Machine Learning Project Description

Joshua Saxton

As a side note, I was looking for data to use in my project and so I signed up for Kaggle. Their "Find your dream dataset" tool, uses amusing categories to try to find interesting data sets.

I decided to use the Allstate Claims Severity data to study machine learning techniques on.

1. Problem Definition

Allstate has provided sample data for predicting which claim will be a loss.

2. List of methods Allstate said they would evaluate submissions on MAE(Mean Absolute Error) from the predicted loss and the actual loss so that is one of the techniques I plan to evaluate as well.

I plan to perform linear regression since there is enough data that the dependent matrix X should be non-singular. I would like to compare the performance to ridge regression.

3. Brief description of the data (e.g., # of samples and features)

The train claims severity dataset has 188318 instances and 132 features or attributes. 116 of these features are categorical variables and 14 are continuous variables.

Part of the problem some people mentioned in the forums was determining if the test and train data had the same categorical variables (i.e. A is present in test but nowhere in training).

4. Description of Data settings or preprocessing (if need)

A first step would be to generate a correlation matrix between variables and see which ones are more highly correlated to one another.

Since we have just learned Principal component analysis, it might be worthwhile to compare that method as well.