رگرسیون بیزی و بیت کوین

علیرضا کشاورز (a. keshavarz@khu. ac. ir)

پروژه یادگیری ماشین-ترم اول ۱۴۰۱

مقاله: http://arxiv.org/pdf/1410.1231.pdf

دانشکده ریاضی و کامپیوتر-دانشگاه خوارزمی

معرفی

این مقاله تحت عنوان رگرسیون بیزی و بیت کوین روش رگرسیون بیزی و اثربخشی آن برای پیش‌بینی تغییرات قیمت بیت‌کوین(یک ارز دیجیتال) را مورد بحث قرار می دهد.

رگرسیون بیزی به استفاده از داده های تجربی برای انجام استنتاج بیزی می پردازد. در بسیاری کاربردها از رگرسیون به عنوان "مدل منبع نهفته" (latent source model) استفاده می شود. اولین بار، رگرسیون بیزی برای مدل منبع نهفته به منظور دسته بندی باینری(binary classification) معرفی و مورد بحث قرار گرفت به این صورت که کارایی تئوری روش به خوبی کارایی تجربی آن برای ایجاد دسته بندی باینری، بررسی شد.

در این مقاله، از آن برای پیش بینی مقدار واقعی قیمت بیت کوین استفاده شده است. بر اساس این روش پیش بینی قیمت، یک استراتژی ساده برای معامله بیت کوین طراحی شده است. این استراتژی می‌تواند سرمایه‌گذاری را در مدت کمتر از ۶۰ روز تقریباً دو برابر کند، در صورتی که بر اساس ردیابی داده‌های واقعی اجرا شود.

I . رگرسیون بیزی

طرح مساله رگرسیون:

به ما n داده برچسب دار آموزشی داده شده است $(x_i , y_i)\text{for } 1 \le i \le n$. هدف مساله استفاده از داده های آموزشی برای پیش بینی برچسب(y) برای داده تست(x) است.

رویکر حل مساله:

K منبع نهفته متمایز $s_1, \dots, s_k \epsilon \mathbb{R}^d$ ، یک توزیع نهفته روی $1, \dots, K$ با احتمالات $\{\mu_1, \dots, \mu_k\}$ و K توزیع احتمال پنهان روی $\mathbb{R}$ که با $P_1, \dots, P_k$ نشان داده شده است، وجود دارد. هر نقطه داده ی برچسب دار (x, y) به صورت زیر تولید می شود:

$$T \in \{1, \dots, K\}$$
$$P(T = k) = \mu_k \quad \text{for } 1 \le k \le K$$
$$x = s\,T + \varepsilon$$

که در آن ε یک متغیر تصادفی مستقل d بعدی است که نویز را نشان می دهد (آن را گوسی فرض می کنیم با میانگین بردار $0 = (0, \dots .0) \in \mathbb{R}^d$ و ماتریس کوواریانس شناسایی y که از $\mathbb{R}$ مطابق با توزیع P_T نمونه برداری می شود.)

با توجه به این مدل، برای پیش‌بینی برچسب y مرتبط با مشاهده x، می‌توانیم از توزیع شرطی y با توجه به x به صورت زیر استفاده کنیم:

$$P(y|x) = \sum_{k=1}^{T} P(y|x, T = k)P(T = k|x)$$

$$\propto \sum_{k=1}^{T} P(y|x, T = k)P(x|T = k)P(T = k)$$

$$= \sum_{k=1}^{T} P_k(y)P(\varepsilon = (x - s_k))\mu_k$$

$$= \sum_{k=1}^{T} P_k(y)\exp\left(-\frac{1}{2}\,||x - s_k||_2{}^2\right)\mu_k$$

(نویسنده، هم ارزی در رابطه بالا را به سادگی از روابط قانون بیز نتیجه گرفته است

(posterior $\propto$ likelikhood $\times$ prior

بنابراین، تحت مدل منبع نهفته، مسئله رگرسیون به یک مسئله استنتاج بیزی بسیار ساده تبدیل می شود. با این حال، مشکل، عدم آگاهی از پارامترهای "مخفی" مدل منبع است. به طور خاص، عدم آگاهی از K منابع $(s_1, \ldots, s_k)$، احتمالات $(\mu_1, \ldots, \mu_k)$ و توزیع های احتمال $P_1, \ldots, P_k$.

