正则化Logistic回归预测股价收益率

问题介绍

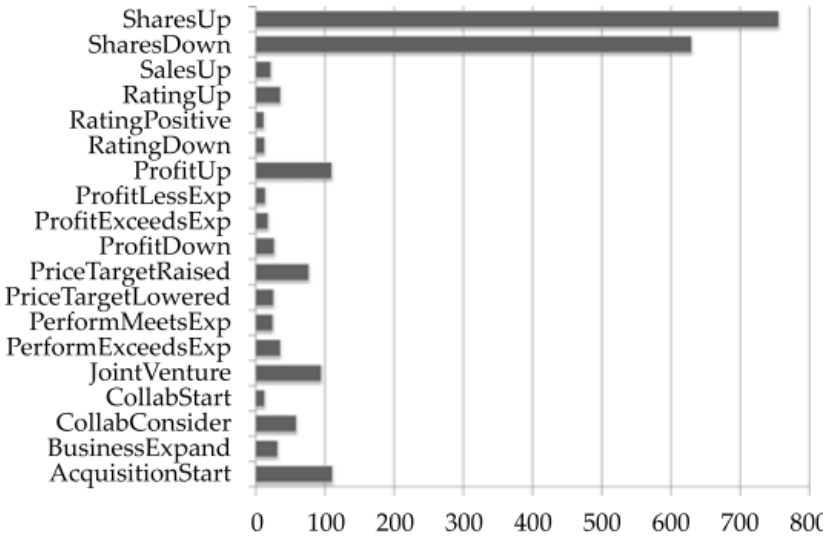
金融市场由信息驱动，得到信息的很重要的一个来源就是新闻。因此我们希望利用新闻里的信息来预测股价的波动。

但新闻的数量巨大并且内容庞杂，因此需要将新闻提取为事件（比如增持、减持等），再通过分析历史数据找到事件发生对股票价格波动的影响。

An Automated Framework for Incorporating News into Stock Trading Strategies

- Relationship between News and Share Prices

News → **Events** → **Impact** ⟷ **Return(R_x, A_x)**



Event	Impact	Freq.	R_0	d	p	R_1	d	p	R_2	d	p	R_5	d	p	R_{10}	d	p
SharesUp	2	756	1.63	85	0.00	1.65	80	0.00	1.53	73	0.00	1.74	72	0.00	2.27	69	0.00
RatingUp	2	36	1.55	83	0.00	1.82	72	0.00	1.89	75	0.00	2.52	69	0.00	2.31	72	0.01
CollabStart	2	13	1.06	46	0.32	1.73	62	0.18	1.95	62	0.14	1.95	77	0.10	1.30	62	0.20
RatingPositive	1	12	0.84	75	0.25	0.96	58	0.32	1.51	75	0.28	3.23	75	0.02	4.26	92	0.02
ProfitExceedsExp	3	18	0.74	50	0.39	1.99	61	0.11	1.72	61	0.19	2.87	56	0.08	2.61	61	0.12
AcquisitionStart	3	111	0.40	62	0.08	0.56	59	0.06	0.53	55	0.12	0.90	59	0.02	0.90	62	0.06
SalesUp	2	22	0.33	64	0.44	0.56	55	0.37	0.35	55	0.65	1.17	64	0.29	1.44	59	0.23
PriceTargetRaised	2	77	0.24	51	0.31	0.54	56	0.13	0.65	58	0.11	0.83	56	0.08	2.35	71	0.00
BusinessExpand	1	32	0.21	53	0.70	0.83	56	0.16	0.69	56	0.34	1.52	66	0.08	2.13	72	0.02
JointVenture	1	95	0.18	53	0.31	0.12	47	0.59	0.10	52	0.72	0.49	56	0.14	0.78	53	0.05
PerformExceedsExp	3	36	0.11	58	0.82	0.35	50	0.52	0.25	47	0.70	1.49	53	0.11	0.58	56	0.53
ProfitUp	2	110	0.08	50	0.80	0.22	50	0.53	0.26	52	0.54	1.25	56	0.03	1.73	64	0.01
CollabConsider	1	59	-0.05	49	0.81	-0.08	54	0.83	-0.17	49	0.69	0.19	59	0.67	0.11	59	0.85
PerformMeetsExp	1	25	-0.21	40	0.66	-0.11	48	0.79	0.31	56	0.57	0.55	64	0.46	0.75	56	0.42
ProfitDown	-2	27	-0.94	48	0.21	-0.52	52	0.49	0.05	48	0.95	0.33	52	0.78	0.83	52	0.56
RatingDown	-2	13	-0.96	54	0.15	-0.98	62	0.14	-1.32	69	0.11	-0.65	54	0.54	0.04	46	0.98
PriceTargetLowered	-2	26	-1.17	73	0.16	-1.44	77	0.12	-1.77	73	0.07	-1.50	65	0.14	-1.51	50	0.20
SharesDown	-2	630	-1.38	81	0.00	-1.49	71	0.00	-1.44	69	0.00	-1.20	62	0.00	-0.91	59	0.00
ProfitLessExp	-3	14	-2.52	64	0.06	-2.33	71	0.04	-2.04	71	0.08	-1.38	57	0.40	-1.29	64	0.35
		2,112		60			60			61			62			62	

