پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

ارائه و بررسي ميزان كارآمد بودن الگوريتمی در كاهش اثر ناخواسته ايجاد شده از واريانس خطا بر مدل سازی

نگارنده

رکسانا درویشی

استاد راهنما: دکتر علی کمالی‌نژاد

پایان‌نامه برای دریافت درجه کارشناسی
در رشته‌ی آمار

تابستان ۱۴۰۱

چکیده

در این نوشته چند الگوریتم معرفی می‌شوند که ما را در تحلیل داده‌ها و مدل سازی یاری می‌دهند. با استفاده از این الگوریتم‌ها می‌توانیم نوع توزیع خطا را حدس زده و پارامترها و تابع توزیع احتمال آن را برآورد کنیم. به علاوه، امکان بررسی ویژگی‌های خطا مانند پراکندگی، تقارن و چولگی فراهم می‌شود. در ادامه، راهکارهایی برای کاهش واریانس خطا پیشنهاد می‌شود که می‌توانند در مراحل اولیه مدل سازی، آنالیزگر را به سمت تشخیص صحیح رابطه‌ی میان متغیرهای پیشگو و پاسخ هدایت کنند.

سپاسگزاری

سپاسگزاری

با سپاس از استاد گرانقدر دکتر علی کمالی‌نژاد که با راهنمایی‌های خود اینجانب را در تدوین و نگارش این پروژه یاری نمودند.

فهرست تصاویر

پیشگفتار

بخش قابل توجهی از چالش‌هایی که در تحلیل داده و مدل سازی آماری مطرح می‌شوند، در وجود متغیر تصادفی خطا و اثر ناخواسته‌ی آن ریشه دارند. در بسیاری از موارد حتی نوع توزیع خطا و ویژگی‌های آن مشخص نیست. بنابراین، امکان تشکیل تابع درستنمایی وجود ندارد. به علاوه، برای استفاده از برخی روش‌های مدل سازی لازم است فرضیاتی را حول توزیع خطا بپذیریم. برای آن که از برقرار بودن این فرضیات اطمینان حاصل کنیم، نیاز داریم نوع توزیع خطا و بعضی از ویژگی‌های آن را بدانیم یا بتوانیم تخمین بزنیم.

یکی از اولین مراحل مدل سازی انتخاب گزینه‌های مناسب برای رابطه‌ی میان متغیرهاست. سپس، آنالیزگر تلاش می‌کند تا با داده‌های موجود، بهترین تابع را انتخاب کند و شکل کلی آن را تقریب بزند. در صورتی که واریانس خطا بزرگ باشد، تشخیص رابطه‌ی میان متغیرهای با داشتن نمودارهای پراکنش دشوار خواهد بود.

در این نوشته سه الگوریتم که اساس همه‌ی آن‌ها تقریبا مشابه است، معرفی می‌شوند. با به کارگیری آن‌ها می‌توانیم تابع توزیع احتمال خطا را تخمین بزنیم. با داشتن این تابع، شکل کلی چگالی خطا مشخص شده و می‌توانیم پارامترهای آن را برآورد کنیم. از این رو، با به کارگیری آزمون‌های ناپارامتری برای سنجش هم‌توزیعی، توزیع آن را حدس می‌زنیم.

یکی دیگر از کاربرهای این الگوریتم‌ها پیدا کردن دور ترین نقاط نسبت به میانگین شرطی متغیر پاسخ است. آنالیزگر می‌تواند با جابه‌جایی یا حذف این نقاط دقت مدل سازی را افزایش دهد. داده‌های پرت نیز از این روش قابل تشخیص خواهند بود.

نتیجه‌ی اجرای الگوریتم بر داده‌های شبیه سازی شده و کارایی آن‌ها در شرایط متفاوت بررسی گردیده است. سپس، تنظیماتی که میزان دقت و حساسیت الگوریتم را تعیین می‌کنند معرفی و تاثیر آن‌ها بر خروجی شرح داده شده است.

فهرست مطالب

فصل ۱

مفاهیم مقدماتی

۱.۱ تعاریف

۱.۱.۱ آزمایش تصادفی

آزمایشی که در شرایط یکسان بتوان آن را تکرار کرد و نتیجه‌ی آن از قبل از انجام آزمایش تعیین شدنی نبوده ولی همه ی نتایج ممکن آن قابل تعیین باشد، یک آزمایش تصادفی گویند.

۲.۱.۱ فضای نمونه

مجموعه‌ی تمام نتایج یک آزمایش تصادفی را فضای نمونه گویند و آن را با نماد S نمایش می‌دهند.

۳.۱.۱ متغیرتصادفی

تابعی حقیقی است که به هر پیشامد در آزمایش تصادفی مورد نظر، عددی حقیقی را نظیر می‌کند. بنابراین، دامنه‌ی آن فضای نمونه خواهد بود.

۴.۱.۱ متغیر پیشگو، کمکی، ورودی یا مستقل

متغیر یا متغیرهای پیشگو که با بردار X نمایش داده می‌شوند، ورودی‌های مدل آماری هستند که برای توضیح دادن تغییرات متغیر پاسخ و یا پیشبینی مقدار آن استفاده می‌شوند. استقلال این متغیرها

از فرض‌های مهم مدل آماری است.

۵.۱.۱ متغیر پاسخ یا وابسته

متغیر مورد علاقه جهت پیشبینی و مدلسازی که مقدار آن وابسته به متغیرهای پیشگو است، با Y نشان داده می‌شود.

۶.۱.۱ امید ریاضی

فرض کنید X متغیری تصادفی با تابع احتمال یا تابع چگالی احتمال $f_X(x)$ باشد. امید ریاضی X یا میانگین X به صورت زیر تعریف می شود :

اگر X گسسته باشد: $$\mu = E(X) = \sum_x x f_X(x)$$

اگر X پیوسته باشد: $$\mu = E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$$

در صورتی که مجموع یا انتگرال بالا همگرا نباشد ، گوییم امید ریاضی X وجود ندارد.

۷.۱.۱ آماره

هر تابع از نمونه‌ی تصادفی مانند $T(X) = T(X_1, \ldots, X_n)$ را یک آماره گویند.

۸.۱.۱ برآورد نقطه‌ای

در این روش از روی مقدار مشاهده شده آماره یعنی $t = T(x_1, \ldots, x_n)$ تنها یک مقدار برای تخمین پارامتر نامعلوم ارائه می گردد.

۹.۱.۱ برآورد فاصله‌ای

در این روش با استفاده از مقدار مشاهده شده‌ی آماره، فاصله‌ای را با یک اطمینان بسیار خوب به عنوان تخمین پارامتر نامعلوم ارائه می‌دهند.

۱۰.۱.۱ ناریبی

برآوردگر $T = T(X_1, \ldots, X_n)$ را برآوردگری ناریب برای پارامتر θ گویند هرگاه برای هر θ در فضای پارامتری آن

$$E(T) = \theta$$

.

اگر برآوردگر T برای θ ناریب نباشد آن را برآوردی اریب گویند و اریبی آن به صورت زیر تعریف می شود .

$$bias(T) = E(T) - \theta$$

۱۱.۱.۱ آزمون فرض‌های آماری

در این روش بر اساس مشاهده‌ها در مورد صحت یا عدم صحت ادعایی که در مورد جامعه یا پارامترهای آن انجام شده است، قضاوت می کنیم .

۱۲.۱.۱ توزیع‌های تبدیل یافته

فرض کنید متغیر تصادفی پیوسته X با میانگین μ و واریانس σ^2 دارای توزیع D باشد. در ازای ثابت‌های دلخواه $\mu' \in \mathbb{R}$ و $\sigma' \in \mathbb{R}^+$ متغیر تصادفی Z را به صورت ترکیب خطی از X به صورت $Z = \sigma' X + \mu'$ تعریف می‌کنیم. داریم:

$$F_Z(z) = F_X(\frac{z - \mu'}{\sigma'})$$

بنابراین، تابع چگالی Z با داشتن f_D یا F_D قابل محاسبه است. در این صورت می‌گوییم متغیر تصادفی Z دارای توزیع تبدیل یافته D^t با تغییر مکان به اندازه μ' و تغییر مقیاس به مقدار σ' است.
در بسیاری از موارد، تبدیل خطی نوع توزیع را تغییر نمی‌دهد و تنها پارامترهای آن عوض می‌شوند. توزیع‌های نرمال، tـاستیودنت و یکنواخت از این دسته هستند.
در توزیع‌هایی که بر دامنه‌ی تابع چگالی محدودیت دارند، مانند توزیع نمایی و بتا که به ترتیب مقادیر مثبت و در بازه‌ی بسته $[۰, ۱]$ را به عنوان ورودی اتخاذ می‌کنند، یک تبدیل خطی می‌تواند شروط دامنه را بر هم زند و متغیر تصادفی جدید دیگر از توزیع قبلی پیروی نخواهد کرد. اما شکل تابع چگالی آن مشابه توزیع اصلی و با جابه‌جایی μ' و تغییر مقیاس به مقدار σ' است.

دو مثال زیر این مطلب را واضح‌تر می‌کنند :

مثال ۱.۱.

$$X \sim N(\mu, \sigma^۲) \Rightarrow M_X(t) = \exp(\mu t + \frac{۱}{۲}\sigma^۲ t^۲)$$

در اینجا M تابع مولد گشتاور است.

$$Z = \sigma' X + \mu' \Rightarrow M_Z(t') = E(e^{t'Z}) = E(e^{\sigma' X + \mu'})$$

با تغییر متغیر $u = t'\mu'$ و با استفاده از خطی بودن امید داریم:

$$M_Z(t') = e^{t'\mu'} E(e^{uX}) = \exp((t' + u\mu) + \frac{۱}{۲}\sigma^۲ u^۲) =$$

$$\exp(t'(\mu + \sigma'\mu') + \frac{۱}{۲}(\sigma\sigma')^۲ t'^۲)$$

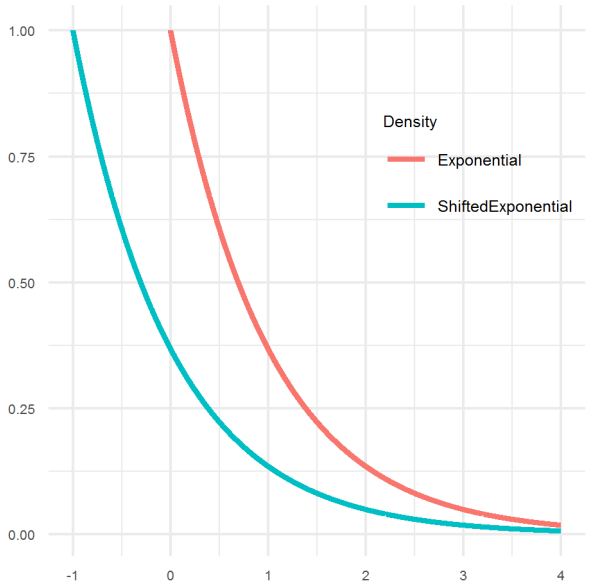
بنابراین Z و X هر دو دارای توزیع نرمال با پارامترهای متفاوت هستند.

مثال ۲.۱.

$$X \sim E(\lambda) \Rightarrow f_X(x) = \lambda e^{-\lambda x}, \quad x > ۰, \lambda > ۰$$

فرض کنید $\mu' < ۰$ و قرار دهید $Z = X + \mu'$. از آنجایی که چگالی Z در ازای مقادیر در بازه‌ی $]\mu', ۰[$ مثبت است، توزیع آن نمی‌تواند نمایی باشد اما چگالی Z به راحتی از f_X به دست می‌آید و با انتقال به اندازه μ' در جهت مثبت محور x بر f_X منطبق می‌شود.

$$f_Z(z) = \lambda e^{-(z-\mu')}$$



شکل ۱.۱: چگالی Z انتقال یافته چگالی X است.

تذکر ۱.۳. *باید به تمایز میان تعریف توزیع‌های تبدیل یافته و خانواده‌ی توزیع‌های مکان‌ـ مقیاس توجه شود.*

۲.۱ داده‌ها

اگر یک متغیر پاسخ y و p متغیر پیشگو داشته باشیم، n مشاهده‌ی ثبت شده به صورت

$$(x_1, y_1), \ldots, (x_n, y_n)$$

خواهند بود که $x_i = (x_{i1}, \ldots, x_{ip})$.

