


```

247 -----
248 --- SUPER RANGES
249
250
251
252 function nb3(data, log)
253   local tmp, xnums = {}
254   local function discretize(c,x, col)
255     if x ~= "?" then
256       col = xnums[c]
257       if col then
258         for _,bin in pairs(col.bins) do
259           if bin.lo <= x and x < bin.hi then return bin.id end end end end
260         return x end
261       local function xnum(c,name)
262         if ako.xnum(name) then return {name=name, xys={}, bins={}} end end
263       local function train(c,x,row)
264         if xnums[c] and x ~= "?" then push(xnums[c].xys, {x=x,y= row[row]}) end end
265       -- start
266       for row in items(data) do
267         if xnums then collect(row, function(c,x) return train(c,x,row) end)
268         else xnums = collect(row,xnum) end end
269       for where,col in pairs(xnums) do col.bins = bins(col.xys,where); print(col.name
270 e,#col.bins) end
271       for j=2,#tmp do tmp[j] = collect(tmp[j], discretize) end
272       return nb1(tmp)
273     end
274
275 --- Find bins
276
277
278 local argmin
279 function bins(xys,where)
280   xys = sort(xys, upx)
281   local triviallySmall = the.cohen*(per(xys,.9).x - per(xys,.1).x)/2.56
282   local enoughItems = #xys / the.bins
283   local out = {}
284   argmin(1,#xys, xys, triviallySmall, enoughItems, -math.huge, where, out)
285   out[#out].hi = math.huge
286   return out end
287
288 function argmin(lo, hi, xys, triviallySmall, enoughItems, b4, where, out)
289   local function add(f,z) f[z] = (f[z] or 0) + 1 end
290   local function sub(f,z) f[z] = f[z] - 1 end
291   local lhs, rhs, cut, div, xpect, xy = {},{}
292   for j=lo,hi do add(rhs, xys[j].y) end
293   div = ent(rhs)
294   if hi-lo+1 > 2*enoughItems then
295     for j=lo,hi - enoughItems do
296       add(lhs, xys[j].y)
297       sub(rhs, xys[j].y)
298       local n1,n2 = j - lo +1, hi-j
299       if n1 > enoughItems and
300          n2 > enoughItems and
301          xys[j].x ~= xys[j+1].x and -- there is a break here
302          xys[j].x - xys[lo].x > triviallySmall and
303          xys[hi].x - xys[j].x > triviallySmall
304       then xpect = (n1*ent(lhs) + n2*ent(rhs)) / (n1+n2)
305          if xpect < div then -- cutting here simplifies things
306            cut, div = j, xpect end end end
307     end -- end if
308     if cut
309     then b4 = argmin(lo, cut, xys, triviallySmall, enoughItems, b4, where, out)
310          b4 = argmin(cut+1, hi, xys, triviallySmall, enoughItems, b4, where, out)
311     else b4 = push(out,{id=cut+1, where=where,
312                    lo=b4, hi=xys[hi].x, n=hi-lo+1, div=div}).hi end
313   return b4 end
314 -----
315 --- nums and syms
316 ---
317 ---
318
319 function create(names)
320   local i = 1,cols()
321   i.names = names
322   for at,name in pairs(names) do
323     local now = ako.num(name) and it.num() or it.sym()
324     now.at, now.name, now.w = at, name, ako.weight(name)
325     push(i.xy, now)
326     if not ako.ignore(name) then
327       if not ako.goal(name) then now.indep = true end
328       if ako.klass(name) then i.klass=now end
329       push(now.indep and i.x or i.y, now) end end
330   return i end
331
332 function update(i,row)
333   local function num(col,x, d)
334     col.lo = math.min(x, col.lo)
335     col.hi = math.max(x, col.hi)
336     d = x - col.mu
337     col.mu = col.mu + d/col.n
338     col.m2 = col.m2 + d*(x - col.mu)
339     col.sd = ((col.m2/col.n < 0 or col.n < 2) and 0) or ((col.m2/(col.n - 1))^0.5) end
340   local function sym(col,x)
341     col.has[x] = 1 + (col.has[x] or 0)
342     if col.has[x] > col.most then
