Texas A&M University
Department of Computer Science & Engineering
Financial Computing 3

A Study of Deep Learning in Micro-Economic Analysis and Computational Finance

Zebo, Xiong & XXX, XXX
Ph.D Student Ph.D Student

Supervisors:
Dr. Anxiao, Jiang

First version

College Station, December 2019

Abstract

During the last decades, computational methods gradually become the essential part of economy and finance researches. With the rise of attention on machine learning techniques, the investors and financial analyst realize the importance of utilize the computers' power to help them make critical business decisions and predict the future's trends.

Table of Contents

1. Collections

This chapter collect the studies materials (including source from technology blogs, etc.) and technical information for the research

1.1 From Student

1. **Financial Time Series Study**
   www.math.pku.edu.cn/teachers/lidf/course/fts/ftsnotes/html/_ftsnotes/index.html#e8afbee7a88be58685e5aeb9
2. **MathJax basic tutorial and quick reference**
   https://math.meta.stackexchange.com/questions/5020/mathjax-basic-tutorial-and-quick-referenc
3. **Math formula editor**
   https://www.codecogs.com/latex/eqneditor.php?lang=zh-cn
4. **Latex Online Formular**
   https://latex.91maths.com/
5. **LaTeX**
   https://blog.csdn.net/simple_the_best/article/details/51244631
6. **List of Greek letters and math symbols** https://www.overleaf.com/learn/latex/List_of_Greek_letters_and_math_symbols

1.2 From Professor

1. **Bryan Kelly**
   https://www.bryankellyacademic.org
2. **Dacheng Xiu**
   https://dachxiu.chicagobooth.edu
3. **Stefan Nagel**
   https://voices.uchicago.edu/stefannagel
4. **Take a look at their recent research papers (which you can find in their webpages, under "research"), to get some basic ideas? (If the papers are hard to understand, that's all right. You just need to get some basic "feeling" at this point.)**
5. **Stefan Nagel - A nice tutorial on ML for asset pricing**
   https://bcf.princeton.edu/wp-content/uploads/2019/05/PrincetonLecturesSlides_Day1_handouts.pdf

2. Research Progress

This chapter main elaborate the research progress about machine learning applications in economics and finance

2.1 Week 1 (Dec 22 - Dec 27)

1. Downloaded seven papers:
    - Autoencoder Asset Pricing Models (Shihao Gu, Bryan Kelly, Dacheng Xiu, 2019 Sep.)
    - Empirical Asset Pricing via Machine Learning (Shihao Gu, Bryan Kelly, Dacheng Xiu, 2019 Sep.)
    - Text as Data (Matthew Gentzkow, Bryan T. Kelly, Matt Taddy, 2017)
    - Market Efficiency in the Age of Big Data (Ian Martin, Stefan Nagel, 2019 Dec.)
    - Asset Pricing with Fading Memory (Stefan Nagel, Zhengyang Xu, 2019 Aug.)
    - High-Frequency Factor Models and Regressions (Yacine Ait-Sahalia, Ilze Kalnina, Dacheng Xiu, 2019 Sep.)
    - Principal Component Analysis of High-Frequency Data (Yacine Ait-Sahalia, Dacheng Xiu, 2017)
2. Studied the hangouts from Dr.Stefan Nagel, and with below take away:
    - Studied the concepts of Stochastic Discount Factor
    - Learnt the key words in the application: Lasso, Ridge, Cross-section, near-arbitrage, etc.
    - Learnt models such as: Fama-French Three Factor Model, Capital Asset Pricing Model, etc.
    - Penalized regression representation: (Two-penalty Specification, allowing sparsity)

$$\mathbf{b} = \arg\min_{\mathbf{b}} \left\{ \left(\bar{f} - \Sigma\mathbf{b}\right)' \Sigma^{-1} \left(\bar{f} - \Sigma\mathbf{b}\right) \right\} + \gamma_1 \mathbf{b}' + \gamma_2 \sum_{i=1}^{H} |\mathbf{b_i}|$$

3. Read the abstracts of the seven papers.
   Key terms are: latent factors, estimates nonlinear conditional exposures, impose the economic restriction of no-arbitrage, trees and neural networks, comparative analysis, statistical methods for text analysis, impose shrinkage (ridge regression) or sparsity (Lasso) when estimate coefficients, fading memory, high-frequency, Fama-French factors, asymptotic distribution
4. Downloaded old papers such as:
    - The Cross-Section of Expected Stock Returns - Eugene F. Fama; Kenneth R. French
    - Econometric Policy Evaluation: A Critique - Eobert E. Lucas, Jr.
    - Seasonality in the Cross-Section of Stock Returns - Steven L. Heston and Ronnie Sadka
5. Reviewed Latex syntax and collected some study material - will continue study with them .