تذکر ۱.۴. *متغیر تصادفی با حروف بزرگ و مقدار مشاهده شده‌ی آن با حروف کوچک نمایش داده می‌شود.*

رابطه میان متغیرها به صورت زیر است :

$$Y|X = g(X) + e, \quad E(Y|X) = g(X) \quad (۱)$$

e متغیر تصادفی منسوب به خطا بوده و پس از جمع آوری داده مقدار آن در ازای مشاهده‌ی iام برابر e_i است. e دارای توزیع D با میانگین ۰ و واریانس $\sigma^۲$ می‌باشد.
از (۱) نتیجه می‌شود که

$$\mathrm{Var}(Y|X) = \mathrm{Var}(g(X)) + \mathrm{Var}(e) = \sigma^۲ \quad (۲)$$

قضیه ۵.۱. *$Y|X = x$ دارای توزیع تبدیل یافته‌ی D با انتقال و تغییر مقیاس به اندازه‌های به ترتیب $g(x)$ و ۱ است. یعنی، $Y|X = x \sim D(g(x), \sigma^۲)$*

اثبات. از خطی بودن رابطه (۱) و تعریف متغیرهای تبدیل یافته نتیجه می‌شود. □

۳.۱ تعاریف و قضیه‌هایی در آنالیز ریاضی

۱.۳.۱ نقطه‌ی حدی

نقطه ی p یک نقطه‌ی حدی مجموعه‌ی E است هرگاه هر همسایگی p شامل نقطه‌ای چون $q \in E$ غیر از p باشد.

۲.۳.۱ مجموعه‌ی بسته

E بسته است هرگاه هر نقطه‌ی حدی E یک نقطه از E باشد.

۳.۳.۱ نقطه‌ی درونی

نقطه p یک نقطه‌ی درونی E است هرگاه یک همسایگی از p مانند N باشد به طوری که $N \subset E$.

۴.۳.۱ مجموعه‌ی باز

E باز است هرگاه هر نقطه‌ی E یک نقطه ی درونی‌اش باشد.

تعریف ۶.۱. *منظور از یک پوشش باز مجموعه‌ی E در فضای متری X یعنی گردآیه‌ای از زیرمجموعه‌های باز X مانند $\{G_\alpha\}$ که $E \subset \cup_\alpha G_\alpha$.*

تعریف ۷.۱. *زیرمجموعه‌ی K از فضای متری X را فشرده نامند هرگاه هر پوشش باز K حاوی زیرپوششی متناهی باشد .*

قضیه ۸.۱. *زیرمجموعه‌های بسته‌ی فضاهای متری فشرده، فشرده‌اند.*

قضیه ۹.۱. *قضیه‌ی هاینه‌ـ بورل:* *هرگاه مجموعه ی E در $\mathbb{R}^k$ یکی از سه خاصیت زیر را داشته باشد، آنگاه از دو خاصیت دیگر نیز بهره مند است :*

(آ) E بسته و کراندار است ؛

(ب) E فشرده است ؛

(پ) هر زیرمجموعه نامتناهی E یک نقطه ی حدی در E دارد .

۵.۳.۱ پیوستگی

فرض کنید X و Y فضاهایی متری بوده، $E \subset X$، $p \in E$، و f مجموعه‌ی E را به توی Y بنگارد. در این صورت، f را در p پیوسته نامند هرگاه برای هر $\epsilon > ۰$، $\delta > ۰$ای باشد بقسمی که به ازای تمام نقاط $x \in E$ که $d_x(x,p) < \delta$ داشته باشیم

$$d_Y(f(x), f(p)) < \epsilon.$$

هرگاه f در هر نقطه‌ی E پیوسته باشد، f بر E پیوسته خوانده خواهد شد.

در اینجا d تابع فاصله یا متری است که بر فضای متری تعریف می‌شود. در این نوشته با توجه به آن که متغیرهای تصادفی متناظر با داده‌ها عموما حقیقی هستند، d فاصله‌ی اقلیدسی در $\mathbb{R}^k$ تعریف می‌شود.

قضیه ۱۰.۱. *فرض کنید f یک نگاشت پیوسته از فضای متری و فشرده X بتوی فضای متری Y باشد ، در این صورت ، $f(X)$ فشرده خواهد بود .*

قضیه ۱۱.۱. *هرگاه f یک نگاشت پیوسته از فضای متری فشرده X به توی $\mathbb{R}^k$ باشد، آنگاه $f(X)$ بسته و کراندار است . لذا f کراندار خواهد بود.*

قضیه ۱۲.۱. *فرض کنید f یک تابع حقیقی پیوسته بر فضای متری و فشرده X باشد، و*

$$M = \sup_{p \in X} f(p), \quad m = \inf_{p \in X} f(p).$$

در این صورت، نقاطی مانند $p, q \in X$ هستند به طوری که $f(p) = M$ و $f(q) = m$.

نتیجه‌ای از این قضیه به صورت زیر نیز قابل بیان است :
نقاطی مانند p و q در X وجود دارند بطوری که به ازای هر $x \in X$، $f(q) \leq f(x) \leq f(p)$؛
یعنی، f به ماکزیمم خود (در p) و به مینیمم خود (در q) می رسد.

۶.۳.۱ پیوستگی یکنواخت

فرض کنید f یک نگاشت از فضای متری X بتوی فضای متری Y باشد، می‌گوییم f بر X به طور یکنواخت پیوسته است هرگاه به ازای هر $\epsilon > ۰$، δی مثبتی باشد بطوری که به ازای هر p و هر q در X که $d_X(p, q) < \delta$ داشته باشیم

$$d_Y(f(p), f(q)) < \epsilon.$$

قضیه ۱۳.۱. *فرض کنیم f یک نگاشت پیوسته از فضای متری و فشرده X بتوی فضای متری Y باشد. در این صورت، f بر X به طور یکنواخت پیوسته است .*

قضیه ۱۴.۱. *فرض کنیم E یک مجموعه‌ی نافشرده در $\mathbb{R}^1$ باشد. در این صورت،*
(آ) تابعی پیوسته بر E هست که ماکزیمم ندارد.
(ب) تابعی پیوسته و کراندار بر E هست که ماکزیمم ندارد.
هرگاه، علاوه بر این، E کراندار هم باشد، آنگاه
(پ) تابع پیوسته‌ای بر E هست که به طور یکنواخت پیوسته نمی‌باشد.

فصل ۲

معرفی مسئله

۱.۲ چند چالش در مدل‌سازی آماری

اگر خطا وجود نداشته باشد (یا واریانس خطا صفر باشد)، طبق عبارت (۱) رابطه‌ی X و Y به صورت $Y|X = g(x)$ خواهد بود.بنابراین، اگر بخواهیم در نقطه‌ی دلخواه $x_. \in S_X$ مقدار Y را ثبت کنیم، با احتمال ۱ این مقدار برابر $g(x_.)$ خواهد بود. در این صورت با خطای ۰ درصد مقدار g در نقطه‌ی $x_.$ با مقدار مشاهده شده‌ی $g(x_.)$ تخمین زده می‌شود.
حتی در این صورت تخمین فرمول دقیق تابع g دشوار خواهد بود و تنها می‌توان به تعداد نقاط متمایز مجموعه‌ی $\{x_1, \ldots, x_n\} \subseteq S_X$ در مورد مقدار دقیق تابع با دقت کامل اظهار نظر کرد. در صورتی که بدانیم g یک چندجمله‌ای از درجه حداکثر n است و تعداد نقاط متمایز $\{x_1, \ldots, x_n\}$ حداقل $n+1$ باشد، می‌توان فرم دقیق تابع را محاسبه کرد. البته که چنین مفروضاتی همیشه برقرار نیستند.

مسئله زمانی پیچیده‌تر می‌شود که σ^2، واریانس خطا، مثبت باشد. در ازای $x_. \in S_X$ معین، توزیع $Y|X = x_.$ توزیع تبدیل یافته D با انتقال به اندازه $g(x_.)$ و تغییر مقیاس به اندازه ۱ است.از رابطه‌ی (۱) داریم:

$$Y|X = x_. = g(x_.) + e$$

از خاصیت خطی بودن امید ریاضی نتیجه می‌شود که

$$E(Y|X = x_.) = E(g(x_.)) + E(e) = g(x_.)$$

به علاوه از (۲) به عبارت زیر می‌رسیم:

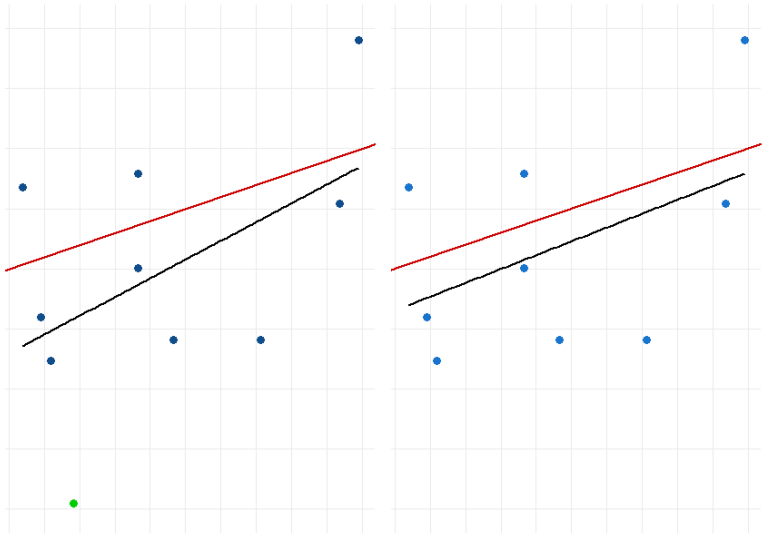
$$\mathrm{Var}(Y|X = x_\cdot) = \mathrm{Var}(g(x_\cdot)) + \mathrm{Var}(e) = \mathrm{Var}(e) = \sigma^۲$$

بنابراین در حالتی که $\sigma^۲ > ۰$، هر مشاهده‌ای که در نقطه‌ی $x_\cdot$ ثبت شود، نمونه‌ای تصادفی از توزیع $D(g(x_\cdot), \sigma^۲)$ بوده و با توجه به شکل تابع چگالی خطا، حول $g(x_\cdot)$ پراکنده است. به علت پیوسته بودن D، احتمال مشاهده کردن مقدار صحیح $g(x_\cdot)$ برابر صفر است و تنها می‌توان گفت که به صورت میانگین فاصله‌ی مقدار ثبت شده از $g(x_\cdot)$ با کاهش σ کمتر می‌شود.
حتی اگر فرم کلی تابع را بدانیم، تشخیص فرمول دقیق آن به علت تصادفی بودن $Y|X$ تقریبا غیر ممکن است و فقط می‌توان ضرایب و پارامترها را در سطح خطای مشخص تخمین زد.

معمولا اطلاعاتی به جز مشاهدات در دسترس نیست. در بیشتر روش‌های مدل‌سازی متداول از روی نمودار پراکنش گزینه‌هایی برای فرم کلی g در نظر می‌گیرند و در هر حالت فرمول تابع را تخمین می‌زنند. سپس، معیاری که توسط آنالیزگر برای تشخیص مدل بهتر انتخاب می‌شود، مناسب‌ترین مدل را مشخص می‌نماید.