- Technical Trading

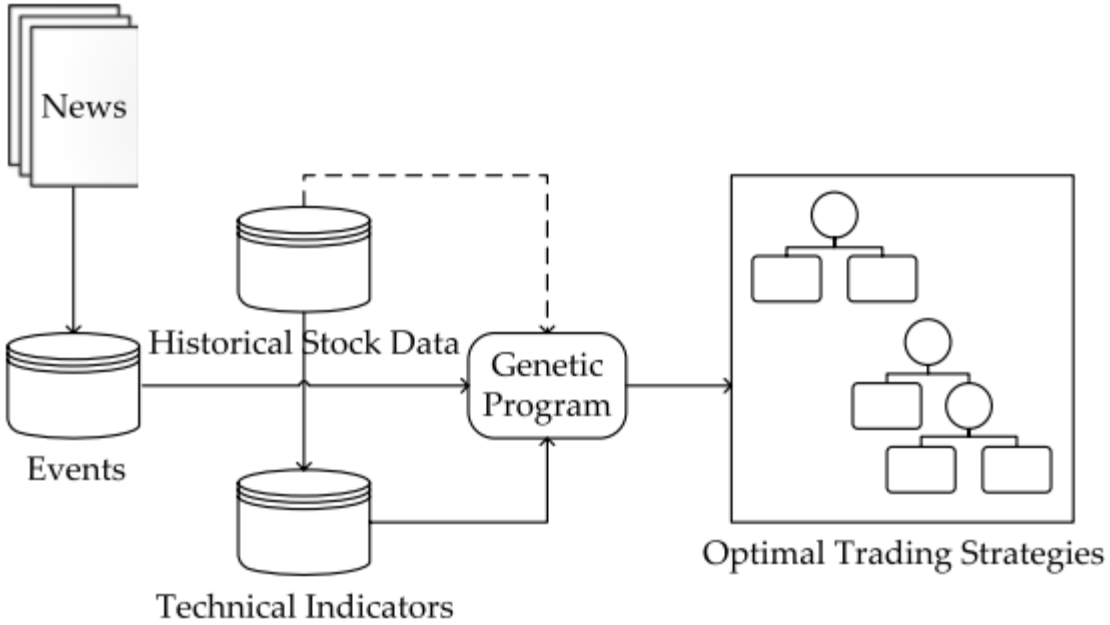
History Data ⟶ Technical Indicator ⟶ Buy/Sell signal ⟷ Return(R_x, A_x)

