Introduction to Learning and Intelligent Systems - Spring 2015

bgrigoro@student.ethz.ch
norlundt@student.ethz.ch
fabmuell@student.ethz.ch

March 15, 2015

Project 1 : Regression

Data Analysis

In a first step, we analysed the data set according to some regularities. We plotted different time values as well as weather parameters against the Y-values to get an intuition of the dependencies in the data set. It turned out, that the time parameters influence the behavior of the travelers much more than the weather. Figure 1 shows how many passenger were traveling at which hour of the week. Based on this knowledge, we tried to find the best fitting regression model.

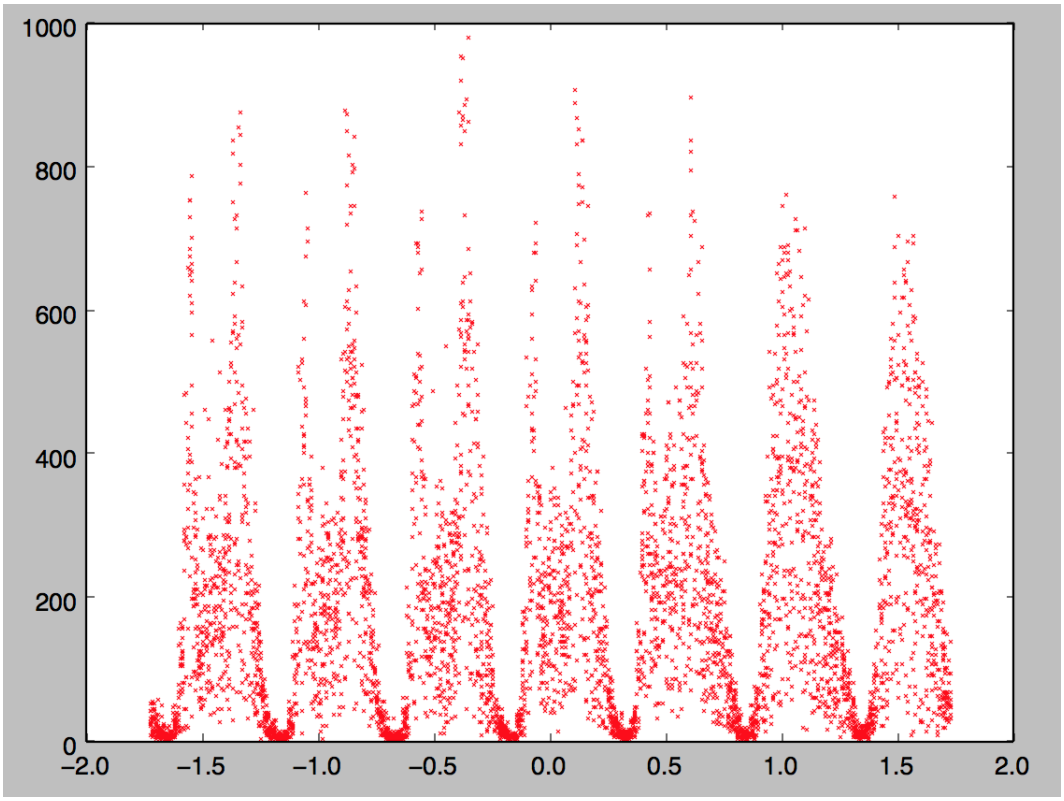

Fig. 1: X: week time in hours, normalized. Y: number of passengers

Model

In a second step, we had to evaluate different models, each with different parameters. We used a k-fold cross validation from the scikit-learn function set and brute forced the parameters or used the grid search function from the same library. When we used too many features for the learning and prediction, all the models seem to overfit the regression. So we tried to extract the features with the highest impact and neglect the others. The SVR-model with an RBF kernel seemed to approach the regularities of the data set best, so we decided to optimize this model by adjusting its input parameters.

Parameters

We tried many different models including: Ridge regression, Linear regression, SVR, Nearest neighbour and Logistic Regression. From all of them SVR performed the best, yielding a score of 0.57 on average. The parameters that gave the best result after the grid search are: epsilon = 1e-1; C = 10e2; gamma = 10e3; d = 15. The final model uses almost all (99.5%) of all input samples as support vectors, which is bad.

Features

We tried many combinations of parameters. For the time we used two separate features: time of the week in hours and month.

We saw there was less variability when plotting the week time occurrences. During the night there were few travelers. During a weekday people used trains more around rush hours in the morning before work and in the afternoon after work. On the weekend, on the other hand, people used the trains at more spread-out times, with most travels focused at midday.

We also used month, because there was a clear rise in travels during the summer months compared to the winder months, but there was still some variability. For example there was more travels around holidays like Christmas and the model took that into account.

When we incorporated the any combination of the weather features, we always got a result that was worse. For this reason we didn't use the weather features for our final model. We would appreciate feedback why that was the case.