Machine Learning for Cities
CUSP-GX-5006-002

Prof. Luis Gustavo Nonato

Assignment I

Release date: Jan. 31st, 2017 Due date: Feb. 21st, 2017

Linear Regression on Raw and PCA Data: a comparative study

Motivation

A common approach in machine learning is to accomplish the learning process not directly on raw data, but on features extracted from the data. The reasoning is to map the original data to a feature space so as to remove redundancy while still reducing data dimensionality and noise, thus making the underlying learning process more stable and accurate. PCA is a typical mechanism used to map data to a feature space, as it ensures attribute decorrelation, noise removal, and dimensionality reduction.

Description

For this assignment students should:

1. Perform linear regression using as training data both original raw data and PCA-featured data;
2. Choose an appropriate number of principal components to be used in the PCA feature space transformation, arguing about the choice; A numerical study as to the impact of the number of components in the regression model is highly recommended;
3. The accuracy of each regression model should be assessed via bootstrap strategy;
4. Comparisons involving more than one regression method is encouraged (pure Least Squares, Ridge and Lasso).
5. Results should be presented in an assignment report (template in NYU Classes) using a combination of tables and graphics.

Data Set

The data set "`manhattan-dof.csv`" (downloadable from NYU Classes) can be used as training set. It is important to make clear in the assignment report which attributes have been taken as dependent and independent variables. If other data set is used in the experiments, a clear description of its attributes (and the chosen dependent variable) should be included in the assignment report.