Machine Learning Engineer Nanodegree

Capstone Proposal

Jose Rodriguez
June 20th, 2017

Proposal

Domain Background

Before my enrollment on the Machine Learning Nanodegree Program I had an Idea of what Supervised Learning was, but now after seen all the scope of what this course offers me, I have a very clear understanding of what to do to solve multiple problems and situations.

Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.

In this Capstone Project, I had the option to select a topic in Kaggle, so found a competition that I think suits my needs for this project, is the "Allstate Claims Severity Challenge": Allstate is currently developing automated methods of predicting the cost, and hence severity, of claims. In this challenge, Allstate enforce you to show off your creativity by creating an algorithm which accurately predicts claims severity.

(https://www.kaggle.com/c/allstate-claims-severity)

Problem Statement

According to Allstate as stated in the Kaggle challenge: "When you've been devastated by a serious car accident, your focus is on the things that matter the most: family, friends, and other loved ones. Pushing paper with your insurance agent is the last place you want your time or mental energy spent. This is why Allstate, a personal insurer in the

United States, is continually seeking fresh ideas to improve their claims service for the over 16 million households they protect."

Using the data supplied by Allstate in the Kaggle challenge, that is a recompilation of real data from claims reported to Allstate (this data was transformed to protect the private and personal information), we can predict the claim cost by using Supervised Learning algorithms, specifically Regression algorithms to predict the continue value of a Claim Loss.

Datasets and Inputs

The data is provided in (https://www.kaggle.com/c/allstate-claims-severity/data).

Each row in this dataset represents an insurance claim. You must predict the value for the 'loss' column. Variables prefaced with 'cat' are categorical, while those prefaced with 'cont' are continuous.

The training Dataset has a total of 587633 claims with a total of 116 categorical features and 14 continues features, one "loss" column (target)that refers to the Loss of the Claim.

To download the data, you need to have an account in Kaggle and accept the rules of the challenge, the dataset size is 22mb aprox.

Solution Statement

To solve this problem, we can train a Supervised Learner with the data we have for the features of the claim using the provided target column, after that we can train and test which of the Regression models is the most accurate on predicting the Cost (Loss), and after that we can use the model to predict the Loss for the test dataset.

The selected algorithms would be:

1) Support Vector Machine (linear)

    - In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. We will be using Sklearn's LinearSVR implementation.

        - https://en.wikipedia.org/wiki/Support_vector_machine

2) K-Nearest Neighbors

- In machine learning, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. We will be using Sklearn's KNeighborsRegressor implementation.
    - https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

3) Boosting

- Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. We will be using Sklearn's AdaBoostRegressor implementation.
    - https://en.wikipedia.org/wiki/Boosting_(machine_learning)

4) Stochastic Gradient Descent

- Stochastic gradient descent (often shortened in SGD), also known as incremental gradient descent, is a stochastic approximation of the gradient descent optimization method for minimizing an objective function that is written as a sum of differentiable functions. In other words, SGD tries to find minima or maxima by iteration. We will be using Sklearn's SGDRegressor implementation.
    - https://en.wikipedia.org/wiki/Stochastic_gradient_descent

Benchmark Model

As benchmark, I will be using a Support Vector Machine model to compare the MSE to the MSE obtained for the Capstone Project. We will be training/testing this model with the same process of the other classifiers following the Project Design described below.

Evaluation Metrics

I will be using the Mean Squared Error. In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated. While MSE is greater it is a worst result.

(https://en.wikipedia.org/wiki/Mean_squared_error)

(http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html)

Project Design

First, we need to do data exploration: we will determine the structure of the data, learn about the distribution of the data, how many features we have and get the minimum maximum and average values for the continuous values.

Then, we need to follow a process to Train a model, Test and validate the effectiveness of the implemented learner, basically we are going to use three Regression algorithms following this process:

1) Process the data: identify the features and target columns, we need to assure that we have numeric values in all columns, it is often the case that the data contains non-numeric features. This can be a problem; as most machine learning algorithms expect numeric data to perform computations with. We also need to encode the categorical features to handle the labels as Boolean features (0,1)

2) Reduce the dimensionality: As we have several features and it will grow a lot when we encode the categorical features, we need to find a way to reduce the dimensionality of the features. We will be using Sklearn's PCA and SelectKBest to achieve this.

3) Training and Testing Data Split: We will need to split the data into training and testing data, this will help us to measure the precision of our model after the training is done.

4) Model Evaluation: Here I will be using the selected models to perform the training and after this will compare the results of each of them based on their performance metrics, I will be using the MSE of every model after the training/testing and based on their error select the best model.

5) Model Tuning: In this step, I will be tuning the selected model (if possible), to see if I can obtain better results. We will be using Sklearn's GridSearch to tune the parameters.

6) For the next step, I will report the results of each of the models and their scores and compare it to the Support Vector Machine results from the pre-tuning face.

7) For the last step of the process I will be reporting the Final MSE obtained for the selected model and then give some conclusions about the obtained results.