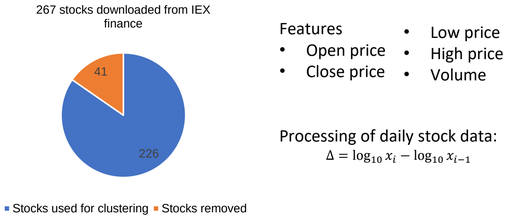
Analyzing public companies in the life sciences and designing an investment portfolio

Rosa Ma, Vandon Duong
{rosaxma, vandon} @ stanford.edu

Abstract

Stocks of public companies in the biotech sector tend to be extremely volatile and largely dissociated from traditional fundamentals (e.g. EPS, P/E ratio). There are numerous obscure factors that drive the valuation of companies in the life sciences. We propose a deep learning-based investment strategy, which includes diversification through clustering and portfolio construction by selecting stocks in each cluster based on Sharpe ratio. Specifically, our clustering model has two phases: (1) parameter initialization with a deep convolutional autoencoder and (2) parameter optimization (i.e., clustering), where we iterate between computing an auxiliary target distribution and minimizing the Kullback-Leibler (KL) divergence to it. Experimental results show the diversification accomplished by our clustering method suggested portfolios that outperform multiple indexes during a downturn in the biotech sector.

Data & Features



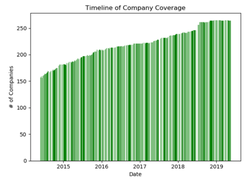
Features
- Open price
- Close price
- Low price
- High price
- Volume

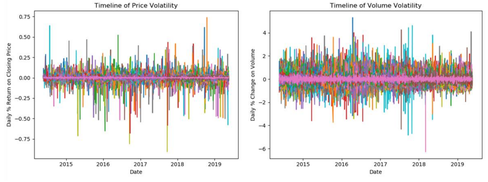
Processing of daily stock data:

$$\Delta = \log_{10} x_i - \log_{10} x_{i-1}$$

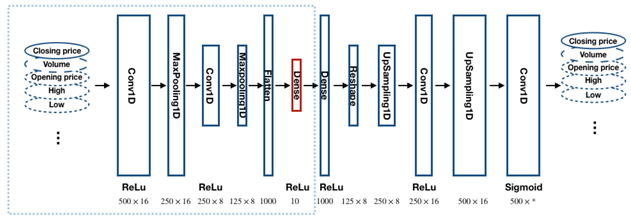
Training and testing paradigm:

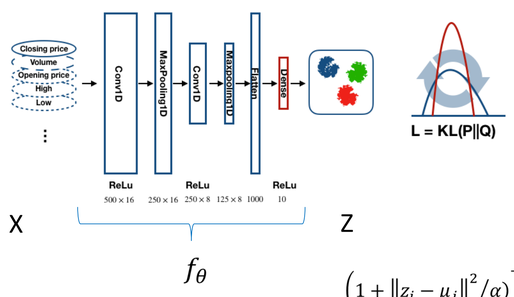
Evaluation of dataset:
- General increase in biotech IPOs over time
- Significant price and volume volatility





Model





f_θ

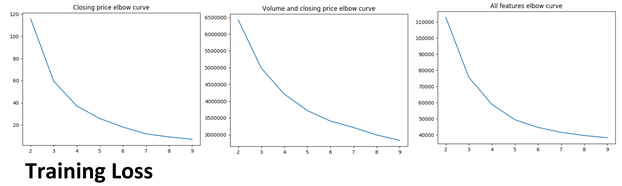
Student's t-distribution: (soft assignment)

$$q_{ij} = \frac{\left(1 + \|z_i - \mu_j\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'}\left(1 + \|z_i - \mu_{j'}\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}$$

Auxiliary target distribution:

$$p_{ij} = \frac{q_{ij}^2/f_j}{\sum_{j'} q_{ij}^2/f_{j'}}$$

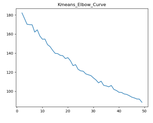
Elbow curves



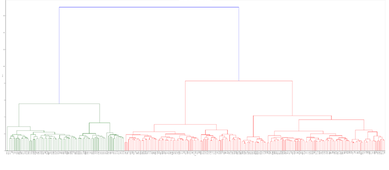
Training Loss

Baseline

K-means clustering
(initialization: K-means ++)

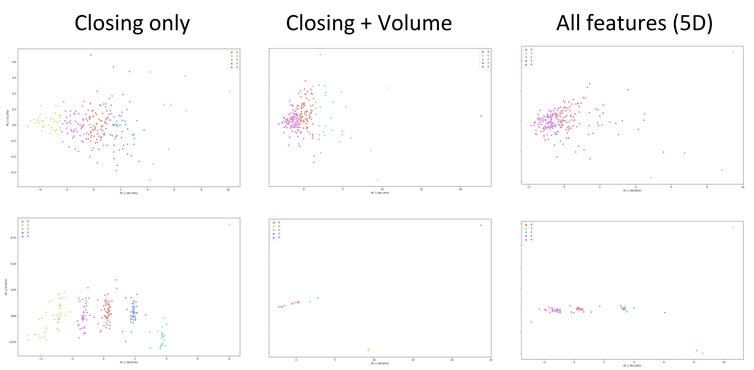


Hierarchical clustering
(linkage: Ward's method)

