

Matlab 7: Commonly Used Datasets for Pattern Recognition



Cheng-Hsin Hsu

National Tsing Hua University

Department of Computer Science

Slides are based on the materials from Prof. Roger Jang

Datasets

- There are numerous datasets for testing machine learning algorithms:
 - [Kaggle](#)
 - [UCI Machine Learning Repository](#)
 - [Image net](#)
 - [MNIST handwritten digit database](#)
 - [Labeled Faces in the Wild](#)
 - Some conferences have dataset tracks
 - Many many more can be found via Google...

Kaggle Dataset

The image shows two overlapping screenshots of Kaggle dataset pages. The top screenshot is for the 'Predict'em All' dataset, and the bottom screenshot is for the '2016 US Election' dataset.

Predict'em All

Predict where Pokemon appear in PokemonGo based on historical data

by SemionKorchevsky · last updated 2 months ago

39

Overview | Kernels | Discussion | Activity | Download (75 MB) | New Notebook | **New Script**

Kernels	Discussion
Pokemon WHOA - DrNDM run 2 months ago 6 votes	Dataset update & questions 22 days ago
Shiny App for Pokemon Go ... run a month ago 5 votes	Map visualization (tests) wit... a month ago
Some Exploratory Data Anal... run a month ago 5 votes	Shiny App for Pokemon Go ... a month ago

Recent Activity

EdirisPiril Created kernel Notebookc7121da5aa

2016 US Election

Explore data related to the 2016 US Election

by Ben Hamner · last updated 5 months ago

153

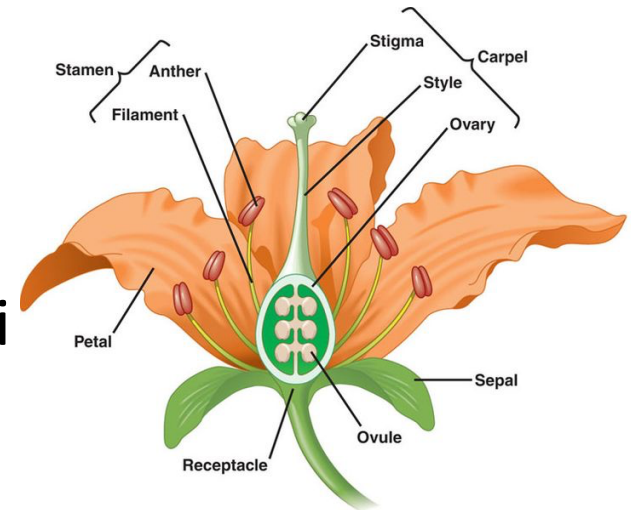
Overview | Kernels | Discussion | Activity | Download (17 MB) | New Notebook | **New Script**

Kernels	Discussion
Predictions in the Republica... run 5 months ago 75 votes	Is there a nationwide poll da... 10 days ago 4 replies
Lets Look at Correlations run 9 months ago 22 votes	Heatmap of votes fraction f... 17 days ago 3 replies
'Clinton: Champion of the Pri... run 6 months ago 20 votes	Predictions in the Republica... a month ago 22 replies

Top Contributors
Ben Hamner 1st
Alexandru Papiu 2nd
DerekElliott 3rd

UCI Dataset: Iris

- Source
 - R.A. Fisher, 1936
- Goal
 - Predict the types of iris in Hawaii
- Problem sizes
 - 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
- Problem size
 - 178 instances, 3 classes, 13 attributes



UCI Dataset: Abalone

- Source
 - Dept. of Primary Industry and Fisheries, Tasmania, Australia
- Goal
 - Predict the age of abalone
- Problem sizes
 - 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

- Source
 - Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981)
- Goal
 - To determine a mushroom is poisonous or edible
- Problem size
 - 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
- Problem size
 - 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
- Problem size
 - 125 instances, 2 classes, 15 attributes

UCI Dataset: House Price Prediction

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

How to Acquire/Visualize the Datasets?