برای غلبه بر این چالش، الگوریتم ساده زیر را پیشنهاد شده است:

از داده‌های تجربی برای تخمین توزیع شرطی y با توجه به x در روابط بالا می توان استفاده کرد. به طور خاص، با توجه به n نقطه داده $(x_i, y_i),\ 1 \le i \le n$ احتمال شرطی تجربی برابر خواهد بود با:

$$P_{emp}(y|x) = \frac{\sum_{k=1}^{T} P(y = y_i) \exp(-\frac{1}{4} ||x - x_i||_2^2)}{\sum_{i=1}^{n} \exp(-\frac{1}{4} ||x - x_i||_2^2)}$$

تخمین تجربی پیشنهادی در رابطه بالا بدین صورت است که در دسته‌بندی باینری، y مقادیر $\{0, 1\}$ را می‌گیرد و بر این اساس قانون دسته بندی زیر را پیشنهاد می کند:

$$\text{compute ratio: } \frac{P_{emp}(y = 1|x)}{P_{emp}(y = 0|x)}$$

$$= \frac{\sum_{i=1}^{n} P(y_i = 1) \exp(-\frac{1}{4} ||x - x_i||_2^2)}{\sum_{i=1}^{n} P(y_i = 0) \exp(-\frac{1}{4} ||x - x_i||_2^2)}$$

اگر نسبت < 1 است، $y = 1$، در غیر این صورت $y = 0$ را اعلام کنید. به طور کلی، برای تخمین امید شرطی y، برای مشاهده x، رابطه زیر پیشنهاد شده است:

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

تخمین در رابطه بالا را می توان به طور معادل به عنوان یک تخمین زننده "خطی" مشاهده کرد:

اگر بردار $X(x) \in \mathbb{R}^n$ به صورت زیر

$$X(x)_i = \exp\left(-\frac{1}{4} ||x - x_i||_2^2\right) / Z(x)$$

$$Z(x) = \sum_{i=1}^{n} \exp(-\frac{1}{4} ||x - x_i||_2^2)$$

و $y \in \mathbb{R}^n$ با i امین عنصر y_i باشد، آنگاه

$$\hat{y} \equiv E_{emp}[y|x]$$

$$\hat{y} = X(x)\, y$$

در این مقاله، از رابطه قبل ($\hat{y}$) برای پیش‌بینی تغییرات آینده قیمت بیت‌کوین استفاده شده است.

هدف کلی محقق در این مقاله، علاقه به درک این موضوع است که آیا اطلاعاتی در داده های تاریخی مربوط به بیت کوین وجود دارد که بتواند به پیش بینی تغییرات قیمت در آینده در بیت کوین کمک کند و بنابراین به توسعه استراتژی کمی سودآور با استفاده از بیت کوین کمک کند. همانطور که قبل از این هم بررسی کردیم، از رگرسیون بیزی الهام گرفته از مدل منبع نهفته، استفاده خواهد شد.

II . تجارت بیت کوین

ارتباط مدل منبع پنهان:

یکی از رویکردهای رایج در صنعت مالی، تحلیل تکنیکال است، که فرض می‌کند حرکات قیمت از مجموعه‌ای از الگوها پیروی می‌کند و می‌توان از حرکات قیمت گذشته برای پیش‌بینی بازده‌های آتی تا حدی استفاده کرد. مطالعات نشان دادند که برخی از الگوهای هندسی توسعه یافته تجربی، مانند $heads - and - shoulders$، $triangle$ و $double - top - and - bottom$، می توانند برای پیش بینی تغییرات قیمت در آینده استفاده شوند.

مدل منبع نهفته دقیقاً در تلاش است تا وجود چنین الگوهای زیربنایی را که منجر به تغییرات قیمت می‌شود، مدل‌سازی کند. تلاش برای توسعه الگوها با کمک یک متخصص انسانی یا تلاش برای شناسایی دقیق الگوها در داده ها، می تواند چالش برانگیز و تا حدی ذهنی باشد. در عوض، استفاده از رویکرد رگرسیون بیزی همانطور که قبل

از این هم گفته شد، به ما امکان می دهد از وجود الگوها برای پیش بینی بهتر بدون یافتن صریح آنها استفاده کنیم.

داده:

محقق در مقاله اصلی، برای انجام آزمایش‌ها، از داده‌های مربوط به قیمت و دفتر سفارش به‌دست‌آمده از Okcoin.com که یکی از بزرگترین صرافی‌های فعال در چین است، استفاده کرده است. داده های استفاده شده مربوط به دوره زمانی بین فوریه ۲۰۱۴ تا ژوئیه ۲۰۱۴ است. کل نقاط داده ی خام بیش از ۲۰۰ میلیون بوده است. اطلاعات دفترچه سفارش، شامل ۶۰ بهترین قیمت است که در یک زمان معین مایل به خرید یا فروش است. (نقاط داده در فاصله زمانی هر دو ثانیه به دست آمده است)

به منظور سهولت محاسباتی، یک سری زمانی جدید با فاصله زمانی ۱۰ ثانیه ساخته شده است. هر یک از نقاط داده ی خام به نزدیکترین نقطه در ۱۰ ثانیه آینده نقشه برداری شده است. با افزایش بازه زمانی، "خطای" جزئی در دقت مدل به وجود خواهد آمد، و از آنجایی که استراتژی معاملاتی ما در مقیاس زمانی بزرگتر عمل می کند، این خطا ناچیز خواهد بود.