2.2 Week 2 (Dec 28 - Jan 4)

1. Dive into paper "High-Frequency Factor Models and Regressions"
    - Un-systematic Risk (individual firm risk): Random events such as CEO step down, company being prosecuted, company news, etc; Systematic Risk (market wide risk): nationally cut interests rate, etc;

- When portfolio size increase, the volatility will decrease because "Unsystematic Risk" averaging out.
- Total Risk = Systematic Risk + Unsystematic Risk
- How to measure Systematic risk? Beta
- Beta = the percentage change in an asset's return GIVEN a 1% change in the market portfolio (market portfolio = all firms) - how individual return tied to market portfolio change

2. Data from Thomson Reuters

   - Login Page:
     - `https://php.library.tamu.edu/resources/datastream/` or
     - `http://eikon.thomsonreuters.com/login`
   - Credentials:
     - Account: student3@tamu.edu — Password: TAMUwcl1
     - Account: student2@tamu.edu —Password: WCLtamu1
     - Account: student1@tamu.edu — Password: WCLtamu1

3. Data from Eikon Data API

   - eikon PyPI `https://pypi.org/project/eikon/`
   - Quick Start `https://developers.refinitiv.com/eikon-apis/eikon-data-api/quick-start`
   - Tutorials `https://developers.refinitiv.com/eikon-apis/eikon-data-api/learning`

4. Sample Minutes Price (3 months)

Tesla Inc | Price History | Thomson Reuters Eikon 15:58, 04-Jan-2020

TSLA.O
Interval: 1 Minute
History Period: 31-Dec-2019 17:47 - 03-Jan-2020 17:47

VAP: Total	27.35M	
Price	Volume	Count
450 - 455	2.0513M	15
445 - 450	9.2229M	248
440 - 445	5.4682M	288
435 - 440	752,853.0000	53
430 - 435	1.5431M	273
425 - 430	6.1464M	406
420 - 425	2.1561M	401
415 - 420	6,255.0000	36

TSLA.O Price Statistics 1 Minute 31-Dec-2019 17:47 - 03-Jan-2020 17:47

Price			Volume				Up/Down (C-C)		Price Change	Close - Close	
High	454.010000	#########	Max	414,418.00	#########	Advancing	14.85M	Up	807	Up	+0.9043%	#########
Low	418.330000	#########	Min	1.00	#########	Declining	12.39M	Down	784	Down	-1.2074%	#########
Avg	432.758003		Avg	15,899.48		Total	27.35M	Unch	128	Period	+5.6454%	

