Commonly Used Datasets for ML

J.-S. Roger Jang (張智星)
jang@mirlab.org
http://mirlab.org/jang
MIR Lab, CSIE Dept.
National Taiwan University

Datasets

- Numerous datasets for testing ML algorithms
  - Kaggle
  - UCI Machine Learning Repository
  - Image net
  - MNIST handwritten digit database
  - Labeled Faces in the Wild
  - Many many more…
- Looking for a specific dataset?
  - Google search engine
  - Google dataset search

Try "license plate dataset" in Google!

UCI Dataset: Iris

- Source
  - R.A. Fisher, 1936
- Goal
  - Predict the types of iris in Hawaii
- Dataset specs
  - 150 instances, 3 classes
  - 4 attributes (features)
    - sepal length
    - sepal width
    - petal length
    - petal width

UCI Dataset: Wine

- Source
  - Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.
- Goal
  - Using 13 chemical constituents to determine the origin of wines
- Dataset specs
  - 178 instances, 3 classes, 13 attributes

UCI Dataset: Abalone

- Source
  - Dept. of Primary Industry and FIsheries, Tasmania, Australia
- Goal
  - Predict the age of abalone (鮑魚)
- Dataset specs
  - 4177 instances, 29 classes
  - 8 attributes (features): sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight
  - 1 output: rings (+1.5 gives the age in years)

UCI Dataset: Mushroom Classification

- Source
  - Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981)
- Goal
  - To determine a mushroom is poisonous or edible
- Dataset specs
  - 8124 instances, 2 classes, 22 attributes

UCI Dataset: Liver Disorder

- Source
  - BUPA Medical Research Ltd.
- Goal
  - Use variables from blood tests and alcohol consumption to see if liver disorder exists
- Dataset specs
  - 345 instances, 2 classes, 6 attributes (the first five are results from blood tests, the last one is alcohol consumption per day)

UCI Dataset: Credit Screening

- Source
  - Chiharu Sano, csano@bonnie.ICS.UCI.EDU
- Goal
  - Determine people who are granted credit
- Dataset specs
  - 125 instances, 2 classes, 15 attributes

UCI Dataset: House Price Prediction

- Source
  - CMU StatLib Library
- Goal
  - Predict house price near Boston
- Dataset specs
  - 506 instances, 13 attributes

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centres
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's

MNIST Digit Dataset (1/2)