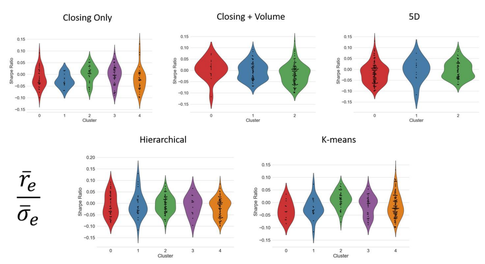


Results

Evaluation of clustering:
- Clustering based on closing & volume is similar with clustering based on all features (possibly redundancy)
- Parameter optimization using Deep embedded clustering (DEC) makes the clusters well-separated



Risk-adjusted performances of clusters during training



Sharpe Ratio $\frac{\bar{r}_e}{\bar{\sigma}_e}$

Returns for portfolios with 1, 3, or 5 holdings

Training: 2016.05.25 - 2018.05.20

Indexes	Return	IRR
SPY	14%	7%
XBI	24%	11%
IBB	7%	4%
UBIO	1%	1%
BBC	25%	12%
CNCR	8%	4%

Close Only	Encoder + DEC		Encoder + K-means	
	Return	IRR	Return	IRR
1 HpC	116%	48%	118%	48%
3 HpC	96%	40%	95%	40%
5 HpC	78%	34%	68%	30%

Close Volume	Encoder + DEC		Encoder + K-means	
	Return	IRR	Return	IRR
1 HpC	130%	52%	103%	43%
3 HpC	135%	54%	104%	43%
5 HpC	118%	48%	95%	40%

Baseline	Hierarchical		K-means	
	Return	IRR	Return	IRR
1 HpC	120%	49%	125%	50%
3 HpC	108%	44%	92%	39%
5 HpC	100%	42%	76%	33%

5D	Encoder + DEC		Encoder + K-means	
	Return	IRR	Return	IRR
1 HpC	130%	52%	103%	43%
3 HpC	116%	47%	104%	43%
5 HpC	106%	44%	88%	37%

Testing: 2018.05.21 - 2019.06.07

Indexes	Return	IRR
SPY	3%	3%
XBI	-6%	-6%
IBB	-2%	-2%
UBIO	-18%	-17%
BBC	-10%	-9%
CNCR	-14%	-14%

Close Only	Encoder + DEC		Encoder + K-means	
	Return	IRR	Return	IRR
1 HpC	-9%	-9%	-12%	-11%
3 HpC	-12%	-12%	-5%	-5%
5 HpC	-5%	-5%	-14%	-14%

Close Volume	Encoder + DEC		Encoder + K-means	
	Return	IRR	Return	IRR
1 HpC	-15%	-14%	-24%	-23%
3 HpC	-10%	-10%	-8%	-8%
5 HpC	-8%	-8%	-10%	-10%

Baseline	Hierarchical		K-means	
	Return	IRR	Return	IRR
1 HpC	-13%	-12%	-18%	-17%
3 HpC	-10%	-10%	-12%	-11%
5 HpC	-9%	-8%	-11%	-10%

5D	Encoder + DEC		Encoder + K-means	
	Return	IRR	Return	IRR
1 HpC	-14%	-14%	-24%	-23%
3 HpC	-10%	-10%	-8%	-8%
5 HpC	-6%	-6%	-10%	-10%

Discussion & Future Directions

- Compared to the general market ($SPY), the biotech sector ($XBI, $IBB, $UBIO, $BBC, $CNCR) had a strong performance during the training period, but significant retraction in the testing period.
- Baseline clustering (hierarchical and K-means) tend to overfit to the training period.
- Using the autoencoder with DEC has a robust portfolio performance across the different clustering models. Clustering based on closing price only is the most effective in mitigating losses.
- The model based on both closing price and volume, and the model based on all features tend to overfit to the training period compared with the model based only on closing price.
- The current clustering method in the custom layer requires pre-specification of the number of clusters and initial node selection. Thus, other methods that do not require pre-specification such as modularity optimization could be incorporated in the future.
- We will interpret the model to understand the nuances of the markets. i.e., what do the features extracted by the model represent. We will also apply the model to other sectors such as automobile, energy, and real estates or even the general market. We expect the model to have similar performance in these markets.

Reference

Thakor, Richard T., et al. "Just how good an investment is the biopharmaceutical sector?." *Nature biotechnology* 35.12 (2017): 1149.
Goetzmann, William N., and Alok Kumar. "Equity portfolio diversification." *Review of Finance* 12.3 (2008): 433-463.
G. Hu *et al.*, "DEEP STOCK REPRESENTATION LEARNING: FROM CANDLESTICK CHARTS TO INVESTMENT DECISIONS."
J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," vol. 48, 2015.