TSLA.O Price History 1 Minute 31-Dec-2019 17:47 - 03-Jan-2020 17:47

Local Date	Local Time	Close	Net	%Chg	Open	Low	High	Mid	Volume	Turnover - USD	Approx VWAP	O-C	H-L	%CVol	%CTurn
03-Jan-2020	17:47	442.200000	-0.270000	-0.0610%	442.210000	442.200000	442.210000		20			-0.010000	0.010000	-80.3922%	
03-Jan-2020	17:45	442.470000	+0.360000	+0.0814%	442.300000	442.300000	442.470000		102			0.170000	0.170000	+2.0000%	
03-Jan-2020	17:44	442.110000	0.000000	0.0000%	442.110000	442.110000	442.110000		100			0.000000	0.000000	+900.0000%	
03-Jan-2020	17:43	442.110000	0.000000	0.0000%	442.110000	442.110000	442.180000		10			0.000000	0.070000	-50.0000%	
03-Jan-2020	17:42	442.110000	-0.300000	-0.0678%	442.410000	442.110000	442.410000		20			-0.300000	0.300000	-90.0000%	
03-Jan-2020	17:41	442.410000	+0.020000	+0.0045%	442.390000	442.390000	442.470000		200			0.020000	0.080000	+2,757.1429%	
03-Jan-2020	17:40	442.390000	-0.010000	-0.0023%	442.390000	442.390000	442.390000		7			0.000000	0.000000	+75.0000%	
03-Jan-2020	17:39	442.400000	0.000000	0.0000%	442.400000	442.400000	442.400000		4			0.000000	0.000000	-96.1905%	
03-Jan-2020	17:36	442.400000	0.000000	0.0000%	442.400000	442.380000	442.400000		105			0.000000	0.020000	+3,400.0000%	
03-Jan-2020	17:34	442.400000	0.000000	0.0000%	442.400000	442.400000	442.400000		3			0.000000	0.000000	-97.1698%	
03-Jan-2020	17:33	442.400000	+0.300000	+0.0679%	442.400000	442.400000	442.400000		106			0.000000	0.000000	-87.1515%	

5. Columns names

   - Local Date
   - Local Time
   - Close
   - Net
   - %Chg

- Open
- Low
- High
- Mid
- Volumn
- Turnover - USD
- Approx VWAP
- O-C
- H-L
- %CVol
- %CTurn

2.3 Week 3 (Jan 5 - Jan 11)

2.4 Week 4 (Jan 12 - Jan 18)

3. Terminology

1. **Book-To-Market Ratio Definition**
   https://www.investopedia.com/terms/b/booktomarketratio.asp
2. **Price-to-Book ratio, or P/B ratio**
   https://en.wikipedia.org/wiki/P/B_ratio
3. **Market to Book Ratio**
   Reference Link 1
4. **Fama French Three Factors Model**
   Reference Link 1
   Reference Link 2
5. **Arbitrage**
   https://www.investopedia.com/terms/a/arbitrage.asp
6. **Near-Arbitrage**
   http://people.stern.nyu.edu/adamodar/pdfiles/invphilslides/session28.pdf
7. **Sharpe Ratios**
   https://www.investopedia.com/ask/answers/010815/what-good-sharpe-ratio.asp
   Reference Link 2
   Reference Link 3
8. **Factor Return**
   https://www.nasdaq.com/glossary/f/factor-return
9. **Hansen-Jagannathan Distance: Geometry and Exact Distribution**
   Hansen-JagannathanDistance:GeometryandExactDistribution
10. **Stochastic discount factor**
    Reference Link 1 Reference Link 2
11. **HansenJagannathan bound** https://en.wikipedia.org/wiki/Hansen%E2%80%93Jagannathan_bound
12. **Non-parametric Time Series Regression Model**
13. **Local martingale**
    https://en.wikipedia.org/wiki/Semimartingale
14. **Semimartingale**
    https://en.wikipedia.org/wiki/Local_martingale
15. **Linear Conditional Beta Specification**
16. **Book-to-Market**
    https://www.investopedia.com/terms/b/booktomarketratio.asp

3.1 Economics and Finance

3.2 Neural Network

4. Milestones

This Chapter marks down the major progress and important finding etc.

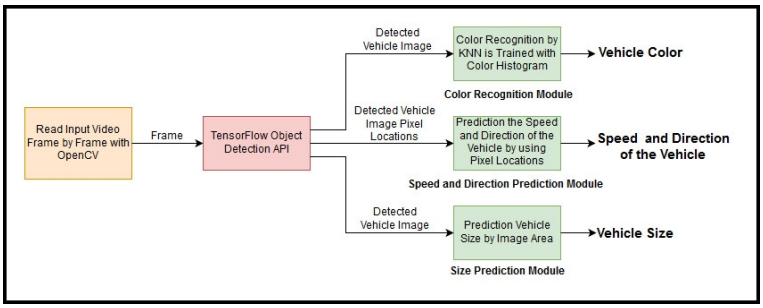
4.1 Before 2019/12

1. Replicated and improved below LTMS project
    - https://github.com/zebointexas/LSTM-Neural-Network-for-Time-Series-Prediction
    - Conducted experiments for Apple Inc. 's 2009-2018 price in NSDAQ, archived 75% accuracy when testing the result in 2019
2. Data Collection
    - Be able to collect seconds data via Robinhood (every three seconds) by developing with Robinhood API.
    - Be able to collect minutes data from Bloomberg for lag 3.
    - Be able to collect news, time series, constituents from Eikon by using Eikon library