استراتژی تجارت و پیش بینی تغییر قیمت:

روش اصلی برای میانگین تغییر قیمت (Δp) در بازه ی ۱۰ ثانیه‌ای، رگرسیون بیزی است.

روش بیان شده در مقاله بدین صورت است که از یک سری زمانی تاریخ دار از تغییرات بیت کوین استفاده می کند. این سری زمانی از تجمیع بازه های زمانی ۱۰ ثانیه ای در چند ماه بدست آمده است. با تجمیع این بازه های زمانی یک سری زمانی(یک بردار) بسیار بزرگ بدست می آید. از این سری زمانی استفاده می شود و از آن سه زیر مجموعه داده سری زمانی با سه طول مختلف تولید می کنیم:

S_1 با طول زمانی ۳۰ دقیقه، S_2 با طول زمانی ۶۰ دقیقه و S_3 با طول زمانی ۱۲۰ دقیقه.

اکنون در یک نقطه زمانی معین، برای پیش‌بینی تغییر آتی Δp، از داده‌های تاریخی سه طول زیر استفاده می‌کنیم: ۳۰ دقیقه قبل، ۶۰ دقیقه قبل و ۱۲۰ دقیقه قبل که به ترتیب با x^1, x^2, x^3 نشان داده می شوند.

پیش بینی میانگین تغییر قیمت با استفاده از رگرسیون بیزی انجام می شود. در همین راستا، رگرسیون بیزی از x^j با نمونه های تاریخی S^j برای Δp^j ($1 \le j \le 3$) و

$r = (v_{bid} - v_{ask})/(v_{bid} + v_{ask})$ استفاده می کند (v_{bid} کل حجمی است که افراد مایل به خرید در ۶۰ سفارش برتر هستند و v_{ask} کل حجمی است که افراد مایل به فروش در۶۰ سفارش برتر بر اساس داده های دفتر سفارش فعلی هستند). تخمین نهایی Δp به صورت تولید می شود:

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

$w = (w_0, \ldots, w_4)$ نرخ های یادگیری هستند.

اکنون در مورد یافتن S_j ، $1 \le j \le 3$ و یادگیری.

برای یافتن S_j ، $1 \le j \le 3$ و نرخ های یادگیری w ، مدت زمان را باید به سه دوره تقریباً مساوی تقسیم کرد. از اولین دوره زمانی برای یافتن الگوهای S_j استفاده می شود. دوره دوم برای یادگیری پارامترهای w و دوره سوم آخر برای ارزیابی عملکرد الگوریتم استفاده می شود. یادگیری w به سادگی با یافتن بهترین تناسب خطی روی همه گزینه ها با توجه به انتخاب S_j، $1 \le j \le 3$ انجام می شود.

انتخاب S_j ، برای این کار، تمام سری های زمانی ممکن با طول مناسب را باید انتخاب شوند(برای تاثیر بیشتر بردارهای با ابعاد ۱۸۰، ۳۶۰ و ۷۲۰ به ترتیب برای S_1 ، S_2، S_3). هر کدام از این x_i ها برچسب مربوط به آن(y_i) با مشاهده میانگین تغییر قیمت در بازه زمانی ۱۰ ثانیه

پس از پایان مدت زمان x_i محاسبه می‌شود. این مخزن داده بسیار بزرگ است. برای تسهیل محاسبات روی ماشین تکی با رم ۱۲۸G و ۳۲ هسته، الگوها در ۱۰۰ خوشه با استفاده از الگوریتم $k - means$ خوشه‌بندی شده است. از بین این ها،۲۰ خوشه موثر انتخاب و الگوهای نماینده از این خوشه ها برداشته شده است.

پس از محاسبه Δp باید استراتژی معاملاتی که بسیار ساده است را توضیح داد: در هر زمان، ما موقعیت ۱+ بیت کوین، ۰ بیت کوین یا ۱- بیت کوین را حفظ می کنیم. در هر نمونه زمانی، میانگین حرکت قیمت را در بازه ۱۰ ثانیه پیش‌بینی می‌کنیم. اگر $\Delta p > t$ (t یک آستانه است) و همچنین موقعیت فعلی بیت‌کوین کمتر یا مساوی ۰ باشد، یک بیت‌کوین می‌خریم. اگر $\Delta p < -t$ و همچنین موقعیت فعلی بیشتر یا مساوی ۰ باشد، یک بیت کوین می فروشیم. و در غیر این صورت هیچ کاری انجام نمی دهیم.

