Machine Learning web service using Python,Bottle and Scikit-Learn

Software as a service (SAAS) is a nice way to provide analytics capabilities to people who are not experts in machine learning and/or do not have time to build the necessary tools. Here, I implemented a simple web service utilizing the python based machine learning toolkit (scikit-learn) that applies simple dimensionality reduction algorithms (Principal Components Analysis and Linear Discriminant Analysis) to a dataset of user's choice and returns 2D visualizations of the data.

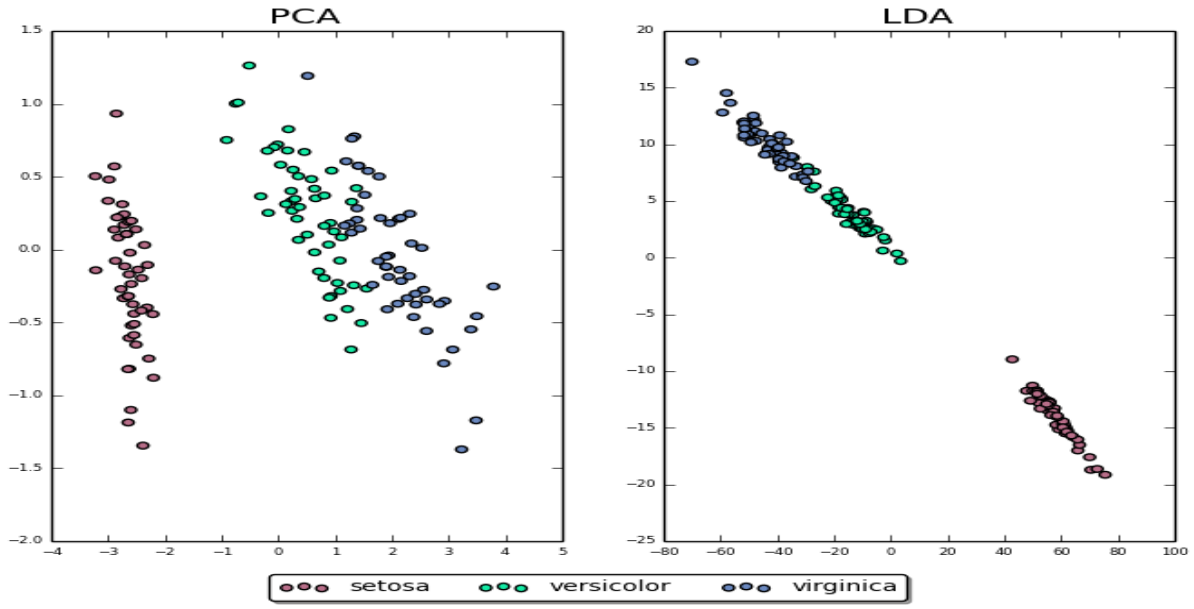
The implementation is totally python based and it uses the Bottle Web Framework. The service is located at:http://mindwriting.org:8073/

First, it will allow you to upload your dataset as a comma separated .csv file. One restriction of the application is that it needs the class labels to be provided as the last column of the dataset. You will also need to provide a header line with attribute names. An example dataset looks like this:

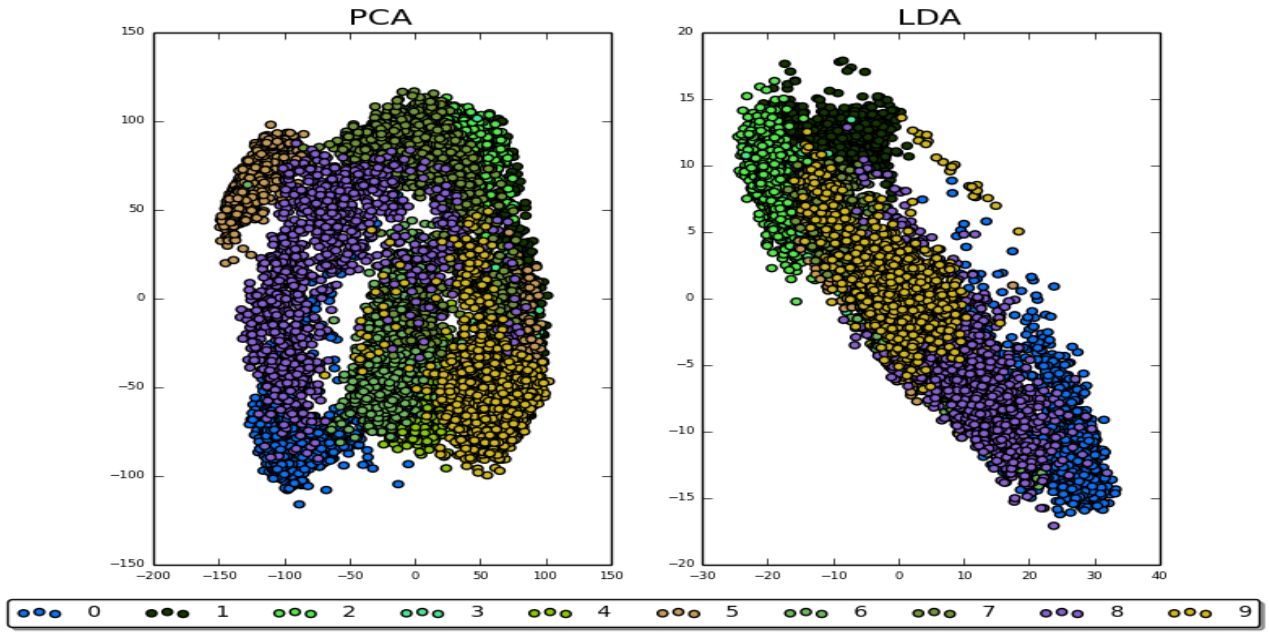
	A	B	C	D	E	F
1	FL	RW	CL	CW	BD	type
2	8.1	6.7	16.1	19	7	1
3	8.8	7.7	18.1	20.8	7.4	1
4	9.2	7.8	19	22.4	7.7	1
5	9.6	7.9	20.1	23.1	8.2	1
6	9.8	8	20.3	23	8.2	1
7	10.8	9	23	26.5	9.8	1
8	11.1	9.9	23.8	27.1	9.8	1
9	11.6	9.1	24.5	28.4	10.4	1
10	11.8	9.6	24.2	27.8	9.7	1
11	11.8	10.5	25.2	29.3	10.3	1
12	12.2	10.8	27.3	31.6	10.9	1

