both.eqs) do
          placed = false
          for _,cut in pairs(cuts) do
407
408
            if cut.when(eg.cells[cut.at])
            then cut.has = cut.has or self:clone()
409
                 cut.has:add(eq)
411
                 placed = true
412
                 break end end
413
          if not placed then push (todos, eq) end end
        return guess (todos, cuts) end
     function score(cut, m,n)
m,n = #(cut.has.eqs), #both.eqs; return -m/n*log(m/n,2) end
     local best, cutsx, tmp = math.huge
for pos,col in pairs(both.cols.xs) do
          cutsx = col:splits(other.cols.xs[pos])
          tmp = sum(divide(cutsx), score)
420
          if tmp < best then best, cuts = tmp, cutsx end end
421
     return cuts end
424 function SAMPLE: tree (top)
425 top = top or self
     one, two = self:twain(self.eqs, top.cols.xs)
     for _, cut in pairs (one:splits (two, self)) do
       if cut.stats.n > (#top.eqs) THE.Tiny then
          cut.sub= cut.has:tree(top) end end end
431 function SAMPLE: show (tree,
      vals=function(a,b) return a.val < b.val end
      function show1(tree,pre)
        if #tree.kids==0 then io.write(fmt("==> %s[%s]",tree.mode, tree.n)) end
434
        for _,kid in pairs(sort(tree.kids,vals)) do
435
            io.write("\n"..fmt("%%%s",pre, showDiv(i, kid.at, kid.val)))
show1(kid.sub, pre.."|..") end
      show1(tree, ""); print("") end
```

```
442 --
            443
446 local go={}
   function go.ls()
     print ("\nlua " ..arg[0] .." -todo ACTION\n\nACTIONS:")
      for _,k in pairs(keys(go)) do print(" -todo",k) end end
451 function go.pass() return true end
   function go.the() shout(THE) end
   function go.bad( s) assert(false) end
   function go.sort( u,t)
455
     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)
assert(sum(t,function(x) return x*100 end)==505000)
     assert(t[1] == 10)
assert(t[#t]==100)
     u=copy(t)
t[1] = 99
      assert(u[1] ~= 99) end
466 function do out ( s)
      assert ("{:age 21 :milestones {1 2 3 4} :name tim}"==out (
             {name='tim', age=21, milestones={1,2,3,4}}))end
470 function go.file(n)
     473
        assert (coerce (t[1]) == t[2])
474
        assert (type (coerce (t[1])) ==t[3]) end
475
476
      n = 0
      for row in csv(THE.file) do
477
        n = n + 1
478
        assert (#row==8)
479
        assert (n==1 or type(row[1]) =="number")
assert (n==1 or type(row[8]) =="number") end end
483 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))
      assert (#t==4)
      assert(t[1]==7)
      assert (79.5 == rnds(shuffle(u))[1])
   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))
     r2 = _roots(2.5, 5, 20,10,1.1, .9)

assert(rnd(r2,2)==3.8)
      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
      for _, cut in pairs (x:splits(y)) do shout (cut) end end
503 function go.sym( cut, min, w, z)
     w= sym{"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 shout (cut) end end
509 function go.sample( s,egs,xs,ys,scopy)
     s=SAMPLE(THE.file)
      scopy=s:clone(s.egs)
      print(s.cols.all[1]:spread(), scopy.cols.all[1]:spread())
      xs,ys= s.cols.xs, s.cols.ys
      assert (4 == #xs)
      assert (3 == #ys)
      eas=s:sorted()
      shout (rnds (s:mid(ys),1));
518
      shout (rnds (map (s:spread (ys), function (_,x) return .35*x end), 1)); print ("")
      for i=1,10 do shout(rnds(egs[i]:cols(ys),1)) end; print("")
520
      for i=#egs, #egs-10,-1 do shout (rnds(egs[i]:cols(ys),1)) end end
function go.dist( s,xs,sorted, show)
s=SAMPLE(THE.file)
      xs= s.cols.xs
      for k,eq2 in pairs(s.eqs) do
        if k > 20 then break end
      print(s.egs[1]:dist(eg2, xs)) end
sorted = s:neighbors(s.egs[1], s.egs,xs)
      show=function(i)print(rnd(sorted[i][1],2), out(sorted[i][2]:cols(xs))) end
      for i=1.10
                                  do show(i) end: print("")
```

```
for i=#sorted-10, #sorted do show(i) end end
function go.far( s,xs,d,eg2)
    s = SAMPLE(THE.file)
     xs = s.cols.xs
     for k,eg1 in pairs(shuffle(s.egs)) do
       if k > 10 then break end
       d,eg2 = s:distance_farExample(eg1, s.egs, xs)
       print(rnd(d), out(eq1:cols(xs)), out(eq2:cols(xs))) end end
   function go.twain( s,lo,hi)
    s = SAMPLE(THE.file)
     lo, hi = s:twain(s.egs, s.cols.xs)
     print(#lo.egs, #hi.egs)
547 function go.splits( s,lo,hi)
548 s = SAMPLE(THE.file)
    lo, hi = s:twain(s.egs, s.cols.xs)
    for _, cut in pairs(lo:splits(hi,s)) do print(cut.txt) end
function go.kordered( s,n)
     s = ordered(slurp())
     n = #s.egs
     shout (s.heads)
     for i=1,15 do shout(s.egs[i].cells) end
     print ("#")
     for i=n,n-15,-1 do shout(s.eqs[i].cells) end end
561 function go.ksymcuts( s,xpect,cuts)
     s=ordered(slurp())
     print (out (s.cols.xs), out (s.cols.ys))
     xpect,cuts = symcuts(7,s.egs, "origin")
for _,cut in pairs(cuts) do print(xpect, out(cut)) end end
567 function go.knumcuts( s,xpect,cuts)
     s=ordered(slurp())
     xpect, cuts = numcuts(s, 2, s.egs, "Dsiplcment")
     if xpect then
       for ,cut in pairs(cuts) do print(xpect, out(cut)) end end end
573 function go.katcuts(s,cuts,at,ynum)
    s=ordered(slurp())
     ynum=NUM(a); map(s.eqs, function(_,eq) add(ynum, eq.klass) end)
     at, cuts = at cuts(s,eqs,sd(ynum)*THE.epsilon, (#s.eqs)^THE.Tiny)
    for _,cut in pairs(cuts) do print(at, out(cut)) end end
                 582 local fails, defaults, todos, ok, msq
583 fails, defaults = 0, copy (THE)
584 go[ THE.debug ] ()
586 todos = THE.todo == "all" and keys(go) or {THE.todo}
587 for _, todo in pairs(todos) do
     THE = copy(defaults)
     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
593 btw(hue(33,"-- %s errors"), fails)
594 for k, v in pairs (_ENV) do
if not b4[k] then btw(hue(31,"--rogue?%s%s"),k,type(v)) end end
596 os.exit(fails)
```

page 8