III . نتیجه گیری

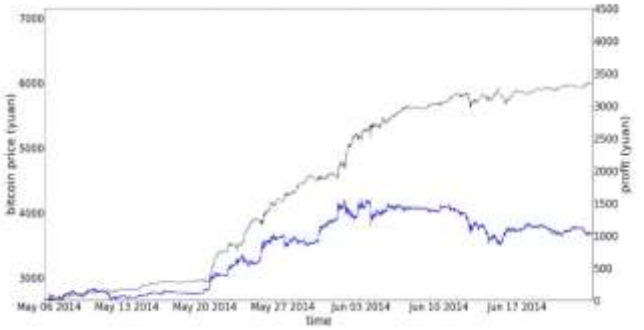
در مقاله اصلی، استراتژی معاملاتی شرح داده شده بر روی یک سوم کل داده ها در مدت زمان ۶ می ۲۰۱۴ تا ۲۴ ژوئن ۲۰۱۴ به صورت علنی شبیه سازی شده است تا عملکرد استراتژی مورد بررسی قرار بگیرد. داده های آموزشی استفاده شده تماماً تاریخی هستند (یعنی قبل از ۶ می ۲۰۱۴ جمع آوری شده اند).

در این بررسی از آستانه های t مختلف استفاده شده است تا تغییرات عملکرد استراتژی مشاهده شود. بررسی ها نشان داده است که آستانه های مختلف عملکرد های مختلفی به همراه دارند. به طور مشخص، با افزایش آستانه، تعداد معاملات کاهش می یابد و میانگین زمان نگهداری افزایش می یابد. در عین حال، میانگین سود هر معامله نیز افزایش پیدا می کند.

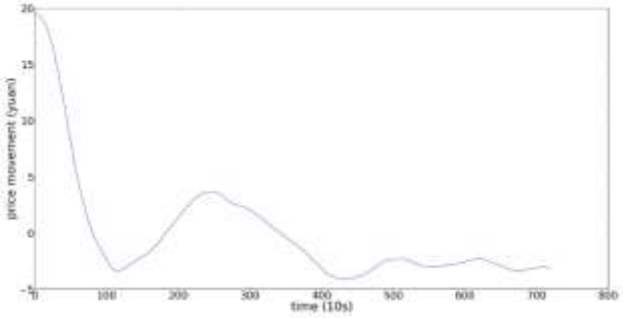
بر اساس شواهد، مشخص است که سود کل در ۳۳۶۲ یوان با ۲۸۷۲ معامله به طور کلی، با میانگین سرمایه گذاری ۳۷۸۱ یوان، به اوج خود خواهد رسید.

شکل زیر عملکرد بهترین استراتژی را در طول زمان نشان می دهد. قابل ذکر است، زمانی که نوسانات بازار بالا است، استراتژی در بخش میانی بهتر عمل می کند.

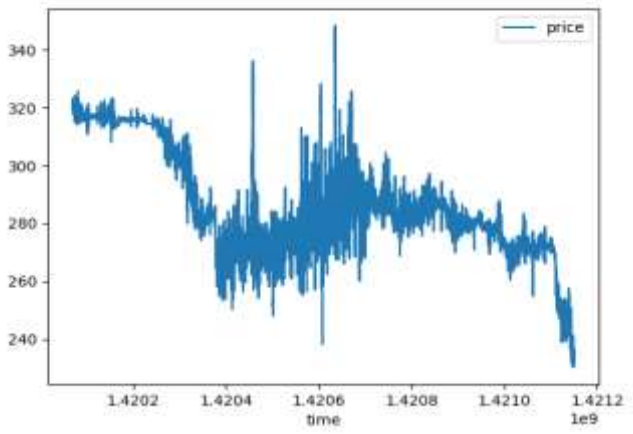
علاوه بر این، استراتژی همچنان سودآور است حتی زمانی که قیمت در آخرین بخش از دوره آزمایش کاهش می یابد.



IV . مقایسه نتایج



نمودار موجود در مقاله که در آن نقاط داده های بدست آمده بر حسب میانگین تغییر قیمت بیت کوین، با استفاده از دو پارامتر زمان و قیمت رسم شده است.



نمودار بدست آمده پس از اجرای کد و گرفتن جواب که نشاندهنده میزان تغییرات قیمت داده ها بر حسب زمان و قیمت بیت کوین است. لازم به ذکر نمودار دوم در ابعاد بزرگتری رسم شده است اما از مقایسه نمودارها با یکدیگر می توان متوجه شد که نقاط داده ها پس از محاسبه میانگین تغییر قیمت و با افزایش زمان، رفته رفته قیمتشان کم شده است (می توان این نتیجه را از هر دو نمودار دریافت).