343       col.mode,col.most = x,col.has[x] end end
344   -- start
345   for _,col in pairs(i.cols.xy) do
346     local x = row[col.at]
347     if x ~= "?" then
348       col.n = col.n + 1
349       (col.nump and num or sym)(col,x) end end
350   return row end
351
352 function mid(i,cols)
353   local function mid(col) return col.nump and col.mu or col.mode end
354   return map(cols or i.cols.y, mid) end
355
356 function div(i,cols)
357   local function div(col) return col.nump and col.sd or ent(col.has) end
358   return map(cols or i.cols.y, div) end
359
360 function clone(old,rows)
361   local i={rows={}, cols=create(old.cols.names)}
362   for _,row in pairs(rows or {}) do update(i,row) end
363   return i end
364
365 function sorted(i)
366   return sort(i.rows, function(a,b) return better(i,a,b) end) end
367
368 function better(i,row1,row2)
369   local s1, s2, n, e = 0, 0, #i.cols.y, math.exp(1)
370   for _,col in pairs(i.cols.y) do
371     local a = norm(col.lo, col.hi, row1[col.at])
372     local b = norm(col.lo, col.hi, row2[col.at])
373     s1 = e^(col.w * (a - b) / n)
374     s2 = e^(col.w * (b - a) / n) end
375   return s1 / n < s2 / n end
376 -----
377 ---
378 ---
379 ---
380 ---
381
382 function xplain(data)
383   local i = {rows={}, cols=nil}
384   local function num(col, best, rest)
385     local tmp = {}
386     for klass,rows in pairs(rest,best) do
387       for _,row in pairs(rows) do
388         local x = row[col.at]
389         if x ~= "?" then push(tmp, {x=x, y=klass}) end end end
390     return bins(tmp, col.at)
391   end
392   local function sym(col, best, rest)
393     local tmp = {}
394     for klass,rows in pairs(rest,best) do
395       for _,row in pairs(rows) do
396         local x = row[col.at]
397         if x ~= "?" then
398           tmp[x] = tmp[x] or {id=x, where=col.at, lo=x, hi=x, n=0, tmp={}}
399           local r = tmp[x].tmp
400           r[klass] = 1 + (r[klass] or 0) end end end
401     for x,t in pairs(tmp) do t.div,t.n = ent(t.tmp) end
402     return tmp
403   end
404   for row in items(data) do
405     if not i.cols then i.cols=create(row) else push(i.rows,update(i,row)) end end
406   d
407   i.rows = sorted(i)
408   local n = (#i.rows)^the.best
409   local best = slice(i.rows, 1, n)
410   local rest = many(i.rows, n*the.rest, n+1)
411   for _,col in pairs(i.cols.x) do
412     print""
413     print(col.at)
414     map((col.nump and num or sym)(col, best, rest),oo) end
415   return i end
416
417 function dist(i,row1,row2)
418   local function sym(_,x,y) return x==y and 0 or 1 end
419   local function num(c,x,y)
420     x=="" then y = norm(c.lo, c.hi, y); x==<.5 and 1 or 0
421     elseif y=="" then x = norm(c.lo, c.hi, x); y==<.5 and 1 or 0
422     else x,y = norm(c.lo, c.hi, x), norm(c.lo, c.hi, y) end
423     return math.abs(x-y) end
424   local function dist(c,x,y)
425     return x=="" and y=="" and 1 or (c.nump and num or sym)(c,x,y) end
426   local d, n = 0, #i.cols.x
427   for _,c in pairs(i.cols.x) do d = d + dist(c, row1[c.at], row2[c.at])^the.e end
428   return (d/n)^(1/the.e) end
429 -----
430 --- MISC
431 ---
432 ---
433 ---
434 ---
435 ---
436 ---
437 function per(t,p) return t[ (p or .5)*#t//1 ] end
438
439 function ent(t)
440   local n=0; for _,m in pairs(t) do n = n+m end
441   local e=0; for _,m in pairs(t) do if m>0 then e = e+m/n*math.log(m/n,2) end end
442   return -e,n end
443
444 function norm(lo,hi,x) return math.abs(hi-lo)<1E-9 and 0 or (x-lo)/(hi - lo) end
445
446 ---
447 ---
448 ---
449 function ish(x,y,z) return math.abs(x-y) <= (z or 0.001) end
450