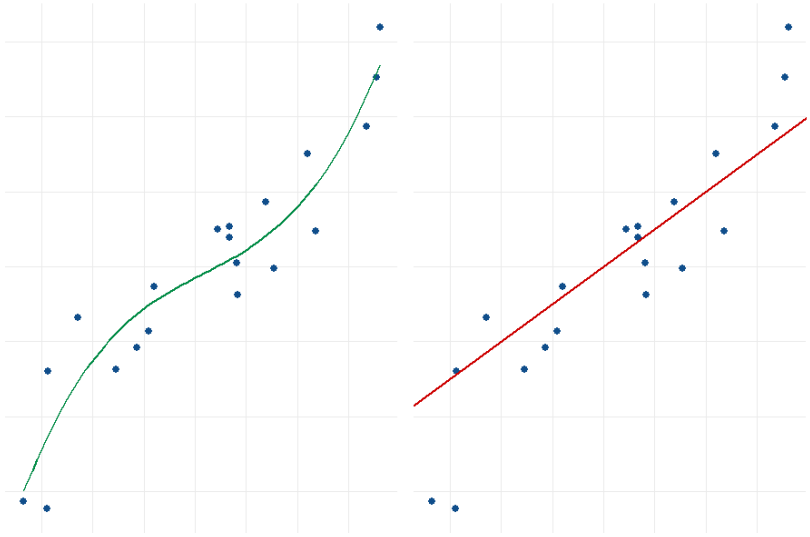
داده‌های پرت می‌توانند معیار سنجش مدل را تحت تاثیر قرار دهند. از طرفی موجب فاصله گرفتن برآوردها از مقدار واقعی آنها می‌شوند. برخی موارد ممکن است وجود این داده‌ها در نمودار پراکنش باعث شود گزینه‌های نادرستی را برای شکل کلی g در نظر بگیریم. از این رو شناسایی و نحوه‌ی استفاده از اطلاعات این نقاط از مراحل مهم برازش مدل محسوب می‌شود.

برای مثال در شکل ۱.۲ وجود داده‌های پرت سبب بیش برآورد شیب خط در مدل رگرسیون خطی ساده شده است. اگر این داده وجود نداشت یا قابل تشخیص بود، مدل برازش یافته به واقعیت نزدیک‌تر می‌شد.
خطوط قرمز و مشکی به ترتیب g و خط برازش داده شده را مشخص کرده‌اند. نقطه‌ی سبز در نمودار سمت چپ داده‌ی پرت بوده که در نمودار سمت راست پس از حذف آن خط رگرسیونی برازش یافته جدید رسم شده است.



شکل ۱.۲: تاثیر حذف داده‌ی پرت بر شیب خط رگرسیونی

در شکل ۲.۲ وجود داده‌های پرت ممکن است باعث گمراه شدن آنالیزگر حین حدس زدن شکل تابع شود. تابع g خطی است اما ممکن است چندجمله‌ای از درجه ۳ تصور شود.



شکل ۲.۲: تشخیص نادرست شکل تابع g در حضور داده‌های پرت

در بسیاری از موارد مقادیر بزرگ $\sigma^۲$ عامل دیگری است که باعث می‌شود نتوانیم از روی نمودار پراکنش حدس‌های درستی از فرم کلی g بزنیم.
اگر به ازای $x_۰ \in S_X$ مشخص، $Y|X = x_۰$ را توزیع شرطی پاسخ و $y|X = x_۰$ را نمونه‌ی مشاهده شده از این توزیع در نقطه‌ی $x_۰$ فرض کنیم، برای $\delta > ۰$ ثابت قرار دهید :

$$P(|Y|X = x_۰ - g(x_۰)| \geq \delta) = A$$

متغیر تصادفی $Y'|X = x_۰$ را هم توزیع $Y|X = x_۰$ با تغییر مقیاس به اندازه $k > ۱$ در نظر بگیرید :

$$Y'|X = x_۰ = k(Y|X = x_۰ - g(x_۰)) + g(x_۰) \Rightarrow$$

$$\mathbf{Var}(Y'|X = x_۰) = k^۲\mathbf{Var}(Y|X = x_۰) = (k\sigma)^۲ > \sigma^۲,$$

$$E(Y'|X = x_۰) = kE(Y|X = x_۰ - g(x_۰)) + E(g(x_۰)) = g(x_۰) \Rightarrow$$

$$Y'|X = x_۰ \sim D(g(x_۰), (k\sigma)^۲)$$

بنابراین $Y'|X = x_۰$ نسخه‌ای مشابه $Y|X = x_۰$ است که واریانس آن افزایش یافته است. در این حالت داریم :

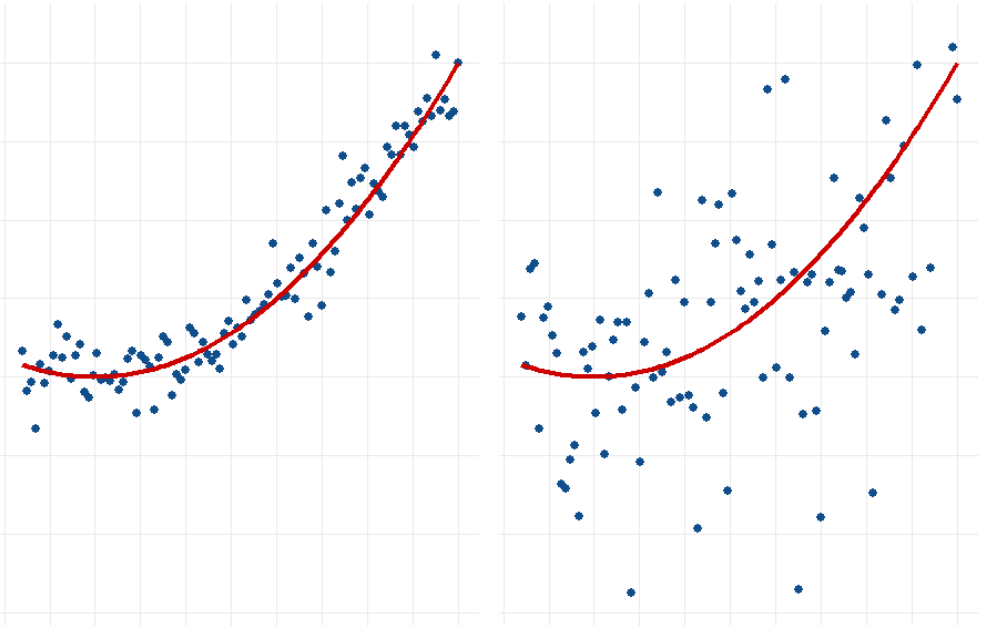
$$P(|Y'|X = x_۰ - g(x_۰)| \geq \delta) =$$

$$P(|(k(Y|X = x_۰ - g(x_۰)) + g(x_۰) + g(x_۰)))| \geq \delta) =$$

$$P(k|Y|X = x_۰ - g(x_۰)| \geq \delta) = P(|Y|X = x_۰ - g(x_۰)| \geq \frac{\delta}{k}) = B$$

می‌دانیم $k > ۱$. برای همین، $\delta > \frac{\delta}{k}$. پس داریم $A \leq B$.

از توضیحات بالا نتیجه می‌شود که در شرایط ثابت، افزایش واریانس خطا باعث می‌شود احتمال آن که در هر نقطه مقدار مشاهده شده متغیر پاسخ با مقدار g از $\delta > ۰$ مشخص بیشتر شود، افزایش یابد. بنابراین، احتمال بیشتری وجود دارد که نقاط روی نمودار پراکنش با فاصله‌ی زیادی از g قرار بگیرند. از این رو، تشخیص شکل کلی تابع g با افزایش $\sigma^۲$ دشوارتر می‌شود.
در شکل ۳.۲ مشاهده می‌شود که حدس زدن تابع g در نمودار پراکنش سمت راست پیچیده‌تر از نمودار سمت چپ است. شاید حتی درجه دوم بودن تابع قایل تشخیص نبوده و مدل خطی برای داده‌ها برازش داده شود.



شکل ۳.۲: تاثیر افزایش واریانس بر نمودار پراکنش

توزیع خطا چالش دیگری است که در مدل سازی آماری مطرح می‌شود. در حالت کلی D برای آنالیزگر ناشناخته است. این که خطا از چه توزیعی پیروی می‌کند در ساختن تابع درستنمایی و استفاده از بیشتر ابزارهای محاسباتی حائز اهمیت است. معمولا برای آن که بسیاری از روش‌های مدل سازی قابل پیاده سازی باشند، توزیع و ویژگی‌های خاصی برای خطا فرض می‌کنند. بدیهی است که الزامی برای برقرار بودن این فرضیات وجود ندارد.
برای مثال در رگرسیون خطی و مدل‌های طرح آزمایش، نرمال بودن خطا از فرض‌های اولیه مدل است و در صورتی که خطا نرمال نباشد بسیاری از برآوردها ، نتایج و تفاسیر بی اعتبار می‌شوند.

در این نوشته راهکارهایی برای مواجهه با چالش‌های اشاره شده معرفی می‌شوند که با کمترین فرضیات و محدودیت‌ها می‌توانند آنالیزگر را در جهت تحلیل بهتر داده‌ها و انتخاب مدل مناسب هدایت کنند.
در ادامه الگوریتمی برای برآورد F، تابع توزیع خطا، و بررسی ویژگی‌های پر اهمیت آن معرفی می‌شود.
برای آن که مجاز به استفاده از روش‌های مدل سازی‌ای شویم که در آنها نرمال بودن خطا از فرضیات مهم است، راهکاری ارائه می‌شود. در این روش، داده‌ها طوری تغییر پیدا می‌کنند که در تئوری،

توزیع خطا نرمال یا بسیار شبیه نرمال می‌شود.

۲.۲ معرفی الگوریتم‌ها

عناوین و توضیح مختصری از الگوریتم‌های مورد بحث این نوشته به شرح زیر است :

۱.۲.۲ الگوریتم برچسب گذاری

این الگوریتم دورترین نقاط از تابع g را در نمودار پراکنش شناسایی می‌کند. میزان حساسیت آن برای مشاهداتی که در بالا یا پایین منحنی قرار دارند به صورت جداگانه قابل تنظیم است. این موضوع بررسی میزان چولگی توزیع خطا را ممکن می‌سازد. این که آنالیزگر از این برچسب‌ها چه استفاده‌ای می‌کند می‌تواند نسبت به داده‌ها و هدف مسئله متفاوت باشد. جابه‌جایی و حذف نقاط از گزینه‌های ممکن است.
نقاط پرت نیز پس از اجرای این روش مشخص می‌شوند. با توجه به پراکندگی داده‌ها و اندازه‌ی نمونه ممکن است پس از تکرار الگوریتم تحت تنظیمات مناسب شکل کلی تابع g در نمودار پراکنش واضح‌تر شود. بنابراین، این الگوریتم در شرایطی که واریانس خطا بزرگ است مراحل مدل سازی را ساده‌تر می‌کند.

۲.۲.۲ روش ناپارامتری برآورد F، تابع توزیع خطا

این روش زمانی که توزیع خطا برای ما ناشناخته است بسیار کاربرد دارد. پس از آن که $\widehat{F}$ تعیین شد، با بهره‌گیری از آزمون فرض‌های مناسب می‌توان میزان شباهت D را با توزیع‌های دلخواه سنجید. واریانس، میزان چولگی و شکل تقریبی تابع توزیع و چگالی با داشتن $\widehat{F}$ قابل بحث است.

۳.۲.۲ تغییر توزیع خطا

به هنگام استفاده از روش‌های مدل سازی‌ای که در آنها الزام داشته باشد که خطا از توزیع خاصی، مثلا نرمال، پیروی کند، می‌توان تبدیلی بر روی داده‌ها اعمال کرد که تا حد خوبی به شرایط مطلوب نزدیک شوند.

تمرکز اصلی این نوشته به روی حالت تک متغیره بردار X است. البته که امکان تعمیم مطالب به

مدل‌های چند متغیره نیز وجود دارد.

۳.۲ فرضیات و اساس اصلی الگوریتم‌ها

تعریف ۱.۲. *R_X را بست حوزه‌ی مقادیر مشاهده شده‌ی X در $\mathbb{R}$ تعریف می‌کنیم :*

$$R_X = [min(\{x_1, \ldots, x_n\}), max(\{x_1, \ldots, x_n\})]$$

فرض ۲.۲. *$E(Y|X)$ موجود و متناهی باشد.*

فرض ۳.۲. *مجموعه‌ی مقادیر متغیر تصادفی X فشرده باشد.*

فرض ۴.۲. *تابع حقیقی و پیوسته‌ی g وجود داشته باشد به طوری که*

$$Y|X = g(X) + e, \quad E(Y|X) = g(X)$$

تذکر ۵.۲. *در فرض ۲.۴ اگر g در محدوده‌ی Xهای مشاهده شده نیز پیوسته باشد کفایت می‌کند زیرا در مدل‌سازی فقط قسمتی از تابع که از آن اطلاعات در اختیار است مورد بررسی قرار می‌گیرد و پیشنهاد می‌شود خارج از بازه‌ی مشاهدات تفسیری انجام نگیرد.*
بنابراین می‌توانیم مجموعه‌ی مقادیر X را با بست محدوده‌ی مشاهدات در فرض‌های ۲.۳ و ۲.۴ جا به جا کنیم. اگر این مجموعه فشرده باشد امکان اجرای الگوریتم فراهم می‌شود.

