« Previous (/course/cs357-f15/flow-session/74262/1/)

下一页 » (/course/cs357-f15/flow-session/74262/3/)

跳转至最末 »

1 2 3 (/course/cs357-f15/flow-session/74262/0/) (/course/cs357-f15/flow-session/74262/1/)
4 5 (/course/cs357-f15/flow-session/74262/3/) (/course/cs357-f15/flow-session/74262/4/) (/course/cs357-f15/flow-session/74262/5/)
6

What Information Do We Have?

We have the following data to work with:

- There are 2 data files named breast-cancer-train.dat (/course/cs357-f15/file-version/59babe8be91b4665456aae2abfdc1e1e732643ec/media/least-squares-cancer/breast-cancer-train.dat) and breast-cancer-validate.dat (/course/cs357-f15/file-version/59babe8be91b4665456aae2abfdc1e1e732643ec/media/least-squares-cancer/breast-cancer-validate.dat) residing on the class server. There are many ways to load this text. One is by way of `numpy.loadtxt`. The second colum is a string with `M` and `B`, so make sure to handle this as a float (-1 or 1) using `converters`.

- There is a list of labels in labels.txt (/course/cs357-f15/file-version/59babe8be91b4665456aae2abfdc1e1e732643ec/media/least-squares-cancer/labels.txt) that contains names (type: string). This can be loaded with `numpy.genfromtxt`.

- Consider a second list called `subset_labels` that contains the names of 4 columns that you will use when creating a quadratic least squares representation:

```
subset_labels = ["radius (mean)", "perimeter (mean)","area (mean)", "symmetry (mean)"
]
```

A snapshot of the data set is shown below.

You will notice that the first column is called *patient ID*. For each patient there is an entry in the 'Malignant/Benign' column indicating whether their tumor was malignant or benign. The remaining 30 columns give characteristics of the tumor.

1. Load your data.

    - `tumor_data` should have 300 rows and 32 columns

    - `validate_data` should have 260 rows and 32 columns

    - `labels` should have 32 labels

    - Now take `labels` and make `labelsid`, a dictionary that maps the label name to the column number. This is easy with `enumerate`

2. Take a look at some the data using plot commands
    - Generate a histogram of one of the data columns. Check out `radius (mean)`.
    - Generate a plot of same data.
    - For both the histogram and plot, include titles, x-labels, and y-labels. What should they be?