Machine Learning Engineer Nanodegree
Capstone Proposal
John Rangel
April 19th, 2017

Domain Background

Since the early 1970s, the video game industry has been consistently growing every year with the potential for huge sales increasing with it. Often times, it can be very hard to know how well a game will sell when released and which regions it will be most popular. Even highly rated video games can make very little for what they offer while lower rated. Regardless, it can be extremely useful if fairly accurate predictions can be made on how much sales a game will produce. Previous research has been done such as on the European market and showed that with the right formulas, sales predictions can be fairly accurate[1]. This research did mostly statistical analysis rather than machine learning techniques. One day I would like to get involved in the video game industry myself and would definitely like to know if a game idea on certain platforms would generate profitable sales.

Problem Statement

The problem is that the sales of a newly released game is unknown to most people but I believe that by combining different relevant features with unsupervised learning and then using supervised learning on those new features, the total sales of a game can be predicted. This is both a feature selection and a regression problem. Feature selection is needed because there are too many different features that could have influence on the generated sales and the actual predictions are a regression problem because we want to understand the relationship between the relevant features and the sales of the game.

Datasets and Inputs

I will be using a video games sales and ratings dataset available on Kaggle[2]. This dataset will provide me with ~6900 different titles with potentially useful features such as platform, title, metacritic ratings, genre, etc, which I can then compare to the sales of different regions. This dataset is actually based on a previous Kaggle dataset that was lacking some very useful rating information from metacritic[3]. The core data from the original dataset was pulled from VGChartz.[4]. All this data is very useful for solving the problem at hand and has very similar inputs used by previous research in this domain.

Simply put, I will be using the most relevant inputs or features from the dataset and finding the relationship these features have to the game sales of different regions, measured in millions of units.

Solution Statement

As mentioned in the problem statement section, I believe a solution to predicting video game sales can be achieved through machine learning. First, preprocessing must be done in order to have consistent data with no missing data in any video game's features or inputs. Next, I plan to use unsupervised learning techniques to extract the most relevant features and even combine features if necessary. Once I have the desired inputs, I'll split the data into training and testing data. Finally, I'll fit the training data to its sales with different regression techniques and then make predictions on the testing data. Once optimized, I should be able to make fairly accurate sales predictions on any new data.

Benchmark Model

As a benchmark model, I plan to use a very simple linear equation that takes video game critic rating as an input and outputs sales. The basic idea of this model is that the more highly rated a game is, the more sales it is likely to make. Since predictions are being made for the total sales as well as region sales, a different equation will be needed for each region. Each linear equation will be the same format and the constants will be determined by fitting critic ratings vs sales. I don't expect this alone to be a very good way for predicting sales but it is probably better than random guessing.

Evaluation Metrics

The primary evaluation metric that will be used to quantify the performance of the benchmark and solution model is going to be mean squared error or MSE. MSE is the preferred choice because the estimations aren't expected to be entirely spot on. The sales predictions of the models are intended to be as close as possible to the real values but it is very much okay for them to be off by seemingly large amounts. MSE works perfect for this since it will penalize outliers more than the smaller error values. This is because as the error values get larger, the squares of the errors increases dramatically. The similar alternative known as mean absolute error is less preferred since it does not square the error and treats all errors equally.

Project Design

Before attempting to solve the problem of predicting video game sales prices, analysis of the dataset must first be performed to better understand the data and what information might be useful or important to know. First I will explore the data in python to understand the ranges and means of the values the model will be working with try to see if any trends exist just by looking at graphs of the data. However, it is going to come down to feature selection meaning the model should only be using the most relevant features to predict sales.

After analysis, I'll begin experimenting with different regression models such as support vector regression and decision tree regression. I might also experiment with neural networks to see if that produces more optimal results. If I definitely need to use more features to improve the model, feature transformation in the form of principal component analysis will be used to create better features for the model.

It is possible that none of the proposed methods evaluate to a model that can accurately predict video game sales. In this case, I will attempt to find more features than the dataset provides that will hopefully lead to better predictions. The first goal at the very least is to surpass the benchmark model. From there, it is all about improving the model until we have the lowest error possible.

References:

1. Beaujon, Walter Steven. "Predicting Video Game Sales in the European Market." (n.d.): n. pag. Web. <https://www.few.vu.nl/nl/Images/werkstuk-beaujon_tcm243-264134.pdf>.
2. Smith, Gregory. "Video Game Sales." Video Game Sales | Kaggle. Kaggle, n.d. Web. 20 Apr. 2017. <https://www.kaggle.com/gregorut/videogamesales>.
3. Kirubi, Rush. "Video Game Sales With Ratings." Video Game Sales with Ratings | Kaggle. Kaggle, n.d. Web. 20 Apr. 2017. <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>.
4. "Video Game Charts, Game Sales, Top Sellers, Game Data." VGChartz. N.p., n.d. Web. 20 Apr. 2017. <http://www.vgchartz.com/>.