451 local fails=0
452 function ok(test,msg)
453   print("", test and "PASS"or "FAIL",msg or "")
454   if not test then
455     fails = fails+1
456     if the and the.dump then assert(test,msg) end end end
457
458 function rogues()
459   for k,v in pairs(_ENV) do if not b4[k] then print("?",k,type(v)) end end end
460
461 ---
462 ---
463 ---
464 ---
465 ---
466 ---
467 ---
468 ---
469 ---
470 ---
471 ---
472 ---
473 ---
474 ---
475 ---
476 ---
477 ---
478 ---
479 ---
480 ---
481 ---
482 ---
483 ---
484 ---
485 ---
486 ---
487 ---
488 ---
489 ---
490 ---
491 ---
492 ---
493 ---
494 ---
495 ---
496 ---
497 ---
498 ---
499 ---
500 ---
501 ---
502 ---
503 ---
504 ---
505 ---
506 ---
507 ---
508 ---
509 ---
510 ---
511 ---
512 ---
513 ---
514 ---
515 ---
516
517 function aux(s)
518   local t = {}
519   for i = 1, #s do
520     for j = 1, #t do
521       t[#t+1] = {s[i],table.unpack(t[j])} end end
522   return t end
523
524 function sort(aux(s), function(a,b) return #a < #b end) end
525
526 function sort(t,f) table.sort(t,f); return t end
527
528 function upx(a,b) return a.x < b.x end
529 function upl(a,b) return a[1] < b[1] end
530 function downl(a,b) return a[1] > b[1] end
531
532 function slots(t, u)
533   local function public(k) return tostring(k):sub(1,1) ~= "." end
534   u={};for k,v in pairs(t) do if public(k) then u[l+#u]=k end end
535   return sort(u) end
536
537 function a(a,lo,hi)
538   lo,hi = lo or 1, hi or #a; return a[ (lo+(hi-lo)*math.random())//1 ] end
539
540 function many(a,n,lo,hi, u)
541   u={}; for j=1,n do push(u,any(a,lo,hi)) end; return u end
542
543 function slice(a,lo,hi, u)
544   u,lo,hi = {},lo or 1,hi or #a; for j=lo,hi do u[l+#u]=a[j] end; return u end
545
546 ---
547 ---
548 ---
549 ---
550 ---
551 ---
552 ---
553 ---
554 ---
555 ---
556 ---
557 ---
558 ---
559 ---
560 ---
561 ---
562 ---
563 ---
564 ---
565 ---
566 ---
567 ---
568 ---
569 ---
570 ---
571 ---
572 ---
573 ---
574 ---
575 ---
576 ---
577 ---
578 ---
579 ---
580 ---
581 ---
582
583 function words(s,sep, t)

```

```

516 sep="([^\n .. (sep or ",") .. "]+)"
517 t={}; for y in s:gmatch(sep) do t[1+#t] = y end; return t end
518
519 function things(s) return map(words(s), thing) end
520
521 function thing(x)
522 x = x:match("^%s*(~)%s*$"
523 if x=="true" then return true elseif x=="false" then return false end
524 return tonumber(x) or x end
525
526 function items(src,f)
527 local function file()
528 src,f = io.input(src),f or things
529 return function() x=io.read();if x then return f(x) else io.close(src) end e
530 nd end
531 local function tbl( x)
532 x,f = 0, f or function(z) return z end
533 return function() if x< #src then x=x+1; return f(src[x]) end end end
534
535 --- things 2 string
536
537 ---
538
539 fmt = string.format
540
541 function oo(t) print(o(t)) end
542
543 function o(t, seen, u)
544 if type(t)~="table" then return tostring(t) end
545 seen = seen or {}
546 if seen[t] then return "..." end
547 seen[t] = t
548 local function show1(x) return o(x, seen) end
549 local function show2(k) return fmt("%.5s",k, o(t[k],seen)) end
550 u = #t>0 and map(t,show1) or map(slots(t),show2)
551 return (t.s or "").."{"..table.concat(u, " ").."}" end
552
553 function rnds(t,f) return map(t, function(x) return rnd(x,f) end) end
554 function rnd(x,f)
555 return fmt(type(x)=="number" and (x~x//1 and f or "%.5f") or "%s",x) end
556
557 --- cli
558 ---
559
560 function cli(help)
561 local d,used = {},{}
562 help:gsub("(~)[^%s+)][%s]+(~)[^%s+)]^[%s+)][%s+)]",
563 function(long,key,short,x)
564 assert(not used[short], "repeated short flag ["..short.."]")
