Memoria TFM

Juan Luis Ruiz-Tagle
Título tentativo: Predicting stock market prices through text media sentiment analysis

Reunión 1 - 12/2/2020

Se aborda el tema del tfm, las posibilidades. Sentiment analysis y predicción de valores.
Emilio me explica los plazos y deja claro que:

- Hay que reunirse cada 2-3 semanas.
- El TFM tiene que aportar algo novedoso para ser válido. No vale replicar resultados.
- Definir el scope del proyecto
- Hay que leer el estado del arte.

Avances

Algunos papers leídos sobre el estado del arte:

- Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction
- Deep learning for sentiment analysis: A survey
- Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts
- Wasserstein Learning of Deep Generative Point Process Models
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Generative Adversarial Networks
- Conditional Generative Adversarial Networks
- Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets

Arquitectura del sistema:
Quiero usar una arquitectura de tipo cGAN (conditional GAN) para las predicciones. Las cGAN son GAN en las que tanto el generador como el discrimnador reciben el mismo input extra, que orienta sus capacidades respectivas de generar y discriminar.

- Dado un activo y una fecha el generador intentará crear una serie temporal prediciendo las tendencias de los siguientes 7 días a la fecha dada. Usaré una LSTM como generador.
- El discriminador determinará si la serie presentada ante él corresponde a la real de los próximos 7 días o ha sido generada. Usaré una red CNN de 1 dimension como discriminador.
- Tanto G como D reciben el input de BERT (red neuronal de Google para NLP tasks en general), que ha procesado un conjunto de noticias relativo a un activo en concreto para la fecha dada. Estos datos pretenden orientar a G y D para generar series más fidedignas y distinguirlas mejor de las reales respectivamente.

Aportación novedosa:
Me he esforzado en buscar papers en los que se aplique la arquitectura de las cGAN para predicción del mercado de valores usando técnicas de NLP (BERT) como apoyo condicional y no he encontrado nada, por lo que parece que este setup es algo novedoso. De todas

formas, para garantizar que hay una novedad, quiero utilizar una nueva Error function para el generador que me he inventado, en la que el error entre la serie real y la predicha se basa en RSME, pero da más peso al error de las predicciones de los primeros valores de la serie que a los últimos. Esto interesa ya que los valores que mejor queremos predecir son los imediatamente posteriores a la fecha actual. Los de más adelante los predeciremos mejor con las noticias de los días posteriores (que todavía no se han publicado). La función quedaría así:

$$\frac{1}{2C}\sum_{i=1}^{n}\frac{1}{i^k}(y_i - \bar{y}_i)^2$$

Scope del proyecto:

Por concretar he elegido algunos activos con los que entrenar las GAN. Pretendo definir una serie de keywords para cada activo y recopilar noticias sobre cada uno de ellos para cada día en los últimos 3 años (este proceso lo automatizaré con python usando la librería news-please). Escogeré los primeros articulos que encuentre en Google Noticias haciendo que aparezcan por "orden de relevancia" (así me aseguraré de que hayan tenido impacto).

Los activos los he escogido del SP500, y he procurado que representen series de alta y baja volatilidad, y de tendencias crecientes y decrecientes, para sí tener un poco de todo.

Tendencia . Volatilidad	Creciente	Regular	Decreciente
Alta	Bitcoin (BTCUSD)	Tesla (TSLA)	Kraft-Heinz (KHC)
Baja	McDonalds (MCD)	PepsiCola (PEP)	Occidental Petroleum (OXY)