

Project One: Exploration and Regression in High Dimensional Space (Due Sunday March 16 2003)

The data set for this project is called Project 1 Data, which can be downloaded from the course website. There are fifty independent variables denoted by x_1 to x_{50} , one dependent variable y and 400 observations. It is a synthetic data set. Assume that we are interested in patterns and dependencies in the data. To accomplish the project, you may use any programming languages and statistical packages as you wish. You are required to use SAS Enterprise Miner at least once for one of the following tasks.

- a) Explore the data and report special patterns. Discuss whether we would encounter the high dimensionality problems in the data.
- b) Three types of regression methods (selection, shrinkage, using derived directions) were discussed in class. Employ at least one from each type to find the dependency between y and the independent variables and compare their performances. Cross validation should be used both for model selection and error estimation.
- c) Mine possible relations between the independent variables. Discuss their influences on part b).

The report of the project should be written in the format of a journal paper. It is due to my office on Sunday March 16.