1. Code Output
- Cross-Validation and ZeroR

```
================================================================================
        outlook,    -$humidity,         windy,         =play,  # notes
#         sunny,         90.00,          TRUE,            no,  # expected
#          0.67,          0.00,          0.67,          1.00,  # certainty
          rainy,             ?,          TRUE,            no,  #
          sunny,            90,          TRUE,            no,  #
          sunny,             ?,         FALSE,            no,  #

        outlook,    -$humidity,         windy,         =play,  # notes
#      overcast,         81.50,         FALSE,           yes,  # expected
#          0.50,          8.70,          0.75,          1.00,  # certainty
       overcast,            90,          TRUE,           yes,  #
          sunny,            70,         FALSE,           yes,  #
       overcast,            86,         FALSE,           yes,  #
          rainy,            80,         FALSE,           yes,  #
================================================================================
        outlook,    -$humidity,         windy,         =play,  # notes
#         rainy,         90.00,          TRUE,            no,  # expected
#          0.50,          0.00,          0.50,          1.00,  # certainty
          rainy,            90,          TRUE,            no,  #
          sunny,            90,         FALSE,            no,  #

        outlook,    -$humidity,         windy,         =play,  # notes
#      overcast,         77.20,         FALSE,           yes,  # expected
#          0.40,         11.90,          0.60,          1.00,  # certainty
       overcast,            75,         FALSE,           yes,  #
          sunny,            70,          TRUE,           yes,  #
          rainy,            80,         FALSE,           yes,  #
       overcast,            65,          TRUE,           yes,  #
          rainy,            96,         FALSE,           yes,  #
================================================================================

        outlook,    -$humidity,         windy,         =play,  # notes
#         sunny,         75.20,         FALSE,           yes,  # expected
#          0.40,         12.15,          0.60,          1.00,  # certainty
          sunny,            70,         FALSE,           yes,  #
          rainy,            96,         FALSE,           yes,  #
       overcast,            75,         FALSE,           yes,  #
          sunny,            70,          TRUE,           yes,  #
       overcast,            65,          TRUE,           yes,  #

        outlook,    -$humidity,         windy,         =play,  # notes
#         rainy,         90.00,          TRUE,            no,  # expected
#          1.00,          0.00,          1.00,          1.00,  # certainty
          rainy,            90,          TRUE,            no,  #
          rainy,             ?,          TRUE,            no,  #
================================================================================
        outlook,    -$humidity,         windy,         =play,  # notes
#         sunny,         90.00,         FALSE,            no,  # expected
#          1.00,          0.00,          0.67,          1.00,  # certainty
          sunny,            90,          TRUE,            no,  #
          sunny,             ?,         FALSE,            no,  #
          sunny,            90,         FALSE,            no,  #

        outlook,    -$humidity,         windy,         =play,  # notes
#         rainy,         84.00,         FALSE,           yes,  # expected
#          0.50,          4.90,          0.75,          1.00,  # certainty
       overcast,            86,         FALSE,           yes,  #
          rainy,            80,         FALSE,           yes,  #
          rainy,            80,         FALSE,           yes,  #
       overcast,            90,          TRUE,           yes,  #
================================================================================
```

2. Zeror Results
- Implement zeror and call it in a cross-val.
- Zeror's accuracies for 'weather1.csv' dataset: 71.43, 57.14, 57.14, 71.43
- Zeror's accuracies for 'soybean.csv' dataset: 11.73, 12.02, 12.90, 11.44

3. Illustration
- xvaltables is a nested dictionary structure that stores all the results from cross-validation process

```
- xvaltables = {'i':{'train':{'0':table0, 'klassname1':table1, 'klassname2':table2,...,'names':list of classnames in table0},
                     'test' :{'0':table0, 'klassname1':table1, 'klassname2':table2,...,'names':list of classnames in table0}},
                ...
                }
```

        * most outlier key i: value from 1 to x*b are the separated groups of training and testing dataset
        * second outlier key: 'train' or 'test' indicate the datasets under the same group i are used for training or testing
        * '0':table1 contain all the data designed to group i's training or testing dataset
        * 'klassnamei':tablei contain all the data in table0 with class value equals to klassnamei
        * 'names': a list that include all the classnames in table0