Then, you submit the dataset and get back the visualization. Here are some examples along with the datasets:
wine dataset (comma-delimited csv)

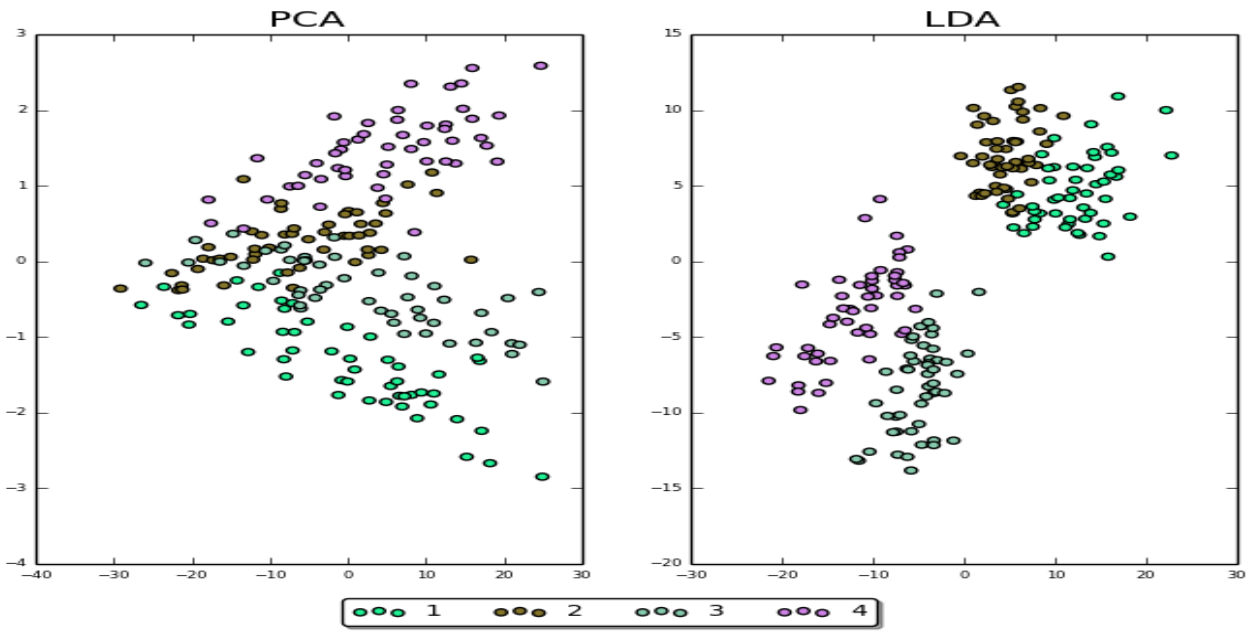
iris dataset (comma-delimited csv)



[digits dataset (comma-delimited csv)]



[Ripley's Leptograpsus crabs dataset (comma-delimited csv)]



Finally, here is how it is done:
A very simple form (upload.html) to upload the dataset and call the service.

This is what the service returns: just an image of the 2D visualizations embedded in html:

The main work is done by the plot() function (pca_lda_viz.py) that receives the uploaded .csv file, extracts the attributes and class variable, applies data transformations, creates 2D visualizations.

The code above creates random colors to represent each class. However, it sometimes does not generate distinct looking colors. I leave it as an exercise to create a set of N distinct colors.

The largest dataset I have tested is the madelon dataset download here:madelon_training set (500 attributes, 2000 rows) (comma-delimited csv) Response time is pretty good but the data transformations we get from these two simple methods are not interesting at all--this was a feature extraction challenge dataset after all!

The service has not been designed to process very large data files--see Mahout(java) for map-reduce implementation. Therefore, it is not ready for the big data frenzy. Nevertheless, I believe it is a good start that can be extended for big learning.

In order to run the example, just change the hostname and port number in pca_lda_viz.py and start the service as:

```
> python pca_lda_viz.py
```

If all goes fine, it should start without any error messages. All done!

Github repository: https://github.com/ilknuricke/bottle_scikit_learn_web_services