۱.۳.۲ افراز حوزه‌ی مقادیر X

قضیه‌ی زیر پایه‌ی اصلی عملکرد هر سه الگوریتم است :

قضیه ۶.۲. *افرازی وجود دارد که R_X را به v بخش $(section)$ به صورت زیر تقسیم می‌کند :*

$$section_1 = [min(\{x_1, \ldots, x_n\}), w_1],$$

$$section_2 = [w_1, w_2],$$

$$\vdots$$

$$\textbf{\textit{section}}_v = [w_{v-1}, max(\{x_1, \ldots, x_n\})]$$

$$s.t \quad min(\{x_1, \ldots, x_n\}) < w_1 < \cdots < w_{v-1} < max(\{x_1, \ldots, x_n\}),$$

$$\bigcup_{i=1}^{v} \textbf{\textit{section}}_i = R_X$$

به طوری که در ازای $\delta > 0$ *دلخواه برای هر* $i \in \{1, \ldots, v\}$ *داشته باشیم :*

$$|max(g(x)|x \in \textbf{\textit{section}}_i) - min(g(x)|x \in \textbf{\textit{section}}_i)| < \delta$$

اثبات. اگر بازه‌ی بسته‌ای زیرمجموعه‌ی محدوده‌ی مقادیرX در نظر بگیریم، از قضیه‌ی ۸.۱ فشردگی آن نتیجه می‌شود. این موضوع همراه با نتیجه‌ی قضیه‌ی ۱۱.۱ وجود مینیمم و ماکزیمم تابع g را در این بازه تضمین می‌کند.
از فرض ۳.۲ و قضیه‌ی ۱۳.۱ ثابت می‌شود که g بر حوزه‌ی مقادیر X به صورت یکنواخت پیوسته است.
طبق تعریف پیوستگی یکنواخت، برای هر $\delta > 0$ دلخواه، $\delta_2 > 0$ وجود دارد به طوری که برای هر p و q در حوزه‌ی مقادیر X که $|p - q| < \delta_2$ داشته باشیم

$$|g(p) - g(q)| < \delta_2$$

اکنون $\delta'_2 < \delta_2$ را بسیاز نزدیک به δ_2 اختیار می‌کنیم و بخش‌ها را با شروع از $min(\{x_1, \ldots, x_n\})$ به صورت بازه‌هایی به طول δ'_2 به این شکل می‌سازیم:

$$[min(\{x_1, \ldots, x_n\}), min(\{x_1, \ldots, x_n\}) + \delta'_2],$$

$$[min(\{x_1, \ldots, x_n\}) + \delta'_2, min(\{x_1, \ldots, x_n\}) + 2\delta'_2],$$

$$\vdots,$$

$$[min(\{x_1, \ldots, x_n\}) + (v-1)\delta'_2, max(\{x_1, \ldots, x_n\})]$$

واضح است که در این تقسیم بندی ممکن است طول آخرین بخش از δ'_2 کمتر شود.

بنابراین، v بزرگترین عدد صحیحی خواهد بود که در نابرابری زیر صدق کند :

$$0 < max(\{x_1, \ldots, x_n\}) - min(\{x_1, \ldots, x_n\}) - (v-1)\delta_2' \leq \delta_2'$$

برای ساده‌تر شدن برای $j \in \{0, \ldots, v-1\}$ قرار دهید :

$$w_j = min(\{x_1, \ldots, x_n\}) + j\delta_2', \quad w_v = max(\{x_1, \ldots, x_n\})$$

بنابراین، kامین بازه‌ی بسته در این تقسیم‌بندی $[w_{k-1}, w_k]$ خواهد بود.

برای هر $i \in \{1, \ldots, v\}$ بازه‌ی $[w_{i-1}, w_i]$ یک مجموعه‌ی بسته از حوزه‌ی مقادیر X و در نتیجه فشرده است. پیوستگی g و فشردگی $[w_{i-1}, w_i]$ ، پیوستگی یکنواخت g را در بخش iام اثبات می‌کند. می‌دانیم تابع در این قسمت به مینیمم و ماکزیمم خود می‌رسد. پس $x_1, x_2 \in [w_{i-1}, w_i]$ وجود دارند به طوری که

$$g(x_1) = min(g(x)|x \in \text{section}_i), g(x_2) = max(g(x)|x \in \text{section}_i)$$

داریم :

$$|x_2 - x_1| \leq |w_i - w_{i-1}| = \delta_2'$$

از پیوستگی یکنواخت تابع در این قسمت نتیجه می‌شود که

$$|max(g(x)|x \in \text{section}_i) - min(g(x)|x \in \text{section}_i)| =$$

$$|g(x_2) - g(x_1)| < \delta$$

□

تذکر ۷.۲. *دقت شود که v بخش $\text{section}_1, \ldots, \text{section}_v$ افرازی از R_X نیستند زیرا اشتراک هر بخش با بخش بعدی مجموعه‌ی تک نقطه‌ای مانند w_j و ناتهی است. بنابراین در تعریف افراز نمی‌گنجند.*
افراز مورد نظر را می‌توان به صورت $[x_{(1)}, w1), [w1, w2), \ldots, [w_{v-1}, x_{(n)}]$ در نظر گرفت که از روی آن می‌توان $\text{section}_1, \ldots, \text{section}_v$ را ساخت. در اینجا منظور از $x_{(i)}$ مقدار مشاهده شده‌ی آماره‌ی ترتیبی i ام از متغیر تصادفی X است.

تذکر ۸.۲. *قضیه‌ی ۶.۲ وجود چنین افرازی را ثابت می‌کند. لزومی ندارد روشی که برای ساختن تقسیم‌بندی بیان شد بهترین و بهینه‌ترین انتخاب باشد. ممکن است افراز دیگری وجود داشته باشد که بخش‌هایی با طول‌های نابرابر ایجاد کند که باعث شود الگوریتم به مراتب بر روی داده‌ی مورد مطالعه عملکرد بهتری داشته باشد.*

تذکر ۹.۲. *در ادامه اثبات می‌شود که فرض ۴.۲ و ۳.۲ با فرض ۱۰.۲ معادل است.*

فرض ۱۰.۲. *تابع حقیقی و پیوسته یکنواخت g وجود داشته باشد به طوری که*

$$Y|X = g(X) + e, \quad E(Y|X) = g(X)$$

قضیه ۱۱.۲. *ترکیب فرض ۳.۲ و ۴.۲ با فرض ۱۰.۲ معادل است.*

اثبات. از قضیه‌ی ۱۳.۱ نتیجه می‌شود که در صورتی که حوزه‌ی مقادیر X فشرده باشد، انگاه g بر این مجموعه پیوسته یکتواخت است. □

تذکر ۱۲.۲. *از قضیه‌ی هاینه- بورل ثابت می‌شود که می‌توانیم فرض ۳.۲ را با فرض زیر جابه‌جا کنیم:*

فرض ۱۳.۲. *مجموعه‌ی مقادیر X بسته و کراندار است.*

تقسیم بندی اشاره شده در قضیه‌ی ۶.۲ پایه‌ی سه الگوریتم مورد بحث در این نوشته است. فرضیات بالا الزام وجود چنین تقسیم بندی‌ای را تضمین می‌نمایند. البته توابعی وجود دارند که فرض‌های ۳.۲ و ۴.۲ یا ۱۰.۲ برایشان برقرار نیست اما تقسیم بندی مورد نظر برایشان وجود دارد. برای همین می‌توانیم فرض ۱۴.۲ را به عنوان جایگزینی از فرض‌های قبلی در نظر بگیریم زیرا آنچه برای اجرای الگوریتم ضرورت دارد در حکم قضیه‌ی ۶.۲ مشخص شده‌است. در صورتی که تشخیص برقرار بودن حکم قضیه‌ی ۶.۲ دشوار باشد باید به دنبال تحقیق در درستی ترکیب فرض‌های ۳.۲و ۴.۲ و یا فرض پیوستگی یکنواخت g باشیم تا مجاز به استفاده از الگوریتم‌ها شویم.

فرض ۱۴.۲. *حکم قضیه‌ی ۶.۲ برقرار باشد.*

حالا برای ساده‌تر شدن محاسبات داده‌ها را به این شکل نمایش می‌دهیم :

اگر n اندازه‌ی نمونه باشد، n_i را تعداد نقاط در بخش iام می‌نامیم. داده‌ها را بر اساس مقدار متغیر پیشگو مرتب کرده و jامین داده در بخش i را به صورت زوج (x_{i_j}, y_{i_j}) و خطای متناظر با آن را با e_{i_j} نشان می‌دهیم ($i \in \{1, \ldots, v\}$ و $j \in \{1, \ldots, n_i\}$).

۴.۲ فرض‌هایی درباره‌ی رفتار g در هر بخش

پس از آن که از وجود تقسیم بندی اشاره شده تحت فرضیات مناسب مطمئن شدیم، یکی از فرض‌های زیر را به عنوان آخرین فرض مهم این بخش می‌پذیریم.

۱.۴.۲ ثابت بودن

فرض ۱۵.۲. *در صورتی که δ را در حکم قضیه ۶.۲ به اندازه‌ی کافی کوچک انتخاب کنیم، با خطای قابل اغماض می‌توانیم g را در هر بخش ثابت فرض کنیم. یعنی برای هر $i \in \{۱, \dots, v\}$ ثابت $c_i \in \mathbb{R}$ وجود دارد به طوری که در بخش iام برای هر $j \in \{۱, \dots, n_i\}$ داریم :*

$$Y_{i_j} = c_i + e_{i_j}$$

۲.۴.۲ خطی بودن

فرض ۱۶.۲. *در بخش iام در صورتی که δ را به اندازه‌ی کافی کوچک اختیار کنیم، با خطای قابل اغماض، g خطی است. $(i = ۱, \dots, v)$*

این خط را با $l_i = a_i x + b_i$ نمایش می‌دهیم.

۳.۴.۲ رفتار دلخواه

امکان آن که برای توضیح دادن رفتار g در بخش‌ها، گزینه‌های دیگر، مانند چندجمله‌ای در نظر بگیریم وجود دارد. البته احتمال دارد که برآوردگرها و رفتار الگوریتم کمی تغییر کند. بنابراین، نیاز است که بررسی‌های لازم در این زمینه صورت پذیرد. توضیحات این حالت از محدوده‌ی این نوشته خارج است.

اگر برای δ مشخص v را تعیین کنیم، با کوچک شدن δ تعداد بخش‌ها یا ثابت می‌ماند و یا بیشتر می‌شود. پس می‌توان گفت v و δ با یکدیگر رابطه‌ی معکوس دارند.
δ به نوعی هزینه‌ای است که آنالیزگر حاضر است بپردازد تا فرض‌های ۱۵.۲ و ۱۶.۲ را با کران بالای خطای δ بپذیرد یا به عبارت دیگر خطایی که پذیرفتن این فرض دارد برایش قابل اغماض باشد. نتیجه‌ی آن ساده‌تر شدن محاسبات و فراهم شدن امکان استفاده او از الگوریتم‌هاست.

این که چه مقداری از δ برای مدل سازی مناسب است بسته به نوع مسئله و هدف آن تغییر می‌کند. لزوما مقادیر کوچک که خطای پذیرش هر یک از فرض‌های ۱۵.۲ و ۱۶.۲ را کم می‌کنند مطلوب نیستند.
در فصل‌های بعدی به تفصیل نحوه‌ی عملکرد الگوریتم‌ها در هر بخش شرح داده می‌شود. به صورت خلاصه می‌توان گفت که آنالیزهایی در هر قسمت صورت می‌گیرند که دقت آنها با اندازه‌ی نمونه رابطه‌ی مستقیم دارد. اگر δ را خیلی کوچک بگیریم ممکن است بخش‌هایی وجود داشته باشند که هیچ داده‌ای در آنها ثبت نشده و یا حجم داده در آنها بسیار کم باشد. شاید بتوان رفتار δ و v در مدل سازی را مشابه مبادله‌ی اریبی و واریانس [۱] خواند.

