Chapter 2

Concepts

2.1 Efficient Market Hypothesis

When looking at mature financial markets it's often said they are efficient. Meaning that products somehow get the correct price given the available information. It is however not easy to define this right price. A weaker statement is that the market can be considered to as a surjective relation, meaning that it results in a price for every set of considered information. This considered information consists not only of fhe fundamental information about a company or its sector but also the state of the market, and general sentiment. This all seems a bit moot and overly theoretical but it is important to underscore that the price set by the market is a variable and not some intrinsic truth.

The system of market prices is kept largely consistent by actors trying to outcompete each other to profit of the inconsistencies. It is however important to realize that they are mostly following monetary incentives which can sometimes cause market prices to seem disconnected of actual value. A good example of this is the negative price of crude oil futures with delivery in May 2020 days before they expired in April. The negative price here was due to the fact that that the holder these futures after the expiry is forced to receive a large amount of physical oil. A lot of these futures are however bought by investors only speculating on the oil price. The preceding crash in demand and price had however caused storage to be filled meaning that investors had to pay for other entities to take the oil off their hands.

2.1.1 Consequences for machine learning

The prediction of financial variables is something a lot of very smart people have spent a lot of time on. If one could could predict a significant change in for example price with a marginally large amount of certainty, he could exploit this change with an arbitrarily highly levered position. A perfect prediction would thus imply arbitrarily high profits. Alas the world is not this simple, these kind op arbitraging possibilities are rare, and taking a position will influence the market which will prevent us from making infinite profits. The reason that these opportunities are rare is expressed by the previously expressed efficient market theorem, it is these effects that doom

the naive application of even very complex machine leanring algorithms to trading. They can simply not compete with the price setting power of the market.

2.2 Implied Volatility

Financial product prices always have an inherent variability. A simple measure of this is volatility or standard deviation. This is a backwards looking measure, indicating the amount of movement in prices over a certain time frame. When trading estimating volatility in the time to come is very important. One place where this is clear is in option pricing. Options are contracts on some underlying product allowing the holder to buy or sell an amount of that product at or before a certain time. Mathematical methods for estimating the price of a certain option generally require at least an estimate of the volatility in the time frame of the option. This means that the price of an option implies some volatility in the future. This implied volatility is therefore an indicator of the volatility expected and priced in by market actors. When we want an aggregate view of volatility expectations we can therefore try to solve an option pricing problem in reverse, by inputting the prices and solving for the volatility. One example of implied volatility is the VIX index calculated by CBOE, reflecting the volatility expected by the market based on 30 day options on the S&P500. This index is sometimes called the fear index as a high value indicates uncertainty in the market, but it is important to understand that it does not imply a direction for the expected movements.

2.2.1 Calculation

Add calculation with black scholes and shiiet

2.3 Sentiment Extraction

Humans are emotional creatures, and most of their communication carries emotions or sentiment. While it's often trivial for a human to evaluate the sentiment of communication this becomes a lot harder when it has to done by algorithms. Sentiment rests on a set of shared understandings and knowledge, and deep understanding requires understanding of the human experience. However it is very attractive to have algorithms that understand the emotional load of some communication. In this thesis we focus on two different modalities of communication, text and visual and try to use them to make some predictions. There are however different ways of extracting sentiment. The most simple one would be a simple polarity, expressing the positivity or negativity of a sample. But just because a message is not positive it doesn't mean it is negative it could also be neutral or non emotional. This implies that there are multiple values we could use to represent sentiment polarity. But sentiment polarity is not the only emotional information in a given message, humans experience a lot of emotions that vary in more than the basic polarity spectrum. A common model is Plutchik's wheel of emotions. Here he uses 8 primary emotions and models a total of

24 'dyads' or composite emotions. Sentiment is often embedded in a vector of these 24 dyads, with each number representing the strength of that emotion.

2.3.1 Textual Sentiment Analysis

Sentiment analysis of text is a common area of research within Natural Language Processing. There are several different common approaches to this. The most simple ones use a bag of words approach, simply looking at all the words in a sentence or text, assigning them a value and aggregating a sentiment from them using some human compiled lexicon. This approach has been demonstrated to work but will still miss a lot of meaning, such as negations, adverbs, homonyms, etc. More complex methods will examine sentences and consider context, an example of this approach is VADER[4]. Here they use a lexicon aimed at social media and microblogs. These approaches remain largely engineered and not extremely complex. More advanced methods are however becoming available, they often leverage some kind of (very simple) embedding on a sentence level and use things like neural networks for predicting the sentiment. Neural text embedding was traditionally done by some kind of recurrent neural network (eg LSTM, GRU, ...) but transformer networks are quickly gaining traction[5]. In this work we elected to use VADER due to its simple nature and its good performance on microblog text.

2.3.2 Visual Sentiment Analysis

Image analysis is a considerable younger field in artificial intelligence, but driven by the GPU revolution it has become a very important field with applications from object recognition to image denoising. Neural networks with images as input commonly use convolutional layers[6] to decrease computational and space complexity while increasing learning performance through the spatiality of features. Visual sentiment analysis is a recent but growing area of research. Compared to other applications of deep learning with images there are a lot of challenges. The most important ones are probably that the amount of generalization necessary is higher than that of for example an image recognition problem. One reason for this is that the intraclass variety is much higher, and classes less clearly distinct. The difficulty is also increased by the lack of large enough, well labeled datasets. While big datasets exist they are often weakly labeled.

Basic neural sentiment analysis often starts with a pretrained classifier[7]. These generally consist of a convolutional part embedding the image and a fully connected layer performing the classification. The fully connected layer from the pretrained classifier is removed, effectively resulting in a pretrained image embedder. Sometimes one layer is added and trained directly to a labeled dataset to predict sentiment[8]. Another technique is to first add a layer, to train towards a visual sentiment ontology like SentiBank[9], which labels images with Adjective Noun Pairs. These adjective noun pairs model some concepts detected in the image. After getting the anp we can try to extract the sentiment from these ANP's, for this we would need an onthology fitting the set of ANP's

2.4 Deep unordered composition

Most deep learning regression tasks have a fixed dimension in their input and output size. For a given imput sample we want to predict some variable. but more recently we have started to use varying sizes of input and or output size for this, most specifically in language learning tasks. Because in language location is very important this often takes the form of a sequence. Such as a sentence or a document. Most approaches to this use some kind of internal state, this is the case in recurrent neural networks. They iterate over the input and maintain some internal state. These models however have some issues. They tend to 'forget' previous values and have issues with longer sequences. They are considered hard to train and need a lot of data points when compared to normal fully connected networks. A new evolution is so called attention networks. Here we allow the states to 'attend' over different states, meaning that it considers the state of other places in the sequence without enforcing some order and sequential internal state. It has to be understood that all of these methods assume a certain and predictable order on their data, which for example is not the case when looking at our problem. They also often require the sequence size to be in the same order of magnitude for each datapoint, another assumption not valid in our case. In NLP however models disregarding the inherent order in their samples are also used. For a sentence they are often called Bag Of Word models, for representing a sentence they will just represent each word and then aggregate those representation. It is a generalization of this concept we'll use to create our deep learning models. This approach is often called deep unordered composition. And while not considering syntactical information it performs well, with small simple models that are really trainable.[10]