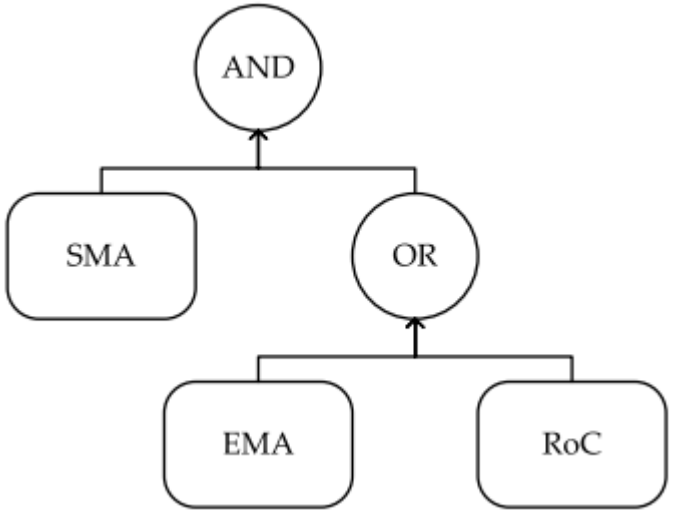
	Buy Signals						Sell signals					
Indicator	Freq.	R_0	R_1	R_2	R_5	R_{10}	Freq.	R_0	R_1	R_2	R_5	R_{10}
SMA(20)	1,663	**1.927**	**2.069**	**2.197**	**2.463**	**3.512**	1,737	**-1.888**	**-1.932**	**-1.923**	**-1.496**	**-0.805**
BB	2,014	-1.608	-1.374	-1.170	**0.110**	**0.180**	2,971	1.563	1.530	1.499	1.595	1.827
EMA(5,20)	870	**1.717**	**1.860**	**1.859**	**2.122**	**3.350**	922	**-1.662**	**-1.654**	**-1.586**	**-0.933**	**-0.488**
RoC(10)	4,387	-0.417	-0.290	-0.245	-0.053	**0.245**	2,937	0.723	0.637	0.953	1.564	2.370
MOM	1,988	**1.499**	**1.444**	**1.637**	**1.938**	**2.759**	2,049	**-1.310**	**-1.151**	**-1.248**	**-1.021**	**-0.629**
MACD(12,26)	667	**1.310**	**1.324**	**1.309**	**1.568**	**2.882**	581	**-1.276**	**-1.106**	**-1.162**	**-0.106**	0.317

- A News-Based Trading Framework

Technical Indicator + News → Trading Rules

Return(R_x, A_x) ← Buy/Sell signal ← Optimized Trading Rule





建模思路

- **找到与股价波动相关的变量作为输入变量**

  事件、宏观经济指标、技术指标、新闻热度以及新闻情分析等

- **量化事件发生后的股价波动作为输出变量**

  平均收益率、平均超额收益率

- **对输入变量以及输出变量建立模型**

  正则化logistic回归、深度学习方法

输入变量

- **宏观经济指标**

  生产者物价指数（PPI）、货币供给M2、制造业采购经理指数（PMI）、居民消费价格指数（CPI）、商品零售价格指数（RPI）、进出口总量等

- **技术指标**

  SMA（20）、EMA（20）、RoC（10）、MACD（12,26）

输出变量

- **事件发生后的收益率(*Ri*)**

$$Y = \begin{cases} 1 & if\ sign(Ri) > 0 \\ 0 & if\ sign(Ri) \leq 0 \end{cases} \qquad i = 0,\ 1,\ 2,\ 5,\ 10$$

- **事件发生后的超额收益率(Ai)**

$$Ri = \alpha + \beta Rim + \varepsilon i \qquad i = 0,\ 1,\ 2,\ 5,\ 10$$

$$Ai = Ri - \alpha - \beta Rim$$

$$Y = \begin{cases} 1 & if\ sign(Ai) > 0 \\ 0 & if\ sign(Ai) \leq 0 \end{cases}$$