۴.۴.۲ برآورد ثابت‌های c_i

ثابت‌های $c_1, \ldots, c_v$ پارامتر هستند. پس، با انتخاب آماره‌ی مناسب می‌توانیم آنها را برآورد کنیم.

قضیه ۱۷.۲. *تحت فرض ۱۵.۲ میانگین نمونه‌ای داده‌های هر بخش برآوردگر ناریب ثابت c در آن قسمت است.*

اثبات. تحت فرض ۱۵.۲ برای هر $i \in \{1, \ldots, v\}$ و هر $j \in \{1, \ldots, n_i\}$ داریم :

$$Y_{i_j} = c_i + e_{i_j}$$

امید ریاضی میانگین نمونه‌ای داده‌های بخش iام برابر است با:

$$E\left(\frac{\sum_{j=1}^{n_i} y_{i_j}}{n_i}\right) = \frac{1}{n_i}\sum_{j=1}^{n_i} E(y_{i_j}) = \frac{1}{n_i}\sum_{j=1}^{n_i} E(c_i + e_{i_j}) =$$

$$\frac{1}{n_i}\sum_{j=1}^{n_i}\big(E(c_i) + E(e_{i_j})\big) = \frac{1}{n_i}(n_i c_i) = c_i$$

□

قضیه ۱۸.۲. *تحت فرض ۱۶.۲ خط برازش یافته‌ی $\widehat{l_i}$ از روش رگرسیون خطی برآوردگر نااریب l_i در هر نقطه‌ی بخش iام است. $i = 1, \ldots, v$*

[۱] bias-variance tradeoff

اثبات. در صورتی که فرض‌های رگرسیون خطی برقرار باشند، برآوردهای شیب و عرض از مبدا در این روش نااریب هستند. بنابراین خط برازش یافته در نقطه‌ی دلخواه x، در هر بخش برآورد نااریب $l(x_.)$ خواهد بود. □

تذکر ۱۹.۲. *شبیه سازی‌های انجام شده نشان می‌دهد که فرض ۱۶.۲ و استفاده از رگرسیون خطی در برآورد l در هر بخش، دقت الگوریتم را افزایش می‌دهد و نسبت به توزیع‌های متفاوت خطا استوارتر عمل می‌کند.*
هزینه‌ی بهره‌مندی از این ویژگی‌های برتر برقراری فرض‌های رگرسیون خطی است. توصیه می‌شود که در صورتی که امکان سنجیدن این مفروضات وجود داشته باشد و از برقراری آن‌ها اطمینان حاصل شود، از ترکیب فرض ۱۶.۲ و خط رگرسیونی برای توضیح دادن رفتار g استفاده شود.

در فصل‌های بعد به تفصیل در مورد نحوه عملکرد هر الگوریتم توضیح می‌دهیم.

فصل ۳

روش برآورد تابع توزیع خطا

۱.۳ ادغام بخش‌ها

۱.۱.۳ تحت فرض ثابت بودن g در هر بخش

طبق عبارت (۱) در فصل اول رابطه‌ی بین متغیرهای پیشگو و پاسخ به صورت $Y|X = g(X)+e$ است. تحت فرضیاتی که در فصل دوم مطرح کردیم، این رابطه در بخش iام، با خطای قابل اغماض، به صورت $Y|(X \in \text{section}_i) = c_i + e$ در نظر گرفته شد. مشاهده‌ی jام در iامین قسمت را نیز با $y_{i_j} = c_i + e_{i_j}$ نشان دادیم.
چون $e_{i_j} = y_{i_j} - c_i$، می‌توانیم با کم کردن c_i از n_i مشاهده‌ی بخش iام مجموعه‌ی

$$E_i = \{e_{i_1}, \dots, e_{i_{n_i}}\}$$

را تشکیل دهیم.

۲.۱.۳ تحت فرض خطی بودن g در هر بخش

با پذیرش فرض ۱۶.۲ داده‌های هر بخش در رابطه‌ی زیر صدق می‌کنند :

$$Y|(X \in \text{section}_i) = l_i(X) + e$$

مشابه حالت قبل، اعضای مجموعه‌ی E_i از رابطه‌ی $e_{i_j} = y_{i_j} - l(x_{i_j})$ تعیین می‌شوند.

۳.۱.۳ تشکیل مجموعه‌ی E

قضیه ۱.۳. *هر عضو E_i مشاهده‌ای تصادفی از توزیع $D(۰, \sigma^۲)$ است.* $(i = ۱, \dots, v)$

اثبات. .

حالت اول : ثابت بودن تابع در هر بخش

از قضیه‌ی ۵.۱ داریم :

$$Y|X = x \sim D(g(x), \sigma^۲)$$

برای هر $i \in \{۱, \dots, v\}$ مقدار g طبق فرض ۱۵.۲ برابر ثابت c_i تعریف شده است. بنابراین، گزاره‌ی اول را می‌توان به صورت زیر نوشت :

$$\forall i \in \{۱, \dots, v\} \quad Y|(X \in \text{section}_i) \sim D(c_i, \sigma^۲)$$

e_{i_j} تبدیلی خطی از y_{i_j} است. از این رو، دارای توزیع تبدیل یافته‌ی D است.

$$E(e_{i_j}) = E(y_{i_j}) - E(c_i) = c_i - c_i = ۰$$

$$\text{Var}(e_{i_j}) = \text{Var}(y_{i_j}) + \text{Var}(c_i) - ۲\text{cov}(y_{i_j}, c_i) = \text{Var}(y_{i_j}) = \sigma^۲$$

پس e_{i_j} مشاهده‌ای تصادفی از توزیع $D(۰, \sigma^۲)$ است.

حالت دوم : خطی بودن تابع در هر بخش

برای هر $i \in \{۱, \dots, v\}$ مقدار g طبق فرض ۱۶.۲ برابر خط $l_i(x)$ تعریف شده است. برای هر نقطه‌ی دلخواه $x.$ در بخش iام داریم :

$$Y|X = x. \sim D(l_i(x.), \sigma^۲)$$

با استدلال مشابه حالت قبل، e_{i_j} دارای توزیع تبدیل یافته‌ی D است. اکنون کافیست میانگین و واریانس آن را محاسبه کنیم :

$$E(e_{i_j}) = E(y_{i_j}) - E(l_i(x_{i_j})) = l_i(x_{i_j}) - l_i(x_{i_j}) = ۰$$

$$\mathrm{Var}(e_{i_j}) = \mathrm{Var}(y_{i_j}) + \mathrm{Var}(l_i(x_{i_j})) - ۲\mathrm{cov}(y_{i_j}, l_i(x_{i_j})) = \mathrm{Var}(y_{i_j}) = \sigma^۲$$

□

باتوجه به هر فرض برآوردگر متفاوتی برای تخمین تابع g در هر بخش ارائه شد. از این پس این برآوردگر را با $\widehat{g}$ نمایش می‌دهیم تا نیاز نباشد در هر قسمت حالت‌های متفاوت را بررسی کنیم. ثابت شد که $\widehat{g}$ معرفی شده متناظر هر یک از فرض‌های ۱۵.۲ و ۱۶.۲ برآوردگری نااریب است.
فرایندی که در هر نقطه، g را برآورد و از مقدار مشاهده شده‌ی متغیر پاسخ کم می‌کند را حذف کردن اثر تابع gدر آن نقطه می‌نامیم.

تذکر ۲.۳. *g را پیوسته فرض کردیم اما واضح است که الزامی برای پیوستگی $\widehat{g}$ وجود ندارد. $\widehat{g}$ تابعی ضابطه‌دار است که فرمول آن از بخشی به بخش دیگر تغییر می‌کند. تنها می‌توان گفت که $\widehat{g}$ در هر بخش پیوسته است.*

تعریف ۳.۳. *مجموعه‌ی اجتماع تمام E_iها را E می‌نامیم. یعنی، $E = \bigcup_{i=۱}^{v} E_i$.*

از قضیه‌ی ۱.۳ نتیجه می‌شود که $E = \{e_۱, \ldots, e_n\}$. همانطور که در فصل اول اشاره شد e_j، مشاهده‌ای از توزیع خطا متناظر با jامین داده است. از این رو، E نمونه‌ی تصادفی به حجم n از توزیع D خواهد بود. اکنون می‌توانیم با به کارگیری روش‌های آماری متداول و آزمون فرض‌های مرتبط به بررسی ویژگی‌های متغیر خطا بپردازیم.

۲.۳ برآورد اعضای E

فرض کنید از متغیر تصادفی دلخواه Ω نمونه‌ای به حجم مشخص اتخاذ کرده‌ایم. Ω_i را متغیر تصادفی متناظر با iامین واحد نمونه و ω_i را مقدار مشاهده شده‌ی آن تعریف می‌کنیم. پیش از آن که نمونه‌گیری انجام گیرد، ω_i بی‌معنی است. اما می‌دانیم زمانی که مشاهدات ثبت شوند، ω_i عددی ثابت بوده که مقدار آن متناسب با توزیع Ω_i یا Ω تعیین می‌شود. درست است که پس از هر بار نمونه‌گیری، ممکن است مقادیر جدیدی را مشاهده کنیم اما بعد از اتمام فرایند جمع‌آوری داده، مقادیر ثبت شده غیر قابل تغییر و در نتیجه غیر تصادفی هستند.
در آمار با تعیین آماره‌ی مناسب می‌توانیم ثابت‌های ناشناخته را برآورد کنیم. این که مقدار برآورد تا چه اندازه به واقعیت نزدیک است به آماره و اطلاعاتی که در دست داریم ربط پیدا می‌کند.

بدون نقض کردن تعاریف می‌توانیم بگوییم که در حقیقت مقدار ω_i مشاهده شده با خطای صفر با ω_i برآورد می‌شود. در اینجا اطلاعاتی که در دست داریم و آماره‌ای که انتخاب کردیم شرایطی را فراهم کرده است که بتوانیم تا این اندازه دقت داشته باشیم. شاید از آنجایی که دقت این تخمین ۱۰۰ درصد است واژه‌ی برآورد گزینه‌ی مناسبی نباشد. برای همین است که در بیشتر متون علمی این کار را تخمین یا برآورد نمی‌نامند. در ادامه مثالی غیر بدیهی‌تر را از تخمین مقادیر مشاهده شده‌ی یک نمونه‌ی تصادفی ارائه می‌دهیم.

هر واحد داده برداری است از اندازه‌گیری‌ها که هر عنصر آن نمونه‌ای تصادفی به حجم یک از توزیع متغیر تصادفی متناظر با آن است. متغیر تصادفی منسوب به خطا به صورت مستقیم قابل اندازه‌گیری نیست. صرفا از وجود آن آگاه هستیم و می‌توانیم جایگاهش را در روابط ریاضی میان متغیرها مشخص کنیم. برای واحد iام داده، یک مشاهده از توزیع D وجود دارد که آن را طبق قراردادهای فصل اول با e_i نمایش می‌دهیم. بر خلاف متغیرهای پیشگو و پاسخ مقادیر e_iها در دسترس نیستند. به دنبال آن هستیم که E که مجموعه‌ی تمام e_iها است را برآورد کنیم.

به علت آن که a_i, b_i و c_iها پارامترهای نامعلوم هستند، امکان تعیین دقیق E وجود ندارد. از جایی که مقادیر متغیر پاسخ ثبت شده‌اند، در صورت انتخاب برآوردگر مناسب برای ثابت‌های a, b, c، e_iها از رابطه‌ی $\widehat{e}_{i_j} = y_{i_j} - \widehat{g}(x_{i_j})$ محاسبه می‌شوند.

قضیه ۴.۳. *$\widehat{e}_j = y_j - \widehat{g}(x_j)$ برآوردگر نااریب jامین نمونه‌ی تصادفی از توزیع خطاست.*

اثبات. برهان این قضیه از ناارییی $\widehat{g}$ نتیجه می‌شود.

