

Homework 1

Due, Tuesday, November 21, 2017

1. Use your own python PCA and linear regression modules to answer the following two questions.

(1) (30 points) The data in `linear_regression_test_data.csv` contains x , y , and y -theoretical. Perform PCA on x and y . Plot y vs x , y -theoretical vs x , and the PC1 axis in the same plot.

(2) (30 points) Perform linear regression on x and y with x being the independent variable and y being the dependent variable. Plot the regression line in the same plot as you obtained in (1). Compare the PC1 axis and the regression line obtained above. Are they very different or very similar?

2. (40 points) Perform linear regression on the diabetes dataset using `sklearn`. This dataset can be accessed via the following python code:

```
from sklearn import datasets

diabetes = datasets.load_diabetes()
```

and the `linear_model` module in `sklearn` can be accessed via the following python code:

```
from sklearn import linear_model
```

This diabetes dataset contains 10 features/variables. Select `diabetes.data[:,2]` as x for linear regression. The dependent variable y is `diabetes.target`. Split x and y into training and testing sets by randomly selecting 20 points for testing and the remaining for training. Plot the testing x vs testing y , and the testing x vs predicted y in the same plot.