- Source

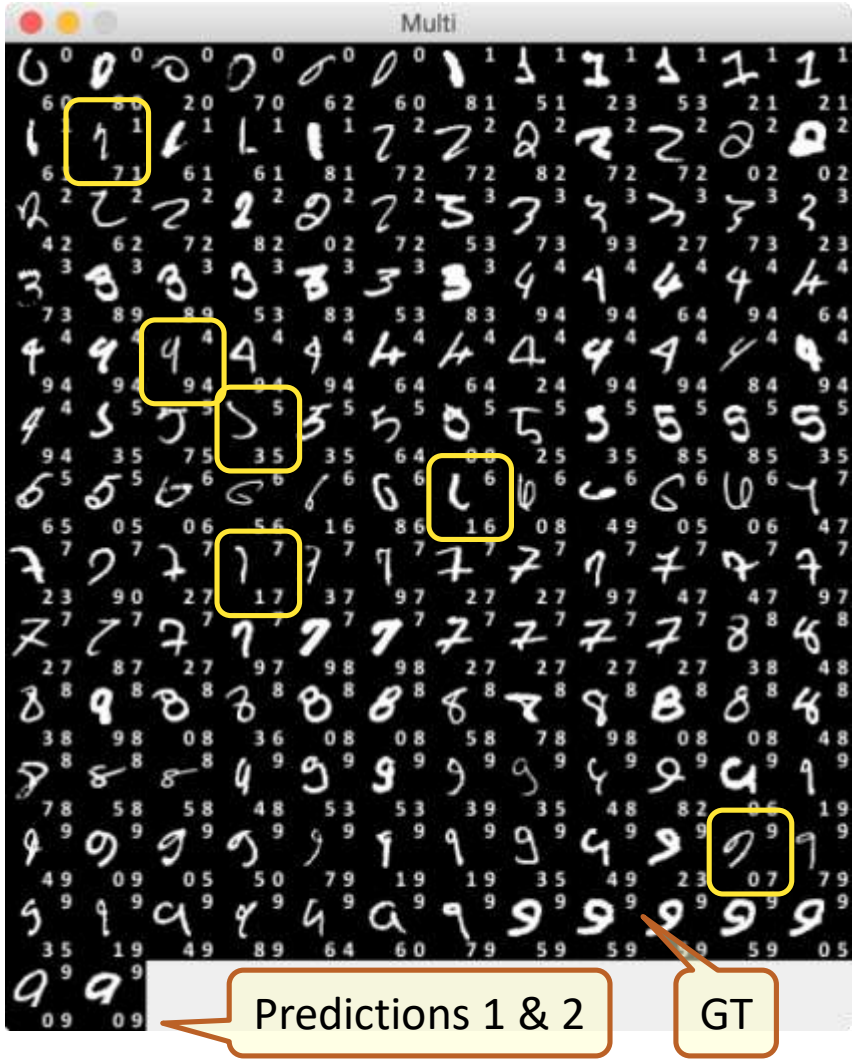
  Quiz: Full name of NIST?

  - NIST's Special Database 3 (collected among Census Bureau employees) and Special Database 1 (collected among high-school students)
- Goal
  - Recognize isolated hand-written digits of 0-9
- Dataset specs
  - 70000 instances
    - 60000 for training (30000 from SD-3 and 30000 from SD-1) of about 250 writers
    - 10000 for test (5000 from SD-3 and 5000 from SD-1)

    Disjoint writers!
  - Normalized to 28x28 gray-scale image, centered by gravity

MNIST Digit Dataset (2/2)

- Links
  - Data source
  - Wikipedia
- Examples



How to Acquire/Visualize the Datasets?

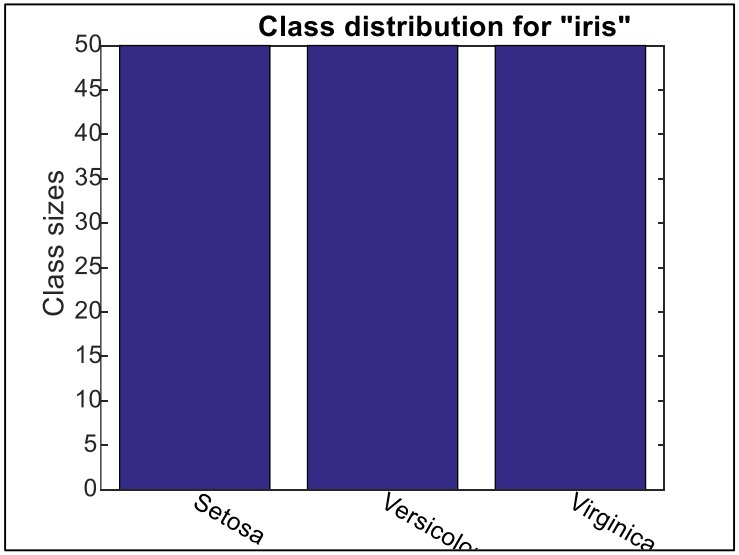
- Acquire the datasets
  - prData.m for acquiring PR data
  - dcData.m for acquiring DC data
- Visualize the datasets
  - Please refer to Chapter 2 of DCPR tutorial
- Example:

You need to download Machine Learning Toolbox to try these commands.