□

$\widehat{E}$ را مجموعه‌ی تمام $\widehat{e}$ها می‌نامیم. زمانی که E قابل محاسبه نباشد از $\widehat{E}$ به عنوان جایگزین آن استفاده می‌کنیم.

۳.۳ استنباط بر پایه‌ی $\widehat{E}$

اعضای E به صورت پنهان در داده‌ها وجود دارند. نزدیک‌ترین جایگزینی که برای این مجموعه می‌شناسیم $\widehat{E}$ است.

خطای هر برآورد از نداشتن اطلاعات کافی نشأت می‌گیرد. از این رو، در تخمین اعضای E دقت کامل نخواهیم داشت. از طرفی هر استنباطی که از روی نمونه‌ی تصادفی انجام می‌شود، خطا به همراه دارد. بنابراین، در مطالعه‌ی نتایجی که بر پایه‌ی تحلیل $\widehat{E}$ به عنوان نمونه‌ی تصادفی از توزیع

خطا به دست می‌آیند، در دو مرحله خطا مرتکب می‌شویم. البته اگر حجم نمونه بیشتر باشد، میزان خطا کم می‌شود. با این حال، اطلاعاتی که از $\widehat{E}$ به کسب می‌شود ما را تا حد بسیار خوبی در مراحل مدل سازی یاری می‌دهند.

هر روش آماری و آزمون فرضی که بر پایه‌ی نمونه‌ی تصادفی باشد، بر $\widehat{E}$ قابل پیاده‌سازی است. در ادامه چند مورد را بیان می‌کنیم :

۱.۳.۳ توزیع نمونه‌ای خطا

با تشکیل نمودارهای تراکم و هیستوگرام شکل کلی تابع چگالی خطا، پراکندگی، میزان چولگی و قرینگی آن قابل تشخیص است. می‌توانیم توزیع‌های که چگالی آن‌ها مشابه است را حدس زده و با به کارگیری آزمون فرض‌هایی مانند *کای-دو* و *کلموگروف-اسمیرنوف* آن‌ها را آزمون کنیم.

اگر توزیع خطا ناشناخته باشد و علاقه داشته باشیم از روش‌های مدل سازی‌ای استفاده کنیم که در آن‌ها نرمال بودن خطا مهم است، این نمودارها کاربردی هستند. در صورتی که شکل توزیع نمونه‌ای $\widehat{E}$ زنگوله‌ای و شبیه نرمال نباشد یا نشانه‌هایی از عدم قرینگی ببینیم، این روش‌ها انتخاب نامناسبی خواهند بود.

۲.۳.۳ برآورد پارامترهای توزیع خطا

در آمار ریاضی برای برآورد پارامترهای هر توزیع برآوردگرهایی بر پایه‌ی نمونه‌ی تصادفی معرفی شده‌اند. با استفاده از اطلاعات $\widehat{E}$ امکان تخمین این پارامترها، البته با خطای کمی بیشتر، وجود دارد.

یکی از مهم‌ترین پارامترهای توزیع خطا σ^2 است. از آمار ریاضی می‌دانیم که واریانس نمونه‌ای با فرمول $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ برای نمونه‌ای به حجم n از توزیع X برآوردگر نااریب واریانس جامعه است. از این رو، می‌توانیم واریانس نمونه‌ای $\widehat{E}$ را به عنوان برآوردگری از σ^2 در نظر بگیریم.

معمولا روش‌های مدل سازی برآوردگری برای تخمین واریانس D ارائه می‌دهند. برای مثال در رگرسیون آماره‌ی MSE را محاسبه می‌کنند. خوب است مقادیر به دست آمده از برآوردگرهای هر دو روش در گزارش ذکر شده و با یکدیگر مقایسه شوند.

۳.۳.۳ برآورد تابع توزیع خطا

n_x را تعداد اعضای مجموعه‌ی $\widehat{E}$ که کمتر یا مساوی x هستند تعریف می‌کنیم. تابع پله‌ای $\widehat{F}$ که برآوردی از تابع توزیع خطاست را به این شکل تعریف می‌کنیم :

$$\widehat{F}(x) = \frac{n_x}{n}$$

در اینجا n حجم نمونه و برابر تعداد اعضای $\widehat{E}$ است.

قضیه ۵.۳. *$\widehat{F}$ همواره بین صفر و یک و از راست پیوسته است.*

اثبات. برای هر $x \in \mathbb{R}$، n_x با توجه به نحوه‌ی تعریف آن در نابرابری $۰ < n_x < n$ صدق می‌کند. پس، برای هر x دلخواه داریم :

$$۰ \leq \widehat{F}(x) \leq ۱$$

مرتب شده‌ی اعضای مجموعه‌ی $\widehat{E}$ را به صورت زیر در نظر بگیرید :

$$\widehat{e}_{(۱)} \leq \widehat{e}_{(۲)} \leq \cdots \leq \widehat{e}_{(n)}$$

برای نقطه‌ی $a \in \mathbb{R}$ دلخواه دو حالت زیر را در نظر بگیرید :
حالت اول : a برابر یکی از اعضای $\widehat{E}$ باشد.
بنابراین $j \in \{۱, \ldots, n\}$ وجود دارد بقسمی که $a = \widehat{e}_{(j)}$.
اگر $j = n$ باشد، طبق تعریف تابع، در ازای تمام $x \geq a$ مقدار $\widehat{F}$ برابر ۱ است. پس،

$$\lim_{x \to a^+} \widehat{F}(x) = \widehat{F}(a) = ۱$$

اگر $j \neq n$ باشد، برای $۰ < \delta < e_{(j+۱)}$، با نزدیک شدن از راست به a در همسایگی با شعاع δ، n_x برابر n_a خواهد بود. از این رو :

$$\lim_{x \to a^+} \widehat{F}(x) = \lim_{x \to a^+} \frac{n_x}{n} = \frac{n_a}{n} = \widehat{F}(a)$$

حالت دوم : a برابر هیچ یک از اعضای $\widehat{E}$ نباشد

اگر $a > \widehat{e}_{(n)}$ یا $a < \widehat{e}_{(1)}$ اثبات مشابه حالت $j = n$ قسمت بالا می‌شود.
در غیر این صورت، $j \in \{1 \dots, n\}$ وجود دارد به طوری که $\widehat{e}_{(j)} < a < \widehat{e}_{(j+1)}$. در این حالت δ را طوری انتخاب می‌کنیم که در نابرابری $0 < \delta < \widehat{e}_{(j+1)} - a$ صدق کند. ادامه‌ی اثبات مشابه قبل خواهد بود.

□

قضیه‌ی بالا نشان می‌دهد که برآوردگر معرفی شده ویژگی‌های پراهمیت تابع توزیع خطا را حفظ می‌کند.

۴.۳ تنظیمات الگوریتم

همانطور که اشاره شد، با پذیرش هر یک از فرض‌های ۱۵.۲ و ۱۶.۲، خطای تقریب زدن رفتار g با تابع دلخواه را قابل اغماض فرض کردیم. تشخیص میزان این خطا برای ما دشوار است اما به علت آن که تعداد بخش‌ها با آن رابطه‌ی معکوس دارد، آنالیزگر می‌تواند با تعیین v مناسب دقت روش را افزایش دهد.

تعیین v، شکل دقیق افراز را مشخص نمی‌کند اما با روشی که در اثبات قضیه‌ی ۶.۲ بیان شد، می‌توانیم بخش‌ها را با تقسیم R_X به v بخش مساوی بسازیم.

اتفاق خوب آن است که تعداد داده‌ها در هر بخش تقریبا مشابه باشد. در غیر این صورت اگر بخش‌هایی باشند که تعداد مشاهداتشان خیلی کم باشد، می‌توانیم آن‌ها را در ادغام کردن لحاظ نکنیم و یا مرزهای تقسیم‌بندی را کمی جابه‌جا کنیم تا بخشی از داده‌های بخش‌های مجاور به این قسمت‌ها منتقل شود.

کم بودن تعداد مشاهدات هر بخش، دقت ما در برآورد g را تحت تاثیر قرار می‌دهد. اگر $\widehat{g}$ از واقعیت دور باشد، $\widehat{e}$های محاسبه شده نیز از مقدار صحیحشان فاصله می‌گیرند. این موضوع باعث می‌شود تا $\widehat{E}$ نماینده‌ی خوبی برای E نباشد و استنباط‌های انجام شده بر پایه‌ی این مجموعه کمتر قابل اعتماد خواهند بود.

هر یک از E_iها نمونه‌ی تصادفی از توزیع خطا به حجم n_i است. این موضوع نشان می‌دهد که باید برای هر $i, j \in \{1, \dots, v\}$ دلخواه، توزیع نمونه‌ای E_i با E_j یکسان باشد. یکی از دیگر راه‌هایی که می‌تواند در انتخاب v به آنالیزگر کمک کند، بررسی هم‌توزیعی $\binom{v}{2}$ انتخاب از ترکیب‌های دوتایی مجموعه‌های E_i و E_j است.

برخی از آزمون‌های ناپارامتری که برای سنجیدن هم‌توزیعی دو نمونه‌ی تصادفی استفاده می‌شوند

عبارتند از:

- kolmogorov-smirnov test
- chi-square test
- wald-wolfowitz Two-sample Run test
- Hollander test of extreme reactions

جزئیات تست‌های بالا در حیطه‌ی این نوشته نیست اما در منابع به صورت کامل آورده شده است. باید پیش از آن که آزمون مورد نظر را انتخاب کنیم، درمورد آماره‌ی آزمون و صورت دقیق فرض صفر و مقابل هر یک از موارد بالا تحقیق کنیم تا مناسب‌ترین گزینه را برگزینیم.

علاوه بر دو مورد ذکر شده، بررسی برابری پراکندگی دو به دوی E_iها نیز مفید خواهد بود. اگر قرار باشد تمامی آنها هم توزیع باشند، انتظار می‌رود که پراکندگی مشابهی نیز داشته باشند. پراکندگی لزوما σ^2 نیست و ممکن است معیارهای متفاوتی برای آن تعریف شود. آزمون‌های زیر در سنجیدن این ادعا کارایی دارند :

- Ansari-Bradley test
- Moses test

در آخر حالتی را در نظر بگیرید که بدانیم خطا از توزیع تبدیل یافته‌ی نمایی با پارامتر λ با میزان انتقال به اندازه‌ی $\frac{1}{\lambda}$ در عکس جهت مثبت محور پیروی می‌کند. این انتقال تضمین می‌کند که امید توزیع D برابر صفر باشد.

چگالی توزیع نمایی تنها بر روی مقادیر مثبت ناصفر است. بنابراین توزیع تبدیل یافته‌ی آن بر مجموعه‌ی $[-\frac{1}{\lambda}, +\infty)$ تعریف می‌شود. از این رو انتظار می‌رود اعضای مجموعه‌ای که به عنوان نماینده‌ی نمونه‌ی تصادفی خطا مشخص می‌کنیم، در محدوده‌ی ذکر شده باشند. در این شرایط اگر تعداد اعضایی که در مجموعه‌ی $\widehat{E}$ مقدارشان کمتر از $-\frac{1}{\lambda}$ است، زیاد باشد، به درست بودن انتخاب v شک می‌کنیم.

اگر λ را ندانیم، برای بررسی تعداد این اعضا، $\widehat{\lambda}$ محاسبه شده را جایگزین آن می‌کنیم. برای مثال در توزیع نمایی بالا، برآوردگر ماکسیمم درستنمایی از روابط زیر به دست می‌آید :

$$f_D(e) = \lambda e^{-\lambda(e+1/\lambda)}, \quad e \geq \frac{1}{\lambda}$$

$$L(\lambda) = \prod_i f_D(e_i) = \lambda^n \exp(-\lambda \sum_i e_i - n) I_{[-1/\lambda, +\infty)}$$

تذکر ۶.۳. *اگر امکان دسترسی به $\widehat{E}$ وجود داشت، حتی یک مشاهده که در محدوده‌ی $[-\frac{1}{\lambda}, +\infty)$ نباشد، نشان می‌دهد که اشتباهی صورت گرفته است.*

می‌دانیم که در مراحل ساختن $\widehat{E}$، نیاز است ابتدا فرض ثابت یا خطی بودن رفتار تابع g در هر بخش را بپذیریم. امکان آن که در کل این فرضیات برقرار نباشند وجود دارد.

