October 31st, 2016

Machine Learning Engineer Nanodegree

CAPSTONE PROPOSAL

FAISAL HAMID

Domain Background

Trading in financial securities began hundreds of years ago in Europe and revolutionized the way in which economies and global commerce functioned. Since then, people have been dedicating time and resources to research in order to gain an advantage over competitors and other investors.

Global market capitalization of equities was approximately USD 62 trillion with the United states accounting for USD 25 trillion of that[1]. More recently, companies in the United States and around the world have resorted to digital online filings of their financial disclosure. This presents opportunities to perform analysis on data not previously available to most researchers; however, the large amount of data makes this a challenge.

Anytime new information for any domain is available, the potential for reward also goes hand in hand. Identifying differences between successful and unsuccessful companies can provide all stakeholders with valuable information regarding future performance of public companies. In addition, the possibility of identifying fraudulent or failing companies before they can do too much damage is also a great motivation.

Problem Statement

Until recently, financial statement data was not available in the public domain and compiling proprietary databases was expensive and tedious. With the availability of data over the internet, the new challenge is attempting to extract useful knowledge from large volumes of, often incomplete and inaccurate, data. Complicating this challenge is the fact that we are not entirely sure of how much useful information is actually there.

Advances in machine learning, particularly unsupervised learning techniques, have the potential to discover underlying relationships that may not be evident at first. The challenge that needs to be overcome is implementing these techniques in the presence of incomplete and inaccurate data to extract previously unknown relationships.

Datasets and Inputs

In this exercise we will be focusing on US equities only as data is more readily available. The underlying source of the data is the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) managed by the Securities Exchange Commission (SEC). The compiled dataset can be found on Kaggle and is provided by “usfundamentals.com”.

The dataset contains financial statement disclosure filed by US companies with the SEC. Data is available for 2011 onwards and is available in either Quarterly or Annual frequency, with daily updates. Overall there are approximately 8,526 unique indicators reported for over 12,000

[1] Data hosted by The World Bank (http://data.worldbank.org/indicator/CM.MKT.LCAP.CD?locations=US&name_desc=true)

companies. However, there only 20 indicators that are reported across most companies according to the website.
In this project, we are going to explore this dataset and attempt to discover underlying relationships between companies. The goal is to determine if we can classify companies in to unintuitive segments that can enhance our understanding of the types of public companies and their characteristics. However, the challenge is to find a way to do this in the presence of high dimensions, missing data and potential inaccuracies.

Solution

We are going to look at the cross-section of all companies and the most commonly reported features to discover any clustering amongst the companies and the values reported. In addition, we will have to explore techniques for feature transformation, selection as well as dimensionality reduction in order to accomplish our goal. The focus will be on finding clusters in the sample at hand, rather than prediction or out of sample analysis.

Benchmarks

Although there is no specific benchmark model for an unsupervised learning exercise with this particular dataset, the identification of clusters based on relevant evaluation criteria will serve as the ultimate goal. Therefore, the underlying assumption can be that there are no clusters and we will attempt to prove otherwise.

Evaluation Metrics

The primary evaluation metric will be the silhouette coefficient when evaluating the presence of clusters. The silhouette coefficient can be calculated using the "silhouette score" function available in the scikit-learn library. The score ranges between -1 and 1 with negative values indicating a lack of confidence in the clusters that are assigned while high positive values show high confidence clusters detected.

Project Design

Although there will be some overlap regarding where any specific process is applied, a rough outline can be constructed as follows:

- Data Exploration
  - This section will explore all available features and reporting companies as well as the time periods available. The purpose of this part is to understand what the dataset consists of and whether all of it, or a subsample, is actually relevant to the study.
- Feature Selection
  - The next part of the pre-processing stage will be selecting relevant features from the roughly 8500 unique indicators. The primary concern will be to maintain the maximum sample size and to remove any corrupted features.
- Outlier Detection

    - This part may be the most important of all as it will be quite a challenge detecting true outliers without excluding observations that can really help distinguish the underlying relationships.
- Feature Transformation
    - Given that there is a wide variety of features, we will have to make sure that the distributions are normalized so that the algorithms used for dimensionality reduction and cluster detection are not thrown off.
- Final Preprocessing
    - This final preprocessing step will be used to plot and explore the transformed features and make any final removals or adjustments to the selected features. The goal is to use the most independent components, while also respecting sample size.
- Dimensionality Reduction
    - Use principal components analysis (PCA) to reduce the dimensionality of the feature space. Will have to use some judgement to find an acceptable tradeoff to the explained variance and complexity tradeoff.
- Detecting Clusters
    - This section will attempt to find clusters in the PCA transformed data. Other methods including Independent Components Analysis (ICA) will be also be explored to see if any hidden segments can be discovered.
- Exploring the Clusters and their properties
    - This part will explore any clusters found and see what characteristics the companies in each group share with each other. As a bonus exercise, we will look at whether there are any differences in how the equity prices have performed over the recent past.
- Final Conclusions
    - Summarize all findings and discuss the direction for future research.