用贝叶斯线性回归预测比特币价格变化

财通基金　量化投资部

2018 年 7 月 23 日

目标与方法

- 基于过去两年的比特币交易信息预测未来价格变化。
- 用贝叶斯线性回归模型预测，建立虚拟账户评估预测表现。

数据

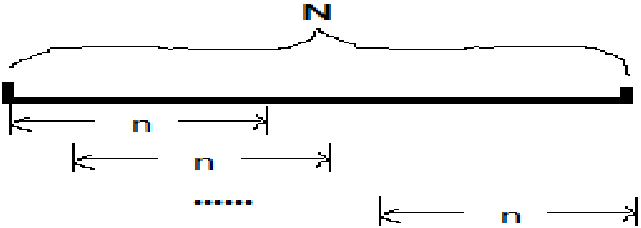
- 数据集1：（目前由于电脑没办法翻墙还没获得）比特币中国交易平台okex.com中2017年1月到2018年6月的价格和定单簿数据

训练集和测试集划分

- 原始价格数据用P表示，包含N个数据，P_i表示P中第i时刻的价格， $i \in [1, N]$
- 将P按时间顺序分为三组长度相同的数据，
  $P_1 = [P_1 : P_{N/3}]$, $P_2 = [P_{(N/3)+1} : P_{2N/3}]$,
  $P_3 = [P_{(2N/3)+1} : P_N]$
- 其中P_1用于生产时间序列及聚类
- P_2用于线性回归拟合系数w_0, w_1, w_2, w_3, w_4
- P_3用于预测比特币价格变动

划分P_1内的数据

- 用长度为n_L， $1 \le L \le 3$的移动窗口
- $n_1 = 180, n_2 = 360, n_3 = 720$生成$H^T, 1 \le T \le 3$, H有$N - n + 1$行， $n + 1$列
- $j \le n$时$H^T(i,j) = P_1[i : i + n + 1]$
- $j = n + 1$时$H^T(i,j) = P_1[i + n + 1] - P_1[1 + n]$



聚类方法

- 将H^t用KMean的方法划分成k个聚类（建议k=100）
- 计算每个聚类中最大值与最小值之差$diff_m$，（$1 \leq m \leq k$）将$diff_m$排序并挑选$diff_m$最大的s个聚类
- 将$diff_m$排序并挑选$diff_m$最大的s个聚类(建议s=20)
- 将每个聚类中心c_q $(1 \leq q \leq s)$作为矩阵的一行，产生一个大小为$s * (n+1)$的矩阵C，C的每一列为聚类中心的一个特征值

贝叶斯线性回归公式

- $E_{emp} = \frac{\sum_{i=1}^{n} y_i * exp(-1/4 * ||x - x_i||_2^2)}{\sum_{i=1}^{n} exp(-1/4 * ||x - x_i||_2^2)}$ [1]

参数意义

- $x = P_1[i - n_L + 1 : i - 1]$
- $x_i = C[i + 1, : n_L + 1]$
- $y_i = C[i + 1, n_L + 1]$

自变量X与应变量Y

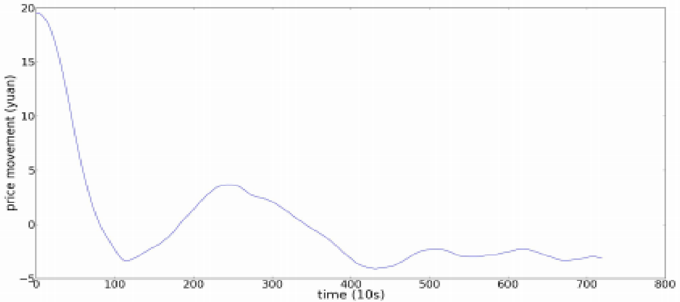
- 自变量X$[\Delta p_1, \Delta p_2, \Delta p_3, r]$，应变量Y$[\Delta p]$
- 分别将$x^L = P_2[i - n_L : i - 1], 1 \leq L \leq 3$, $x_i^L = C_L[i + 1, : n_L + 1]$, $y_i^L = C_L[i + 1, n_L + 1]$代入公式[1]得出$\Delta p_T, 1 \leq T \leq 3$
- $r = \frac{V_{bid} - V_{ask}}{V_{bid} + V_{ask}}$
- $\Delta p = P_2[i + 1] - P_2[i]$

拟合X，Y求系数$w_0 - w_4$

- 用Linear Regression对自变量X和应变量Y进行fit操作，w_0为该线性回归方程与y轴的焦点，$w_1 - w_3$为线性方程回归的系数

根据拟合出的系数$w_0 - w_4$得到$X[\Delta p_1, \Delta p_2, \Delta p_3, r]$和$Y[\Delta p]$之间的线性关系

- $\Delta p = w_0 + \sum_{j=1}^{3} w_j * \Delta p_j + w_4 * r$



公式推导

- $s_1, \ldots, s_K \in \mathbb{R}^d$是$K$个不同的隐藏因子，它们出现的概率为$\mu_1, \ldots, \mu_K$
- $T \in (1, \ldots, K)$是指针，有$P(T = k) = \mu_k$，其中$1 \le k \le K$，$x = s_T + \epsilon$
- $$\begin{aligned} P(y|x) &= \textstyle\sum_{k=1}^{K} P(y|x, T = k)P(T = k|x) \\ &\propto \textstyle\sum_{k=1}^{K} P(y|x, T = k)P(x|T = k)P(T = k) \\ &= \textstyle\sum_{k=1}^{K} P_k(y)P(\epsilon = (x - s_k))\mu_k \\ &= \textstyle\sum_{k=1}^{K} P_k(y)exp(-\frac{1}{2} \parallel x - s_k \parallel_2^2)\mu_k \end{aligned}$$
- 用经验数据作为代替来估计给定x后y的概率分布
- $P_{emp}(y|x) = \frac{\sum_{i=1}^{n} \mathbb{1}(y=y_i)exp(-\frac{1}{4}\|x-x_i\|_2^2)}{\sum_{i=1}^{n} exp(-\frac{1}{4}\|x-x_i\|_2^2)}$

  $\Rightarrow E_{emp}[y|x] = \frac{\sum_{i=1}^{n} y_i exp(-\frac{1}{4}\|x-x_i\|_2^2)}{\sum_{i=1}^{n} exp(-\frac{1}{4}\|x-x_i\|_2^2)}$