به علاوه، برای حذف اثر g و تخمین e_iها نیاز است تابع g را در هر بخش برآورد کنیم. این تخمین خطا نیز به همراه دارد.

برای همین ممکن است هر یک از $\widehat{e}_i$ها با مقدار واقعی‌شان کمی متفاوت باشند. پس اگر تعداد کمی از مشاهدات $\widehat{E}$ نسبت به حجم کل نمونه از محدوده‌ی بالا خارج باشند زیاد جای نگرانی نیست. برقرار نبودن این شرایط ممکن است تنها ناشی از خطاهایی باشد که در دو مرحله مرتکب می‌شویم تا $\widehat{E}$ را به دست آوریم.

اما اگر تعداد زیادی از اعضای $\widehat{E}$ خارج بازه‌ی مشخص شده بوده یا با $-1/\lambda$ تفاوت قابل توجهی داشته باشند، نشان از آن است که با احتمال بالایی باید v را تغییر دهیم تا خطایی که در مفروضات الگوریتم قابل اغماض فرض کردیم کاهش یابد.

خوب است آنالیزگر تمام تست‌های بالا را برای انتخاب‌های متفاوت v اجرا نماید. با در نظر گرفتن پی‌ـ‌مقدارهای این آزمون‌ها و حجم نمونه در هر بخش، می‌تواند کوچک‌ترین v مناسب را برگزیند.

۵.۳ کاربرد الگوریتم

گوشه‌ای از کاربردهای این الگوریتم به صورت تیتروار در زیر آورده شده است :

۱. تعیین شکل تقریبی توزیع خطا
۲. امکان بررسی تقارن و چولگی خطا
۳. امکان آزمودن هم‌توزیعی خطا با توزیع‌های دلخواه
۴. برآورد پارامترهای توزیع خطا
۵. برآورد واریانس خطا
۶. امکان برآورد تابع توزیع احتمال D
۷. استفاده از $\widehat{F}$ به عنوان یکی از ورودی‌های الگوریتم برچسب‌گذاری هنگامی که توزیع D برای

آنالیزگر ناشناخته باشد.

۸. فراهم شدن امکان تغییر توزیع خطا به توزیع دلخواه زمانی که D در مسئله مشخص نشده است.

فصل ۴

تبدیل توزیع خطا

ممکن است نوع توزیع خطا یا برخی از ویژگی‌های آن برای آنالیزگر مهم باشند. برای مثال، روش‌هایی هستند که نسبت به انحراف از تقارن حساسیت نشان می‌دهند. در رگرسیون و مدل‌های طرح آزمایش نیز فرض نرمال بودن خطا حائز اهمیت است.
واضح است که الزامی وجود ندارد که توزیع خطا ویژگی‌های مورد علاقه‌ی ما را داشته باشد. برای حل این مشکل راهکاری در ادامه معرفی می‌شود که می‌تواند داده‌ها را تغییر دهد به نحوی که در مجموعه‌ی داده‌ی جدید، خطا از توزیع مشخصی پیروی کند و یا خصوصیت‌های مورد نظر آنالیزگر را دارا باشد.

۱.۴ مراحل الگوریتم

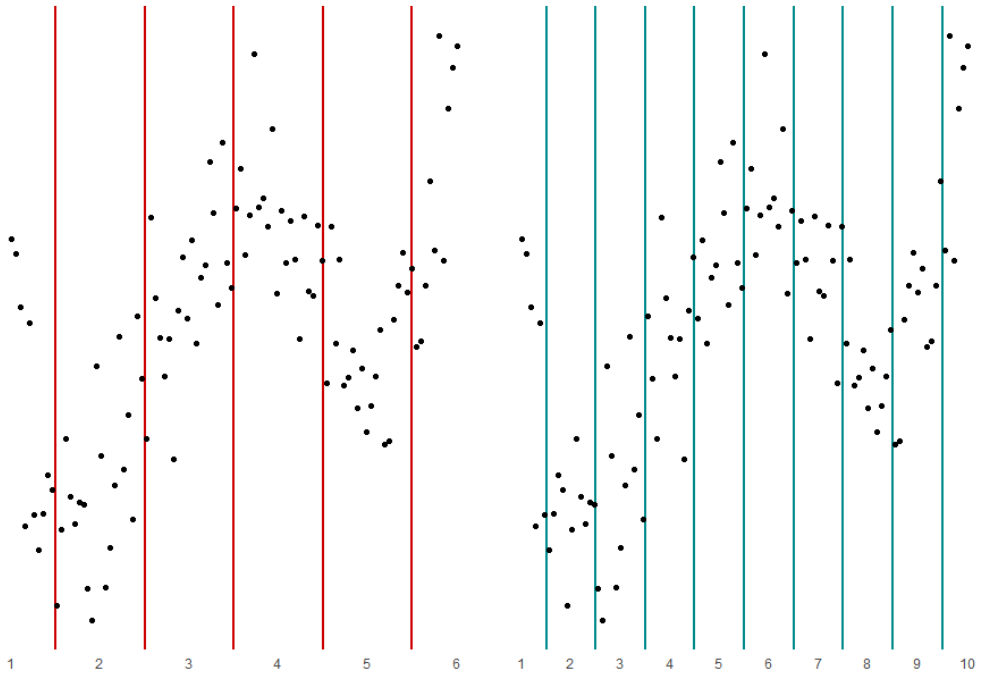
۱. انتخاب v مناسب :

در فصل قبل برای تعیین تعداد بخش‌ها فرایندی بر پایه‌ی تعداد داده‌های هر قسمت و آزمون فرض‌های مناسب معرفی شد. هدف آن است که این تقسیم بندی طوری انجام شود که توزیع نمونه‌ای داده‌های هر بخش مشابه بوده و مقدار پارامترهای آن‌ها، مانند پراکندگی، تفاوت معناداری با یکدیگر نداشته باشند. در این صورت است که می‌توانیم $\widehat{E}$ را جانشین قابل قبولی برای E بدانیم.
در مثال زیر تشخیص v مناسب را بر مبنای تحلیل نمودار پراکنش توضیح می‌دهیم :

مثال ۱.۴.

در شکل ۱.۴ تقسیم بندی سمت راست برای $v = ۱۰$ را به نمودار سمت چپ با $v = ۶$ ترجیح

می‌دهیم. درست است که در هر دو انتخاب v، بخش‌ها به لحاظ تعداد داده‌ها شبیه یکدیگر هستند اما به صورت بصری و پیش از آن که آزمون فرض انجام دهیم، قابل تشخصی است که نمی‌توانیم در تمام بخش‌های نمودار سمت چپ، رفتار g را خطی یا ثابت معرفی کنیم. به عنوا مثال در بخش‌های اول و پنجم این موضوع روشن‌تر است. به علاوه در تقسیم بندی سمت راست پراکندگی داده‌ها به هم شبیه‌تر است.



شکل ۱.۴: انتخاب تعداد بخش‌ها در داده‌هایی که نسبت به پیشگوها تقریبا همگن باشند.

۲. ایجاد تقسیم بندی :

پس از آن که تعداد بخش‌ها را مشخص کردیم، R_X را به v بخش مساوی تقسیم می‌کنیم. ممکن است داده‌ها به صورت همگن نسبت به متغیرهای پیشگو جمع آوری نشده باشند و بخش‌هایی باشند که تعداد مشاهداتی که در آن‌ها قرار می‌گیرند در مقایسه با سایر قسمت‌ها خیلی کم است. حتی احتمال آن که هیچ داده‌های در محدوده‌ی برخی از بخش‌ها قرار نگیرند وجود دارد. در این حالت توصیه می‌شود با توجه به مسئله مدل سازی، از اطلاعات این بخش‌ها چشم‌پوشی کنیم یا با جابه‌جایی مرزهای تقسیم بندی شرایط را بهتر کنیم.

در صورتی که تراکم نقاط نسبت به محور x همگن باشد، بهتر است در تقسیم بندی بخش‌هایی

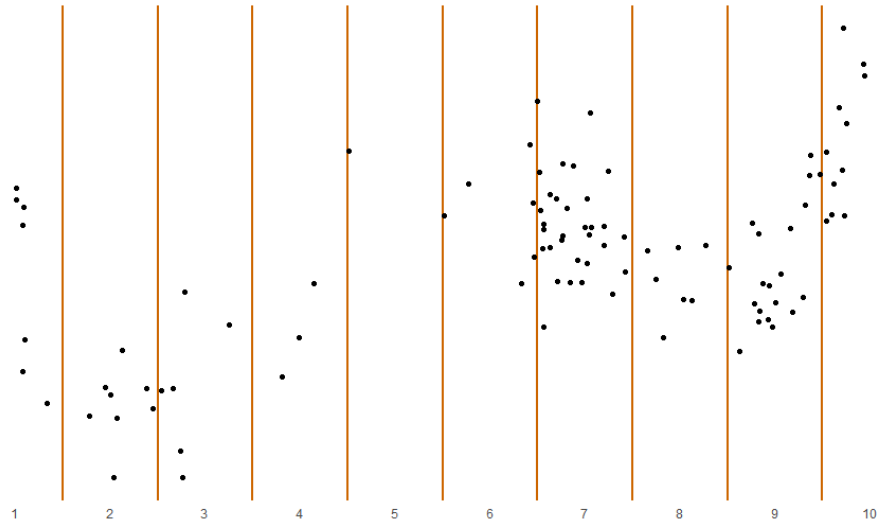
با طول برابر ایجاد کنیم. در این حالت تعداد مشاهدات در هر بخش تقریبا با بخش‌های دیگر برابر می‌شود.

اگر داده‌ها نسبت به محور x همگن نباشد، ممکن است تفاوت تعداد مشاهدات هر بخش زیاد باشد. در این حالت می‌توانیم کمی مرزهای تقسیم‌بندی را جابه‌جا کنیم تا مشکل برطرف گردد. اگر این راه کارساز نبود و بعد از اعمال تغییرات همچنان بخش‌هایی با حجم نمونه‌ی کم وجود داشتند می‌توانیم این قسمت‌ها را از مراحل تشکیل $\widehat{E}$ حذف کنیم.

برای روشن‌تر شدن این موضوع مثال زیر آورده شده است :

مثال ۲.۴.

در شکل ۲.۴ برای $v = ۱۰$ بخش‌های ۴ و ۵ به ترتیب شامل ۳ و ۱ داده می‌شوند. این در حالیت که تعداد مشاهدات در قسمت ۷ام برابر ۳۲ است. در این حالت می‌توانیم از ۴ داده در بخش‌های ۴ و ۵ صرف نظر کنیم و $\widehat{E}$ را از اطلاعات ۹ بخش دیگر بسازیم.



شکل ۲.۴: انتخاب تعداد بخش‌ها در داده‌هایی که نسبت به پیشگوها تقریبا همگن باشند.

شاید بد نباشد که آنالیزگر در این مثال مرز میان قسمت‌های ۶ و ۷ را کمی به راست ببرد و این بخش ر به دو قسمت تقسیم کند. البته اگر نتایج آزمون فرض‌ها این تغییر را بی خطر بدانند.

۳. تشکیل $\widehat{E}_i$ها برای هر بخش :

این مرحله از الگوریتم در فصل قبل به طور کامل شرح داده شده است. جا دارد در این قسمت به تذکر زیر توجه شود :

تذکر ۳.۴. *همانطور که گفتیم برای تعیین اعضای $\widehat{E}_i$ لازم است اثر g را در هر بخش از مشاهدات حذف کنیم. برای این کار ابتدا لازم است برآوردگر مناسب $\widehat{g}$ را تشکیل دهیم. این برآوردگر تحت فرض‌های ۱۵.۲ و ۱۶.۲ به ترتیب از تخمین پارامترهای شیب و عرض از مبدا و ثابت‌های c در هر بخش به دست می‌آید. این که کدام فرض را مبنای محاسبات خود قرار دهیم، از روی نمودار پراکنش پس از تقسیم‌بندی‌ها قابل تشخیص است.*
حتی امکان آن وجود دارد که برای برای بخش‌های متفاوت، فرضی که برای داده‌های آن قسمت مناسب‌تر است را انتخاب کنیم.

