Machine Learning Nanodegree Proposal

Chandra Nayak 07/24/2018

Domain Background

This project explores various Machine learning mechanisms that can be used for predicting future sales. Predicting sales over a future period involves time series forecasting; here, we shall frame this time series problem into a supervised learning problem which allows to apply linear regression algorithms.

We shall examine if using Ensemble methods makes sense, also explore Deep Learning machine learning algorithms. The purpose of this project is to identify the right model such that we reduce the variance between actual and forecasted sales revenue.

Supervised learning is the machine **learning** task of **learning** a function that maps an input to an output based on example input-output pairs. Supervised Learning algorithms can be grouped into “Classification” and “Regression”. Image Recognition, Recommendations and Time Series forecasting have been wildly popular applications that employ supervised learning algorithms;

KNN, SGD, Logistic Regression, Naive Bayes, Decision Tree have been used for years primarily as supervised learning classifiers, while Logistic Regression, Adaboost and ANN have been used for Regression. Deep Learning is a recent technique that has been gaining rapid adoption as a Supervised learning algorithm and I am looking to learn through this process how we can use Deep Learning for sales/revenue forecasting.

My motivation to do this project is that I am working with a couple of friends who intend to do revenue forecasting using on Machine Learning techniques and provide this as a cloud based subscription offering in the Indian market. Accurately forecasting Quarterly sales for a company operating in any industry has been a stubbornly difficult problem to solve. Good and accurate forecasting provides can help you develop and improve your strategic plans by understanding the marketplace better. Companies can make better budgeting when it has a clear sight to revenue, and manage its inventory controls, supply chain and financial planning effectively.

Problem Statement

Sales forecasting for any business has a lot of value because it allows them to do better budgeting, inventory and supply chain management, managing workforce in anticipation of growth, and adequate resource acquisition. But doing accurate sales forecasting needs better methodologies than doing mean average over prior quarters. In this project we will look at various machine learning techniques to predict sales for products across stores in a future quarter.

Datasets and Inputs

https://www.kaggle.com/c/demand-forecasting-kernels-only

The dataset I plan to use for this project is taken from Kaggle competition, which is 5 years of store-item sales data of 50 different items of 10 different stores. The data is provided in csv format which is split into train and test datasets with 80:20 ratio. Store data has following fields DATE, STORE, ITEM and SALES.

Solution Statement

I intend to use Ensemble of various models for solving this project. However, I have a strong motivation to explore the latest innovations in ARIMA & LSTM (Long-Short Term Memory model) , a powerful type of recurrent neural network (RNN) for time series analysis.

Benchmark Model

The benchmark for this model would be mean average of the prior quarters and expecting the future quarter to be at that mean, or equate a future quarter sales to the sales from the same quarter of a prior year.

Evaluation Metrics

The evaluation metric for this project would be Prediction Accuracy , F-Score, Root Mean Squared Error. For time-series forecasting using ARIMA we will look at ACF, PACF, plotting residual ACF plot to see model fit. For RNN we will look at the Error loss , and run epochs to get the least residual loss.

A model that returns with least variance between predicted sales for the quarter and the actual sales for the quarter on the test data will be considered the optimal model for this project.

Project Design

Project Design:

I believe I would be using the following technologies for solving this project:

Scikit-learn, Keras, Tensorflow, Matplotlib, ARIMA, Pandas

I would start with loading the training data available in the .csv format into a pandas dataframe. We will then use pandas describe() to get a statistical description of the data, before we do any preprocessing of the data.

We shall then do a visualization of the data by plotting Sales against Item, sales per year, sales per quarter to see if there is a time based pattern, and we shall also ploy sales against stores against time.

Once we have plotted the data, we can explore various using supervised linear regression algorithms and record the model accuracy by looking at metrics like Mean-Squared Error. If the model is a good fit, we can then run the model against the test data.

If the visualization of data indicated there is a pattern based on time, we will use time-series forecasting methods like ARIMA or Deep Learning forecasting methods like LSTM and run it enough number of epochs such that we arrive at the minimal residual loss, and optimal values for evaluation metrics like ACF and PACF. Once we have reached the minimal loss, we can use the hyper-parameters of the model and run it against test data to forecast the sales in the test quarter.

---

https://machinelearningmastery.com/time-series-forecasting-supervised-learning/

https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/

https://classroom.udacity.com/courses/ud980

https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials

http://www.business-science.io/timeseries-analysis/2018/04/18/keras-lstm-sunspots-time-series-prediction.html

https://www.youtube.com/watch?v=Aw77aMLj9uM