- Acquire the datasets
 - prData.m for acquiring PR data
 - dcData.m for acquiring DC data
- Visualize the datasets, for example:

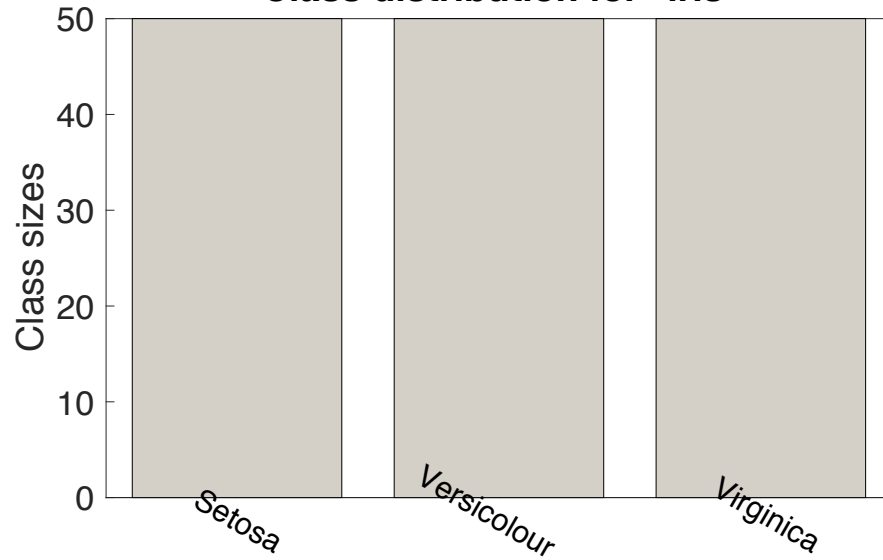
```
>> ds=prData('iris')  
  
ds =  
  dataName: 'iris'  
  inputName: {'sepal length' 'sepal width' 'petal length' 'petal width'}  
  outputName: {'Setosa' 'Versicolour' 'Virginica'}  
  input: [4x150 double]  
  output: [1x150 double]
```

Iris Data Visualization

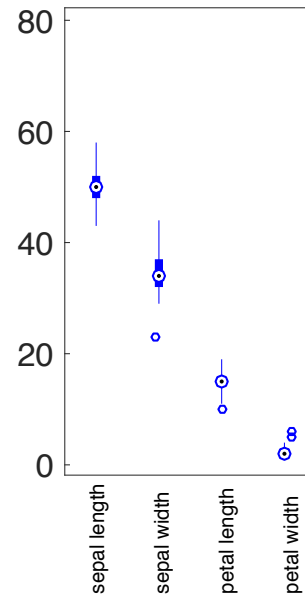
```
ds=prData('iris');  
classSize=dsClassSize(ds, 1);
```

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

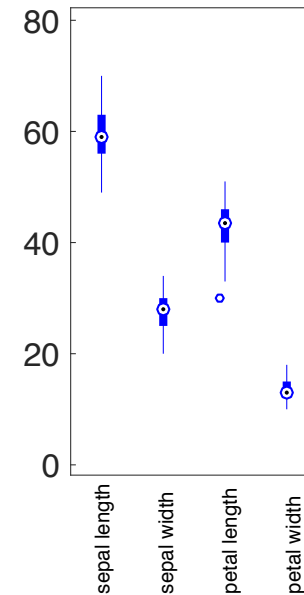
Class distribution for "iris"



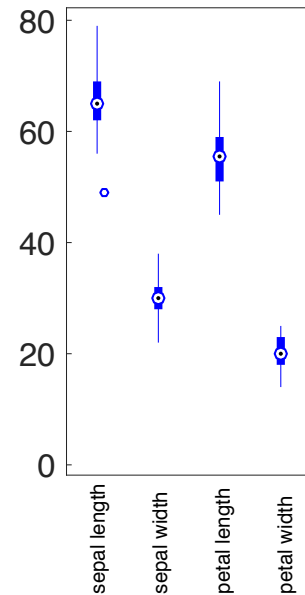
Class 1 (50)