4. Source Codes

```
================================================================================
File <tablestr.py>

import lib
class Table:
    def __init__(self):
        self.data = []     #data[[col1,...],[col2,...]]
        self.name = []     #name of i-th column
        self.order = []    #order of the col
        self.nump = []     #is i-th column numeric?
        self.wordp = []    #is i-th column non-numeric?
        self.indep = []    #list of indep columns
        self.dep = []      #list of dep columns
        self.less = []     #numeric goal to be minimized
        self.more = []     #numeric goal to be maximized
        self.klass = []    #non-numeric goal
        self.term = []     #non-numeric non-goal
        self.num = []      #numeric non-goal
        # for all cols
        self.n = []        #count of things in this col
        # for wordp columns:
        self.count = []    #count of each word
        self.mode = []     #most common word
        self.most = []     #count of most common word
        # for nump columns:
        self.hi = []       #upper bound
        self.lo = []       #lower bound
        self.mu = []       #mean
        self.m2 = []       #sum of all nums
        self.sd = []       #standard deviation# -*- coding: utf-8 -*-
        # table printing format
        self.CONVFMT = '%06d'

def centroid(table):
    "update the mode and most values for wordp type cols or update the mean and sd values for nump cols"
    rows = [[],[]]
    for c in range(len(table.name)):
        s = table.mode[table.wordp.index(c)] if c in table.wordp else table.CONVFMT%table.mu[table.nump.index(c)]
        rows[0].append(str(s))
        if table.n[c] == '0':     # if all the data in this col is "?"
            s = 0.0
        else:
            s = float(table.most[table.wordp.index(c)])/table.n[c] if c in table.wordp else table.sd[table.nump.index(c)]
        rows[1].append(str(table.CONVFMT%s))
    return rows

def tableprint(table, stats=''):
    "print table on the console"
    print ' '
    if stats != '': table.CONVFMT = stats
    print(' ' + lib.rowprint(table.name)+ '  # notes'.ljust(10))
    print('#' + lib.rowprint(centroid(table)[0]) + '  # expected'.ljust(10))
    print('#' + lib.rowprint(centroid(table)[1]) + '  # certainty'.ljust(10))
    for j in range(len(table.data[0])):
        line = []
        for i in range(len(table.data)):
            line.append(table.data[i][j])
        print(’ ’ + lib.rowprint(line)+ ’  #’.ljust(10))

def tableprint_txt(table, f, stats=’’):
    "print table on the indicated txt file with table name"
    f.write(’\n’)
    if stats != ’’: table.CONVFMT = stats
    f.write(’ ’ + lib.rowprint(table.name)+ ’  # notes’.ljust(10) + ’\n’)
    f.write(’#’ + lib.rowprint(centroid(table)[0]) + ’  # expected’.ljust(10) +
’\n’)
    f.write(’#’ + lib.rowprint(centroid(table)[1]) + ’  # certainty’.ljust(10) +
 ’\n’)
    for j in range(len(table.data[0])):
        line = []
        for i in range(len(table.data)):
            line.append(table.data[i][j])
        f.write(’ ’ + lib.rowprint(line)+ ’  #’.ljust(10) + ’\n’)
================================================================================
File <reader.py>

import re
import tablestr
def readcsv(filename, table):
    "read in data from csv and create a table"
    FS = ’,’                     #define field separator
    f = open(filename)
    seen  = 0
    while True:
        str = line(f)
        if str == −1:
            if seen == 0: print("WARNING: empty or missing file")
            return −1
        a = str.split(FS)        #compute the number of attributes in table
        if len(a) > 1:
            if seen: addRow(a, table)
            else: makeTable(a, table)
            seen += 1

def line(f):
    "get one line data (without comments and whitespace)"
    str = f.readline()
    if not str: return −1            #readline finds nothing, output error
    else:
        str = "".join(str.split())    #kill whitespace
        str = re.sub(r’#.*’,’’,str)   #kill comments
        if len(str) >= 1 and str[−1] == ’,’: return str + line(f)
        else: return str

def makeTable(a, table):
    "read table titles and set all corresponding parameters"
    c = 0
    for ite in range(len(a)):
        if a[ite][0] == ’?’: continue  #the col with ’?’ is ignored
        table.order.append(ite)
        x = a[ite]
        table.name.append(x)
        isNum = 1
        if x.find(’=’) != −1:
            table.dep.append(c)
            table.klass.append(c)
            isNum = 0
        elif x.find(’+’) != −1:
            table.dep.append(c)
            table.more.append(c)
        elif x.find(’−’) != −1:
            table.dep.append(c)
            table.less.append(c)
        elif x.find(’$’) != −1:
            table.indep.append(c)
```

```
            table.num.append(c)
        else:
            table.indep.append(c)
            table.term.append(c)
            isNum = 0
        table.n.append(’0’)
        if isNum:
            table.nump.append(c)
            table.hi.append(−1*10**32)
            table.lo.append(10**32)
            table.mu.append(0)
            table.m2.append(0)
            table.sd.append(0)
        else:
            table.wordp.append(c)
            table.most.append(0)
            table.count.append({})
            table.mode.append(’’)
        c += 1
    for i in range(c): table.data.append([])

def addRow(a, table):