565 used[short]=short
566 for n,flag in ipairs(arg) do
567 if flag==short or flag==long then
568 x = x=="false" and true or x=="true" and "false" or arg[n+1] end end
569 d[key] = x==true and true or thing(x) end)
570 if d.help then os.exit(print(help)) end
571 return d end
572

```

```

572 -----
573 --- DEMOS
574
575 ---
576
577 local eg={}
578 function eg.copy( t,u)
579 t={a={b={c=10},d={e=200}}, f=300}
580 u= copy(t)
581 t.a.b.c= 20
582 print(u.a.b.c)
583 oo(t)
584 oo(u)
585 end
586
587 function eg.create()
588 oo(create{"Name","Age","gender","Weight-".y[1]} end
589
590 function eg.clone( i,t,best,rest)
591 i={rows={},cols=nil}
592 the.file = "../etc/data/auto93.csv"
593 i = xplain(the.file) end
594
595 function eg.collect()
596 local function aux(x,y) return x*y end
597 oo(collect({10,20,30},aux)) end
598
599 function eg.ent()
600 ok(ish(ent{a=9,b=7}, .98886), "entropy") end
601
602 function eg.items()
603 for x in items{10,20,30} do print(x) end
604 local n=0
605 for x in items(the.file) do n=n+1; if n<=5 then oo(x) end end end
606
607 function eg.powerset()
608 for _,x in pairs(powerset{10,20,30,40,50}) do oo(x) end end
609
610 local function qq(i,q)
611 print(q[1], fmt("%.5s=%-8s best= %s/%s rest= %s/%s",i.cols[q[2]].name, q[3],q[4],q[5]
612 ],q[6],q[7])) end
613
614 function eg.nb1()
615 local i = nb1(the.file);
616 local acc, out = score(i); print(acc); map(out,function(q) qq(i,q) end) end
617
618 function eg.nb2()
619 the.file = "../etc/data/diabetes.csv"
620 the.goal = "positive"
621 local i = nb2(the.file);
622 abcd(i.log,true)
623 end
624
625 function eg.nb2a()
626 the.file = "../etc/data/diabetes.csv"
627 the.goal = "positive"
628 for _,bins in pairs{2,5,9} do
629 print(bins)
630 the.bins = bins
631 local i = nb2(the.file);
632 abcd(i.log,true)
633 --local acc, out = score(i); print(acc)
634 --map(out,function(q) qq(i,q) end) end end
635
636 function eg.bins( t)
637 local t,n = {},30
638 for j=1,n do push(t, {x=j, y=j<.6*n and 1 or j<.8*n and 2 or 3}) end
639 map(bins(t,20),oo)
640 end
641
642 function eg.many( t)
643 t={};for j = 1,1000 do t[#t+1] = j end
644 print(900,"+", o(many(t,10,900)))
645 print(1,100,o(many(t,10,1,100)))
646 print(300,700, o(many(t,10,300,700))) end
647
648 function eg.nb3()
649 the.file = "../etc/data/diabetes.csv"
650 the.goal = "positive"
651 the.bins = 16
652 local i = nb3(the.file);
653 abcd(i.log,true)
654 local acc, out = score(i); map(out,function(q) qq(i,q) end)
655 end
656
657
658
659

```

```

659 -----
660 ---
661 --- START
662 ---
663 ---
664 fails = 0
665 local defaults=cli(help)
666 local todos = defaults.todo == "all" and slots(eg) or {defaults.todo}
667 for _,todo in pairs(todos) do
668     the = copy(defaults)
669     math.randomseed(the.seed or 10019)
670     if eg[todo] then eg[todo]() end end
671
672 roques()
673 os.exit(fails)
674
675 ---
676 ---
677 ---
678 ---
679 ---
680 ---
681 ---
682 ---
683 ---
684 ---
685
686 -- nb1 and nb2 has "?"
687 -- nb3 needsa new train.

```