Class 2 (50)



Class 3 (50)

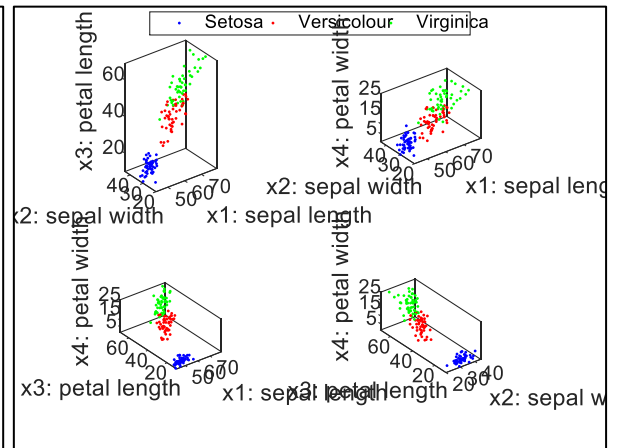
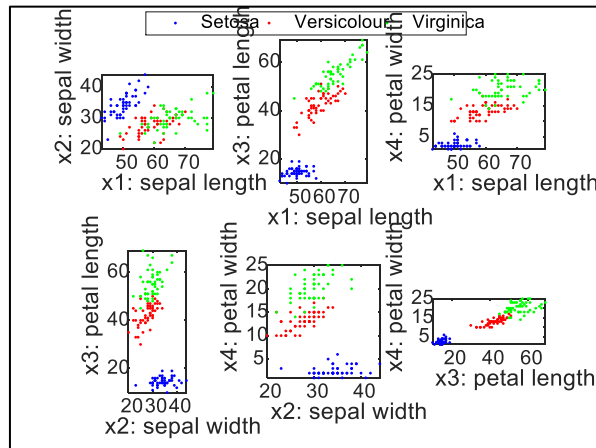
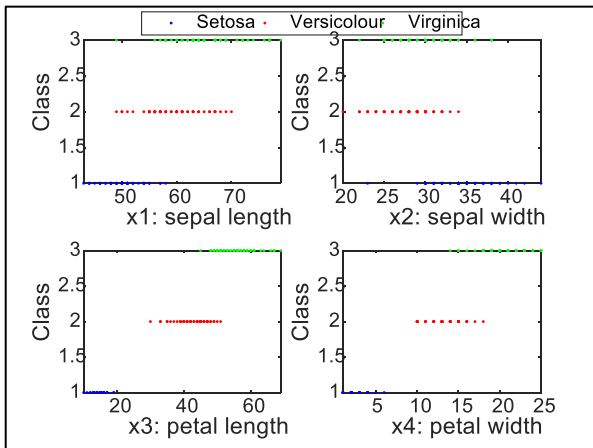


Iris Data Visualization (cont.)

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

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

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



List of Visualization Functions

- `classClassSize(DS)`: Compute the size of each class
- `dsProjPlot1(DS)`: Plot the classes w.r.t. the projected 1D features
- `dsProjPlot2(DS)`: Plot the classes w.r.t. the projected 2D features
- `dsProjPlot3(DS)`: Plot the classes w.r.t. the projected 3D features
- `dsFormatCheck(DS)`: Check the format of the dataset
- `dsNameAdd(DS)`: Add names to the inputs and outputs of a given dataset
- `dsRangePlot(DS)`: Plot the range of each input of a given dataset
- `dsDistPlot(DS)`: Plot the distributions of inputs over different classes of a given dataset
- `dsScatterPlot(DS)`: Scatter plot of a dataset in a 2D space
- `dsScatterPlot3(DS)`: Scatter plot of a dataset in a 3D space
- More details can be found at <https://goo.gl/77kb3v>

Matlab #6 Homework (M6)

1. (1%) Use K-means algorithm with $K=3$ to classify the Iris dataset with all four features (4 dimensions). Plot a figure to visualize the sample data and the classifications due to K-means and the ground-truth. Please submit two files: your .m file and an eps.

Questions?