    "add a row of data to the table"
    for c in range(len(table.name)):
        f = table.order[c]
        x = a[f]
        table.data[c].append(x)
        if x.find(’?’) == −1:
            table.n[c] = int(table.n[c]) + 1
            if c in table.wordp:
                k = table.wordp.index(c)
                if table.count[k].has_key(x): table.count[k][x] += 1
                else: table.count[k][x] = 1
                new = table.count[k][x]
                if new > table.most[k]:
                    table.mode[k] = x
                    table.most[k] = new
            else:
                k = table.nump.index(c)
                if float(x) > float(table.hi[k]): table.hi[k] = x
                if float(x) < float(table.lo[k]): table.lo[k] = x
                delta = float(x) − table.mu[k]
                table.mu[k] += delta/table.n[c]
                table.m2[k] += delta*(float(x) − table.mu[k])
                if table.n[c] > 1:
                    table.sd[k] = (table.m2[k]/(table.n[c] − 1))**0.5
            c += 1

def klasses(table):
    "generate a set of tables based on different classes"
    if len(table.klass) == 0:
        print "No labeled classes in the given data set"
        return −1
    # assume there is only one class feature in the data set
    data = table.data[table.klass[0]]
    classnames = []
    for s in data:
        if s not in classnames:
            classnames.append(s)
    tables = klass1(table, classnames, data)
    tables[’0’] = table
    tables[’names’] = classnames
    return tables

def klass1(table, classnames, data):
    tables = {}
    for s in classnames:
        tables[s] = tablestr.Table()
        makeTable(table.name, tables[s])
        for i in range(len(data)):
            if s == data[i]:
                a = []
                for j in range(len(table.order)):
                    a.append(table.data[j][i])
                addRow(a, tables[s])
    return tables
================================================================================
File <lib.py>

def indexes(data):
    rows = []   #get the indexes for the data
    for i in range(len(data)):
        rows.append(i)
    return rows

def rowprint(a):
    max = len(a)
    line = ’’
    for j in range(max):
        line += (a[j] + ’,’).rjust(15)
    return line

def maybeInt(x):
    return int(x) if x % 1 == 0.0 else float(x)
================================================================================
File <xval.py>

import lib
import tablestr
import reader
import random

def xvals(tables, x, b):
    k = tables[’0’].order.index(tables[’0’].klass[0])
    rows = lib.indexes(tables[’0’].data[k])
    s = int(len(rows)/b)
    xvaltables = {}
    for i in range(x):      # x times
        random.shuffle(rows)
        for b1 in range(b): # b bins
            obj = xval(b1*s, (b1+1)*s, rows, tables)
            xvaltables[i*x+b1+1] = obj
    return xvaltables


def xval(start, stop, rows, tables):
    testT = tablestr.Table()
    trainT = tablestr.Table()
    reader.makeTable(tables[’0’].name, testT)
    reader.makeTable(tables[’0’].name, trainT)
    for r in range(len(rows)):
        d = rows[r]
        a = []
        for j in range(len(tables[’0’].order)):
            a.append(tables[’0’].data[j][d])
        if r >= start and r < stop: #belonging to testing data set
            reader.addRow(a, testT)
        else:
            reader.addRow(a, trainT)
    testT = reader.klasses(testT)
    trainT = reader.klasses(trainT)
    tables = {}
    tables[’train’] = trainT
    tables[’test’] = testT
    return tables
================================================================================
File <zeror.py>

def zeror(testT, trainT, hypotheses):
    k = testT[’0’].klass[0]
    most = 0
    for h in hypotheses:
        these = len(trainT[h].data[k]) if h in trainT[’names’] else 0
        if these > most:
            most = these
            got = h
```

```
    #print "#got", got
    acc = len(testT[got].data[k]) if got in testT[’names’] else 0
    num = 0
    for h in hypotheses: num += len(testT[h].data[k]) if h in testT[’names’] els
e 0
    return got,str(’%4.2f’%(100*float(acc)/num))
================================================================================
File <main.py>

import reader
import tablestr
import zeror
import xval
if __name__ == "__main__":
    filename = ’data/weather1.csv’
    table = tablestr.Table()             #create raw data structure
    reader.readcsv(filename,table )      #read the .csv data set
    f = ’%4.2f’                          #set the formatting for the output
    filename = ’output/table_xval_zeror.txt’
    out = file(filename, ’w’)
    tables = reader.klasses(table)
    tablestr.tableprint_txt(tables[’0’], out, f)
    b = x = 2
    xvaltables = xval.xvals(tables, x, b) #generate the cross validation tables
    for s in range(x*b):
       s += 1
       out.write(’=’*80+’\n’)
       out.write(’Group:’+ str(s) +’\n’)
       out.write(’Training Set \n’)
       for h in xvaltables[s][’train’][’names’]:
           tablestr.tableprint_txt(xvaltables[s][’train’][h], out, f)
       out.write(’Testing Set \n’)
       for h in xvaltables[s][’test’][’names’]:
           tablestr.tableprint_txt(xvaltables[s][’test’][h], out, f)
       got, acc = zeror.zeror(xvaltables[s][’test’], xvaltables[s][’train’], tab
les[’names’])
       out.write(’#Got: ’ + got +’\n’)
       out.write(’#Accuracy: ’ + acc+’\n’)
    out.close()

```