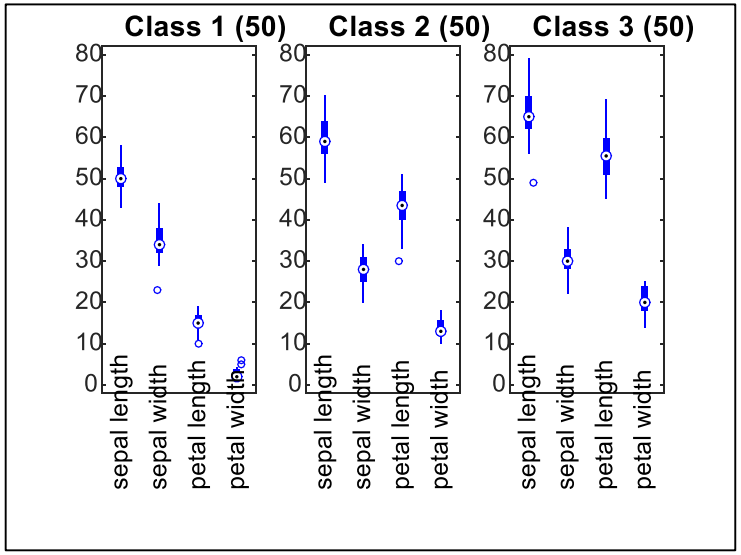
```
>> ds=prData('iris')

ds =
     dataName: 'iris'
    inputName: {'sepal length'  'sepal width'  'petal length'  'petal width'}
   outputName: {'Setosa'  'Versicolour'  'Virginica'}
        input: [4x150 double]
       output: [1x150 double]
```

Iris Dataset Visualization (1/2)

```
ds=prData('iris');
classSize=dsClassSize(DS, 1);
```

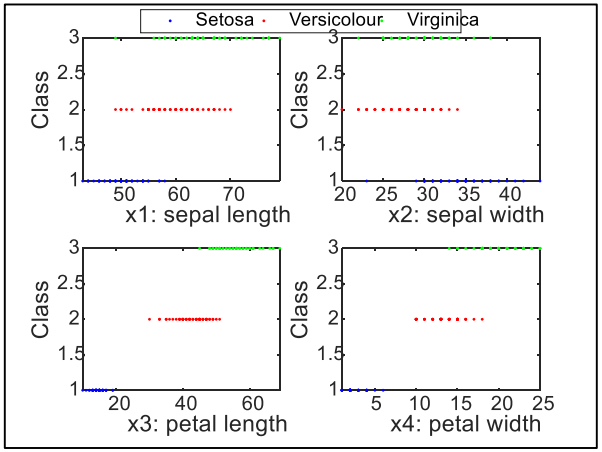



```
ds=prData('iris');
dsDistPlot(ds);
```

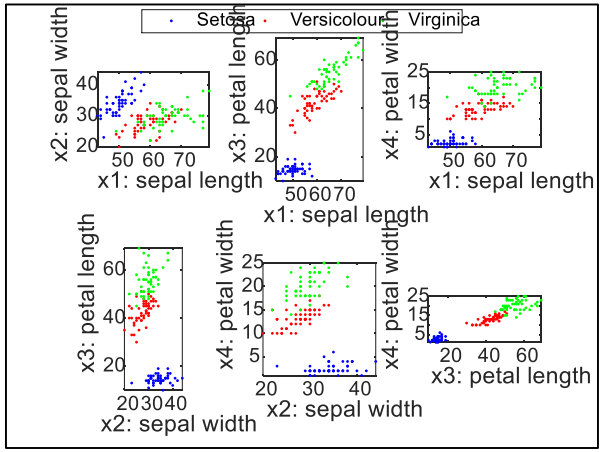


Iris Dataset Visualization (2/2)

```
ds = prData('iris');
dsProjPlot1(ds);
```




```
ds = prData('iris');
dsProjPlot2(ds);
```




```
ds = prData('iris');
dsProjPlot3(ds);
```