Illustration of how to create training and test data sets from the iris data set

Anca Ralescu

September 25, 2014

Please note that the code shown below is only one way that we can create the train and test data sets. Moreover, this is not the BEST code that one can write in Matlab, but I wrote it this way for those who may not know matlab very well. Usually, in Matlab we avoid loops and try to implement them as matrix operations.

Contents

This script shows how to generate train and test data sets

```
clear all;
% load the full data set
data = load('iris.txt');

% check its size: use Matlab function size(). It returns the number of rows
% and columns of a matrix
size(data)

% when ; is omitted the result of the execution of a function / statement
% is returned; for this data set the result shoud be 150 x 5.  The 5th
% column contains the class labels

% length extracts the number of lements in a vector
lengthdata = length(data(:,1));

% set the index variable to take values 1, 2, ..., num of data (here it
% will be 150)
index=1:lengthdata;

% rindex is a random permutation of the values 1, 2, ..., index(end) which
```

```
% is the length of data; here it will be a permutation of 1,2,...,150
rindex=randperm(index(end));

% extract the number of classes: unique is a matlab function which returns
% the list of distinct values of the list which is its argument
numclasses = length(unique(data(:,end))); %number of distinct labels

% let p denote the percentage of the data which will be used for training;
% it is a number in [0,1); for example, if p=0.1, p*150 = 0.1*150 = 15 then
% the size of the training set is 15. see below:

p=0.1;

% let split be defined as round(p*lengthdata)
split = round(p*lengthdata)


% below we generate subsets of the full data set, for train and test
%according to split as follows

 p=0.1:0.1:0.9;   % NOTE: you can take other values for p as well
 N=length(p);
 for i=1:N
```

Split class training data and test data

```
split = round(p(i)*lengthdata);

% train is a cell structure
%train{i} = nearest(data(rindex(1:split), :)); %use nearest(x) if integer values are desirec
%nearest(x)= integer closest to x

% OBTAIN THE TRAINING SET
train{i} = data(rindex(1:split), :);


% OBTAIN THE TEST SET
% Uncomment only one of the two statements below
testdata{i} = data(rindex(split+1:lengthdata),:); % remaining data; generalization
%testdata = train; % for modeling power; can test overfitting

%LENGTH OF THE TEST SET
ltest(i) = length(testdata{i}(:,1));


% Extract data without labels from the training set
```

```
train_without_label{i}=train{i}(:,1:end-1);

% Extract the labels in the training set
train_labels{i}=train{i}(:,end);

% Extract data without labels from the TEST set
test_without_label{i}=testdata{i}(:,1:end-1);

% Extract the labels in the TEST set
truelabels{i} = testdata{i}(:,end); % true lables of the test data



% Split training data into classes along the attributes selected;
% Find the size of each class

for ic=1:numclasses,
    Index{ic} = find(train{i}(:,end)==ic);
    class{i, ic} = train_without_label{i}(Index{ic}, :);
    lclass(ic)=length(Index{ic});
end

classprob = lclass/split; % not used here now

 end

 % Use who to see the variables created
who

% Now see what these variables are
class
train
testdata

ans =

   150     5


split =

    15


Your variables are:
```

```
Index                 i                     numclasses            train
N                     ic                    p                     train_labels
ans                   index                 rindex                train_without_label
class                 lclass                split                 truelabels
classprob             lengthdata            test_without_label
data                  ltest                 testdata
```

```
class =

    [ 8x4 double]    [ 4x4 double]    [ 3x4 double]
    [11x4 double]    [12x4 double]    [ 7x4 double]
    [16x4 double]    [18x4 double]    [11x4 double]
    [24x4 double]    [22x4 double]    [14x4 double]
    [26x4 double]    [28x4 double]    [21x4 double]
    [28x4 double]    [33x4 double]    [29x4 double]
    [33x4 double]    [39x4 double]    [33x4 double]
    [39x4 double]    [42x4 double]    [39x4 double]
    [44x4 double]    [48x4 double]    [43x4 double]
```

```
train =

  Columns 1 through 5

    [15x5 double]    [30x5 double]    [45x5 double]    [60x5 double]    [75x5 double]

  Columns 6 through 9

    [90x5 double]    [105x5 double]    [120x5 double]    [135x5 double]
```

```
testdata =

  Columns 1 through 5

    [135x5 double]    [120x5 double]    [105x5 double]    [90x5 double]    [75x5 double]

  Columns 6 through 9

    [60x5 double]    [45x5 double]    [30x5 double]    [15x5 double]
```