4.2 After 2019/12

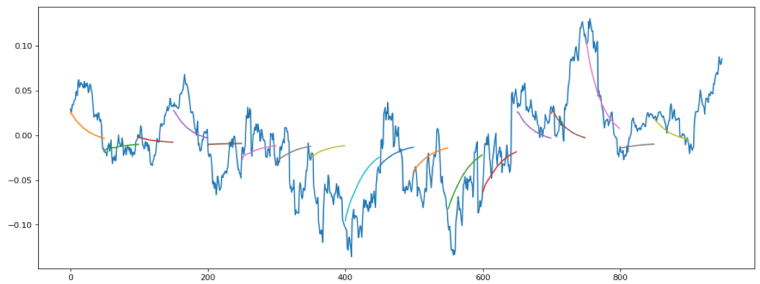
5. Previous & Parallel Works

5.1 Video Steam Vehicle Data Extraction (Parallel)

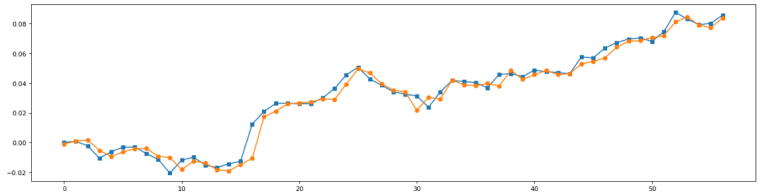
1. Replicated the project and attempting optimize the algorithm:
   h3://github.com/ahmetozlu/vehicle_counting_tensorflow
2. 

5.2 LSTM for Time Series Prediction (Previous)

1. Replicated and improved below LTMS project
   - https://github.com/zebointexas/LSTM-Neural-Network-for-Time-Series-Prediction
   - Conducted experiments for Apple Inc. 's 2009-2018 price in NSDAQ, archived 75% accuracy when testing the result in 2019
2. Sequential by Sequential prediction (4 Epoch, 50 Batch size)



3. Point-by-Point prediction (8 Epoch, 50 Batch size)



5.3 On-Off Campus Ride Sharing Platform (Parallel)

1. Built a responsive web platform with Django/Bookstrap/MySQL to help students post rides information.
2. Utilized GitHub and deployed the project on PythonAnywhere server for public testing.
3. Fine tuned and validated the front-end to meet requirements such as intelligent rides recommendation.

5.4 Facial Recognition & Django Development (Previous)

1. Developed a Twitter like social medial platform with Django which gives object and face recognition on the images uploaded for privacy protection purpose;
2. Incorporated OpenCV and Google Cloud Vision and Python face recognition package into the platform to process the image detection.
3. Designed and adapted machine learning algorithm with Python to train models for better detection (such as language characters and indoor objections) optimization.

5.5 Software Engineering - Program Invariants Detection (Previous)

1. Utilized Daikon software and Symbolic Execution techniques to generate candidate invariants and implemented selfdesigned algorithms to produced better invariants set.
2. Configured and deployed Java/Symbolic PathFinder on Unix environment with Gradle and Eclipse; Employed Homebrew to setup prerequisites such as Graphviz, Texinfo, GNU Compiler Collection (GCC) to support invariants discovery.

5.6 2019 Fall Class Summary (Previous)

