How can machine learning help stock investment?

Xin GUO guoxin@stanford.com

Objective

Build up models to predict the second-day stock price movement trend.

Two movement trends are defined:

1. $\begin{cases} \text{trend} = 1 & \text{if second day close price rises} \\ \text{trend} = -1 & \text{if second day close price falls} \end{cases}$
2. Use K-means classification to classify the price difference between the next trading day and today into two groups.

Data Collection, Cleaning, and Feature Generation

1. The data was downloaded from quandl.com through API. Four stocks are considered

AA	GE	HPQ	IBM

2. Data with incomplete/invalid information is removed.
3. Financial indicators and self-developed indicators are calculated and totally there are **253** features in the feature bag.
4. All the features are normalized as

$$f := \frac{f - mean\ value\ of\ f}{standard\ deivation\ of\ f}$$

5. First 70% and last 30% data is for training and validation, respectively, for cross validation.

Model Qualification Method

Matthews correlation coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$MCC = 1$: A perfect prediction

$MCC = 0$: No better than random prediction

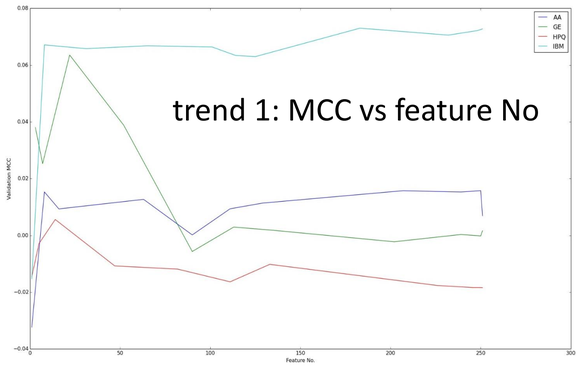
$MCC = -1$: total disagree between prediction and observation

Models

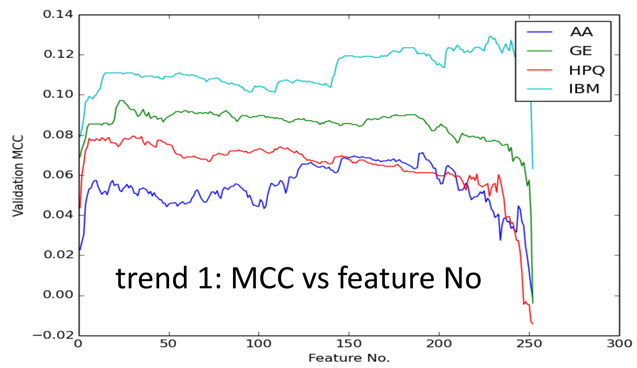
1. Logistical regression (LR)
2. SVM
   a) Linear kernel: $\langle x, x' \rangle$
   b) Polynomial kernel: $(\gamma \langle x, x' \rangle + r)^d$
   c) Rbf kernel: $\exp(-\gamma |x - x'|^2)$
   d) Sigmoid: $\tanh(\gamma \langle x, x' \rangle + r)$
3. Random forest (RF)

Feature Selection

1. Random forest model provides a feature ranking.
   - Models with the top features give the best validation MCC



2. Forward search with logistical regression.
   - Only certain feature combination gives the best validation MCC.



Result

Grid search of all the parameters are performed on all the three models. Models with best validation MCC are given here.

Movement trend 1 (models are chosen based on validation MCC):

	All features		RF feature selection			Forward search (LR)	
	Model	Best MCC	Model	Feat. No.	Best MCC	Feat. No.	Best MCC
AA	RF	0.052	SVM (rbf)	129	0.019	191	0.072
GE	SVM (lin)	0.062	SVM (rbf)	22	0.092	23	0.091
HPQ	LR	0.053	SVM (line)	47	0.053	30	0.071
IBM	LR	0.10	SVM (rbf)	183	0.11	228	0.084

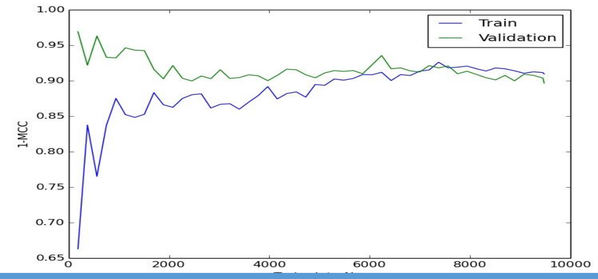
IBM: LR confusion matrix (trend 1)

Predict \Real	-1	1
-1	1276	1077
1	747	960

Movement trend 2 (models are chosen based on validation MCC):

	All features		RF feature selection			Forward search (LR)	
	Model	Best MCC	Model	Feat. No.	Best MCC	Feat. No.	Best MCC
AA	LR	0.051	SVM (sigm)	43	0.079	120	0.11
GE	LR	0.13	LR	43	0.15	91	0.17
HPQ	LR	0.15	LR	149	0.16	36	0.17
IBM	LR	0.14	LR	85	0.10	97	0.15

IBM: LR learning curve (trend 2)



Conclusion and Future Work

1. Models with best validation MCC are built up based on current feature set with the assistant of random forest feature selection and forward search techniques. The best validation MCC is up to 0.17.
2. Stock prediction is quite feature and stock dependent. Different feature subsets and different models are best for different stocks.
3. A good classification of price tendency may help to increase prediction accuracy.
4. For more accurate prediction, more features are needed to provide more useful information.
5. SVM model with very high order polynomial kernel may help prediction.