L1 Regularized Logistic Regression

- Logistic Regression
- L1 Regularized Logistic Regression

Logistic Regression

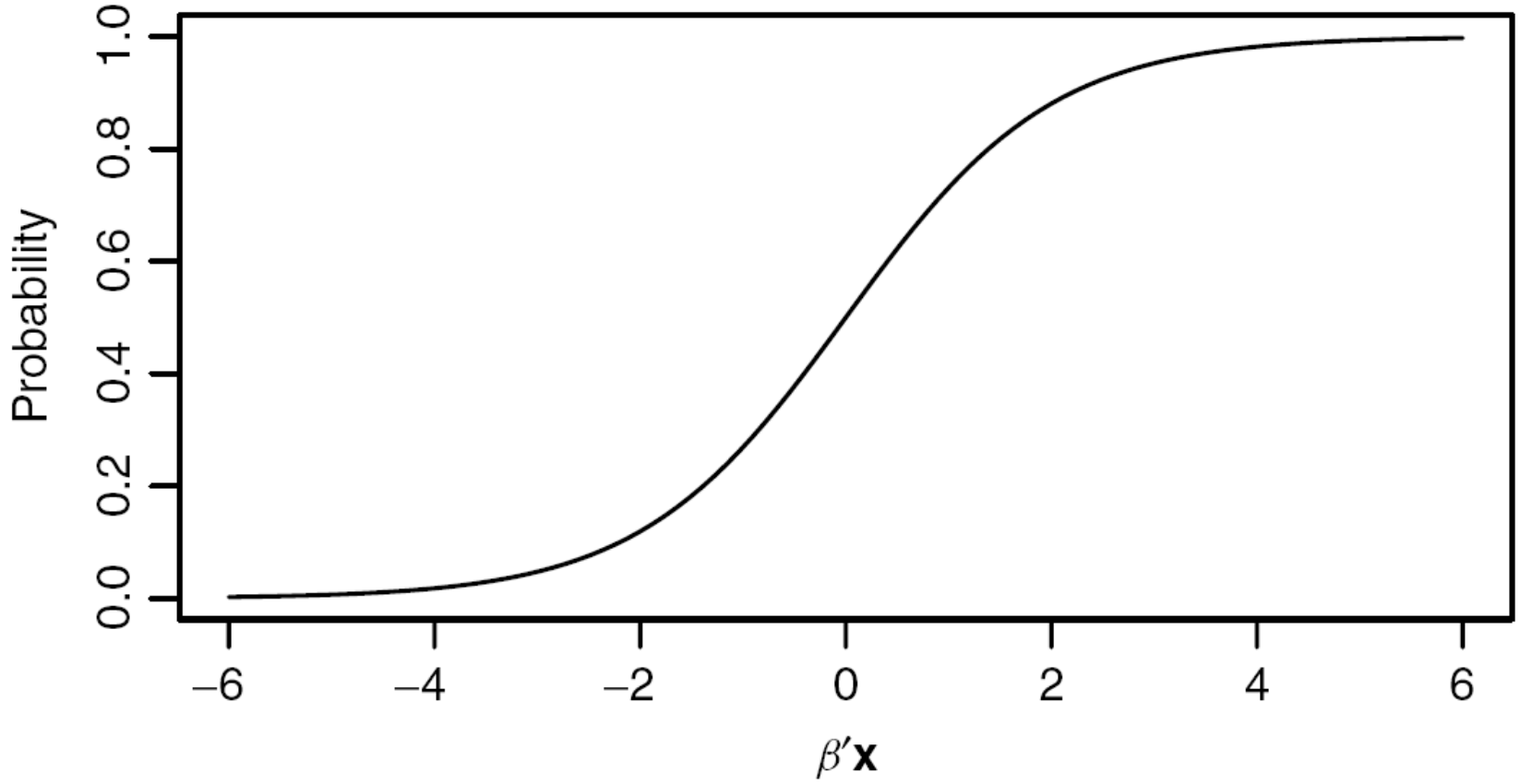
- 多元线性回归模型：$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = \beta_0 + X\beta$

  其中是β_0截距，β是参数向量，X是自变量向量表示n个自变量x与响应变量间的关系，Y为任意实数 ，属于连续变量。

- 当响应变量为离散型变量（0,1）时，为了把y化为二值变量，引入sigmoid函数：

$$g(y) = \frac{1}{1 + e^{-y}}$$

Sigmoid function



基本原理

- $(x_i, y_i)\ i = 1,2,\ldots,N$是N组观测，其中$x_i \in R^p$是输入变量，$y_i \in R$是输出变量并且只能在{0,1}之中取值，$\beta_j \in R\ j = 1,2,\ldots,p$是需要估计的参数。
- 条件分类概率