1. CSCE 676 Data Mining and Analysis (Dr. James Caverlee)
    - Generated association rules (support ¿=minsup, confidence ¿= minconf) using hash-based Apriori to generate the rules)
    - Ranked the graph nodes using the Personal Page Ranking
    - Performed Text Mining, Sentimental Analysis, Topic Modelling, Expectation Maximization on millions of text file crawled from Google regarding corporate social responsibilities
    - Understood MapReduce programming model; Studied the concepts in Hadoop; Analyzed the tweets concerning members of the US congress (such as most mentioned users) on AWS using Spark with Pregel
    - Performed K-means for different clusters; Hierarchical clustering (single link complete link)
    - Tufte's Principles (such as maximizing the data-ink ratio erasing redundant data-ink)
    - Finding similarities using Jaccard distance with shingling, min-hashing, and locality-sensitive hashing (LSH)
    - Understood consistent hashing on Napster; Studied structured P2P approaches such as Chord using Finger table; Used counting bloom filter algorithm to test whether a movie is a member of a set;
    - Used Reservoir sampling in data stream; Adopted Morris algorithm and Flajolet-Martin algorithm to perform probabilistic counting
    - Studied Privacy-Preserving data mining for association rule mining, k-means clustering, and decision trees.
    - Used K-anonymity to remove personally identifying information; Understood differential privacy;
2. CSCE 633 Machine Learning (Dr. Bobak Mortazavi)

- Studied multiple linear regression with model fitting; Understood the concept of convexity;
- Studied multiple logistic regression and perform optimization using cross-entropy (gradient descent with different learning rate)
- Performed model selection with lasso(L2) and ridge(L1) regression
- Studied SVM using hyperplane to classify the data; several techniques are: 1) maximal marginal hyperplane; 2) Enlarge through kernel space; 3) Adopted Hinge Loss and Lagrange Multipliers
- Studied decision tree and used Gini Index to avoid overfitting; Studied random forests and AdaBoost
- Used PCA to reduce the dimension and K-Means to find patterns in clusters
- Studied Expectation Maximization to increase the likelihood
- Studied neural network with the concepts of perceptron, forward/backward propagation calculation.
- Learnt how convolutional neural networks use kernels and max-pooling techniques.
- Understood reinforcement learning and its Hidden Markov Model
- Predicted NASDAQ price with LSTM (RNN) using Keras API and archived 75

3. CSCE 636 Deep Learning (Dr. Shuiwang Ji)

- Studied the concept of Perceptron, threshold and Perceptron Learning Algorithm (PLA).
- Studied linear classification, Pocket Algorithm, Least Squares Linear Regression.
- Adopted Cross Entropy and maximized the likelihood to minimize the total error using Stochastic Gradient Descent.
- Calculated probability for multi-class logistic regression using softmax; Measured the error using corss entropy to calculate individual classs loss; Studied the shift-invariance in parameters (output remain the same); Studied the equivalence of sigmoid when cross entropy cope with binary classes;
- Calculated the derivative of software and derivative of cross entropy with softmax;
- Learnt different loss functions such as: Hinge Loss, Log Loss, Zero-one Loss; Studied convexity;
- Studied how to use regularization (Lasso, Ridge), cross-validation and early stopping techniques to prevent overfitting during model selection;
- Studied the multilayer perceptron (MLP) and the reason to use adopt tanh(.) activation function (original error is not smooth and cannot use gradient descent)
- Utilized Chain Rule to produce sensitivities during the backpropagation process.
- Compared different descents such as steepest descents, conjugate gradients, etc.
- Understood thoroughly about SVD and its relation to Eigen-Decomposition; Learnt compact SVD; Learnt how to compute PCA and the reconstruction process; Studied Autoencoder;
- Studied the concept of fully connected layer, filters, max-pooling, ReLU, stride and padding in CNN and how to calculate the output volume size;
- Studied batch normalization and the suitable position to insert in CNN; Advantages such as improving gradient flow through the network, allowing higher learning rate and reducing the strong dependence on initialization;
- Understood SGD with Momentum, second-order optimization; Understood model ensembles, dropout, data augmentation to prevent overfiting.

- Understood the concept of transfer learning;
- Learnt different CNN such as AlexN;
- Calculated the loss in RNN Language Model and performed the backpropagation with multivariable chain rule;
- Understood the vanishing and exploding gradient problem in RNN; Studied the architecture of LSTM;
- Understood the mechanism in Attention;
- Knowledge in Graph Convolutional Networks; Knowledge in graph pooling with clustering and its limitations;

Bibliography

A. My First Appendix

In this file (Appendices/Appendix_A.tex) you can add appendix chapters, just as you did in the Document.tex file for the 'normal' chapters. You can also choose to include everything in this single file, whatever you prefer.