$$P(y_i = 0|x_i) = \frac{1}{1 + e^{-(\beta_0 + x_i^\top \beta)}}$$

$$P(y_i = 1|x_i) = \frac{1}{1 + e^{(\beta_0 + x_i^\top \beta)}}$$

- Link function

$$ln\ \frac{P(y_i = 0|x_i)}{P(y_i = 1|x_i)} = \beta_0 + x_i^\top \beta$$

例子：一个人在家是否害怕生人来

自变量（x）	不害怕（Y=0）	害怕（Y=1）
文盲（0）	11	7
小学（1）	45	32
中学（2）	664	422
大专以上（3）	168	72

- 用p(x)表示一个人文化程度是x时，害怕生人的概率，考虑模型

$$\ln \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

参数估计

- 对数似然函数

$$\mathcal{L}(\beta_0, \beta) = \sum_{i=1}^{N} [y_i(\beta_0 + x_i^\top \beta) - ln(1 + \exp^{\beta_0 + x_i^\top \beta})]$$

- 极大对数似然函数得到β的估计（通常采用Newton-Raphson迭代法求解）。

L1 Regularized Logistic Regression

- 1996年Tibshirani提出了带一范数的最小二乘问题即Lasso，由于一范数具有筛选变量的能力，所以可以得出一个精简的模型。

$$min\ (y - x\beta)^2 + \lambda||\beta||_1$$

- 比起传统的前向向后法，L1正则化不那么贪婪的选取变量，可以是被视为更光滑和“更民主”的向前逐步选择版本。

- 带一范数的logistic回归实际上是解决这样的问题

$$\min -L(\beta_0, \beta) + \lambda||\beta||_1$$

数值实验
数据处理

- 随机选取了一支股票（code=600000），得到在1999.11-2017.3的历史交易数据。
- 根据收盘价，计算出每天对应的收益率r_i $i = 0,1,2,5,10$ 以及对应的超额收益率a_i $i = 0,1,2,5,10$。
- 根据收盘价，计算出每天对应的技术指标： SMA(20), EMA(20), RoC(10), MACD(12,26)。
- 统计出每个月的交易日个数，计算出每月对应的月平均股价波动率、月平均超额收益率、月平均SMA(20), EMA(20), RoC(10), MACD(12,26)。
- 根据事件发生的日期，得出对应的收益率以及超额收益率以及对应的技术指标

模型描述

- Model 1

将每月的收益率、超额收益率作为输出变量，将月平均技术指标、宏观经济月数据作为输入变量，建立对应的L1 regularized logistic regression，通过CV选出最优正则化参数，并得出模型预测误差。

- Model 2

将与事件对应的收益率、超额收益率作为输出变量，将对应的技术指标作为输入变量，建立对应的L1 regularized logistic regression,通过CV选出最优正则化参数，并得出模型预测误差。

实验结果

Model 1

R1	R2	R5	R10	A1	A2	A5	A10
0.455	0.455	0.415	0.432	0.415	0.432	0.461	0.387

模型总结

- Model 1

  该模型的预测准确率很低，可能有以下几点原因：

  1.在选取输出变量时，直接将收益率的正负取为1,0可能并不合理。

  2.在判断预测目标时，由50%作为分界点，如果样本的刚好选取的时段涨跌不平衡，选取50%就不合理。

  3.通常技术指标都是依据某种规则直接判断涨跌，可能不能直接利用其值进行拟合。

  4.宏观经济指标对股市通常有滞后性，而判断股价的涨跌对时间很敏感。

- Model 2

总结

当前进展：

- 数据

  所有股票的每天的历史交易数据（包括开盘价，收盘价，交易量等）

  所有股票的所有历史公告（文本）

  多个财经网站上的财经新闻（文本）

  上市公司基本情况（还在完善）

- 模型

  Model 1, Model 2得出的结果。

讨论

- 增加哪些输入变量（如上市公司总股本等）？
- 怎样提取一个更完整的事件（如提取增持时可以同时得到增持的比例）？
- 输出变量是选取一个事件对应的收益率还是多个事件的收益率？
- 输入变量中，是否还需要有事件？如果需要用宏观经济指标，至少需要用月数据，那么事件对应的月收益率怎么定义？各种数据具体怎么作为输入变量（如上市公司的基本信息）？
- 由于事件个数较少，将一个行业作为样本进行预测，是否需要加入行业特征，怎样利用？