برای مثال در شکل ۱.۴ فرض خطی بودن تابع در تقریبا تمام بخش‌ها منطقی‌تر به نظر می‌رسد. در شکل ۲.۴ می‌توانیم برای بخش‌های ۲، ۳ و ۸ فرض ثابت بودن تابع و برای سایر بخش‌ها فرض خطی بودن را در نظر بگیریم.

۴. تشخیص توزیع خطا :

اگر در مسئله مدل سازی نوع تابع توزیع خطا داده شده باشد، تنها کافیست پارامترهای آن برآورد شوند. برای این کار می‌توانیم از $\widehat{E}$ به عنوان نمونه‌ی تصادفی از توزیع خطا استفاده کنیم.
در صورتی که هیچ اطلاعاتی از توزیع خطا در دسترس نباشد، بهترین گزینه‌ی آنالیزگر استفاده از $\widehat{F}$ است. با فرایندی که در بخش ۳.۳.۳ شرح دادیم، با بهره‌گیری از الگوریتم فصل قبل، نوع توزیع D و پارامترهای آن قابل حدس خواهند بود.

۵. تشخیص تبدیل مناسب :

توزیع خطا و توزیعی که می‌خواهیم داشته باشد نوع تبدیل را مشخص خواهد کرد. باید به این نکته توجه داشت که توزیع مورد علاقه‌ی ما یا تبدیلی که انتخاب می‌کنیم، ویژگی‌های کلی خطا را تحت تاثیر قرار ندهد. برای مثال، توزیع خطا می‌بایست پیوسته و میانگین آن باید همواره صفر باشد.

از طرفی باید توجه داشت که اگر بر روی دامنه‌ی توزیع خطا محدودیت باشد، مانند توزیع یکنواخت، نمایی، گاما و بتا، توزیع جدید نیز محدودیت‌های مشابه داشته باشد.
برای مثال توزیع نرمال را در نظر بگیرید. اعدادی که به صورت تصادفی از این توزیع تولید می‌شوند، از مجموعه‌ی $\mathbb{R}$ می‌آیند. از این رو می‌توانند هر عدد دلخواه و بسیار از صفر دور باشند. واضح است که likelihood مشاهده‌ی این اعداد با هم متفاوت بوده و بسیار غیر محتمل است که مشاهده‌ای که ثبت می‌شود در بعضی بازه‌ها قرار بگیرد. اما در تئوری برای تبدیل کردن توزیعی به توزیع دیگر باید برای تمام این مقادیر نقطه‌ی متناظری در نظر گرفته شود، هر چند likelihood مشاهده کردن

آنها بسیار کم باشد.
مثلا، این که داده‌های تولید شده از توزیعی مانند نرمال را به توزیع نمایی با محدوده مقادیر $\mathbb{R}^+$ تبدیل کنیم، کمی غیر منطقی بوده و توصیه نمی‌شود.

تعریف ۴.۴. *E' را مجموعه‌ی n عضوی که شامل تبدیل یافته‌های اعضای $\widehat{E}$ است در نظر بگیرید. این مجموعه قرار است در قالب نمونه‌ی تصادفی از توزیع دلخواه باشد.*

اعضای آن را با $e'_1, \ldots, e'_n$ نمایش می‌دهیم. مشابه تعریف E و $\widehat{E}$، این مجموعه اجتماع v مجموعه‌ای است که در هر یک مقادیر تبدیل یافته‌ی $\widehat{E}_i$ها قرار دارند.

در ادامه بعضی از تبدیل‌هایی که متداول‌تر هستند بررسی می‌شوند :

تبدیل D به توزیع تبدیل‌یافته‌ی آن با انتقال به میزان $d > 0$ و تغییر مقیاس به اندازه‌ی ۱

هر نوع تغییر مکان با ثابت نگه‌داشتن سایر ویژگی‌ها، میانگین توزیع جدید را تغییر می‌دهد. به علت آن که میانگین توزیع خطا باید صفر باشد، این کار توصیه نمی‌شود.

تبدیل D به توزیع تبدیل‌یافته‌ی آن با انتقال به میزان ۰ و تغییر مقیاس به اندازه‌ی $a > 0$

توزیع اولیه $D(0, \sigma^2)$ و توزیع جدید $D(0, (a\sigma)^2)$ است. برای هر بخش اعضای E' را به این شکل می‌سازیم :

$$e'_{i_j} = a\widehat{e}_{i_j}$$

برای $i = 1, \ldots, v$ و $j = 1, \ldots, n_i$.

تغییر نوع توزیع

این فرایند زمانی مجاز است که محدودیت‌های دو توزیع مشکل‌ساز نشوند.

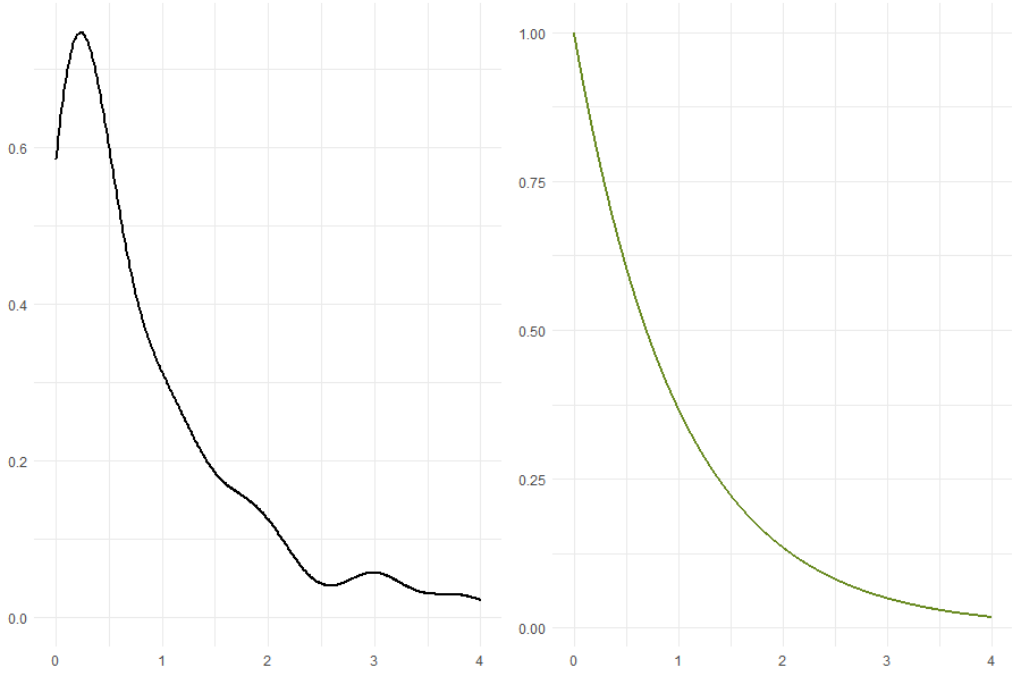
توزیع دلخواه جدید را با D' و تابع توزیع آن را با F' نمایش می‌دهیم. برای تشکیل E' از روی $\widehat{E}$ به روش زیر عمل می‌کنیم :

$$e'_i = F'^{-1}(F(\widehat{e}_i))$$

این تبدیل با استدلال زیر میان اعضای دو مجموعه تناظر برقرار می‌کند :
ابتدا احتمال $e \leq \widehat{e}_i$ را با به کارگیری تابع F محاسبه می‌کند. سپس، e'_i را برابر عددی قرار می‌دهد که باعث می‌شود در توزیع جدید F' برابر این احتمال شود.
در صورتی که F را نداشته باشیم، می‌توانیم $\widehat{F}$ را در رابطه‌ی بالا جایگزین کنیم. اگر نوع توزیع خطا مشخص باشد ولی پارامترهای آن را ندانیم، با برآورد آن‌ها به نوعی F را تخمین می‌زنیم.

مثال ۵.۴.

در این مثال نمونه‌ی تصادفی از توزیع نمایی با پارامتر $\lambda = ۱$ را به توزیع گاما با پارامتر شکل $\alpha = ۲$ و مقیاس $\beta = ۰/۵$ تبدیل می‌کنیم. نمودار سمت راست در شکل ۳.۴ تابع چگالی $E(۱)$ و نمودار سمت چپ نمودار چگالی نمونه‌ی تصادفی را مشخص کرده است.



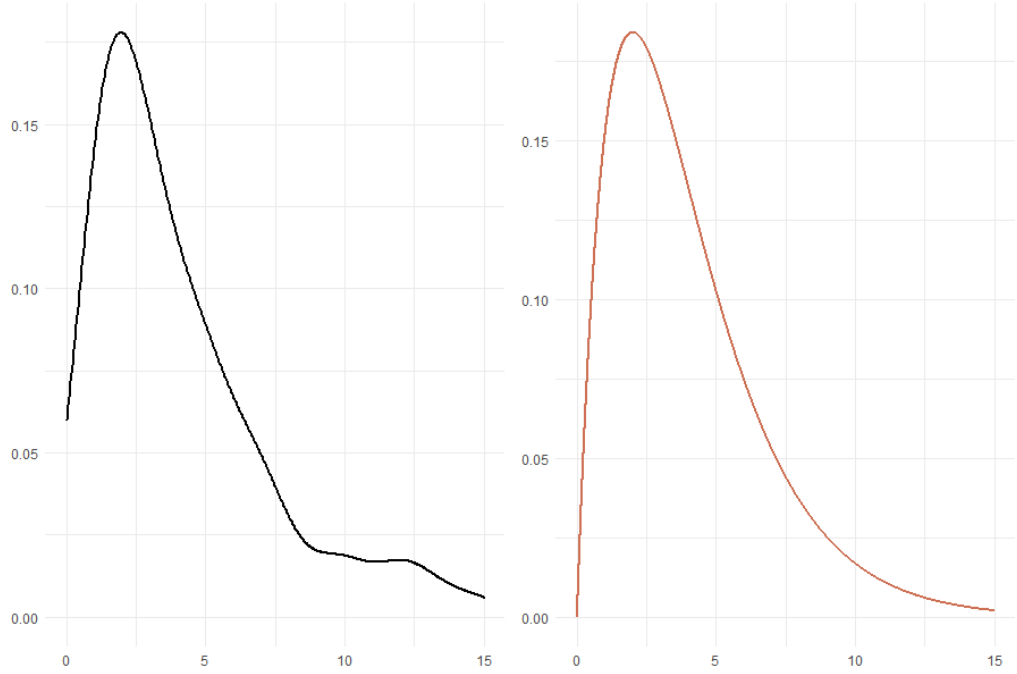
شکل ۳.۴: نمودارهای چگالی اصلی و نمونه‌ای

در نرم افزار R نقاط تبدیل یافته در ازای عضو دلخواه x از نمونه، به این صورت محاسبه شده‌اند :

$$\text{qgamma(pexp(x,rate=1),shape=2,scale=0.5)}$$

در شکل ۴.۴ ، شباهت تابع چگالی نمونه‌ی تبدیل یافته به چگالی توزیع $\text{gamma}(\alpha = ۲, \beta = ۰/۵)$ ملاحظه می‌شود.

علاوه بر آن آزمون کولموگروف–اسمیرنوف برای آزمون کردن فرض هم توزیعی داده‌های تغییر یافته با توزیع گامای متناظر انجام شد. پی–مقدار این آزمون ۰/۶۴۴۶ محاسبه شده که نشان می‌دهد شواهد معناداری بر علیه فرض صفر، در سطح خطای $\alpha = ۰/۰۵$ وجود ندارد.



شکل ۴.۴: نمودارهای چگالی اصلی و نمونه‌ای پس از تبدیل

۶. تغییر مشاهدات و ایجاد مجموعه داده‌ی جدید :

در مرحله‌ی آخر تبدیل مناسب به روی داده‌ها اعمال می‌شود. پس از آن، آنالیزگر می‌تواند مدل مناسب را به روی داده‌ها برازش دهد.

۱.۱.۴ تفاوت روش تغییر توزیع و تغییر مقیاس

یکی از کاربردهای روش تغییر مقیاس کاهش واریانس خطا در نمودار پراکنش است. زمانی که پراکندگی داده‌ها حول g زیاد باشد، امکان تشخیص صحیح فرم کلی آن با استناد بر نمودار وجود ندارد.

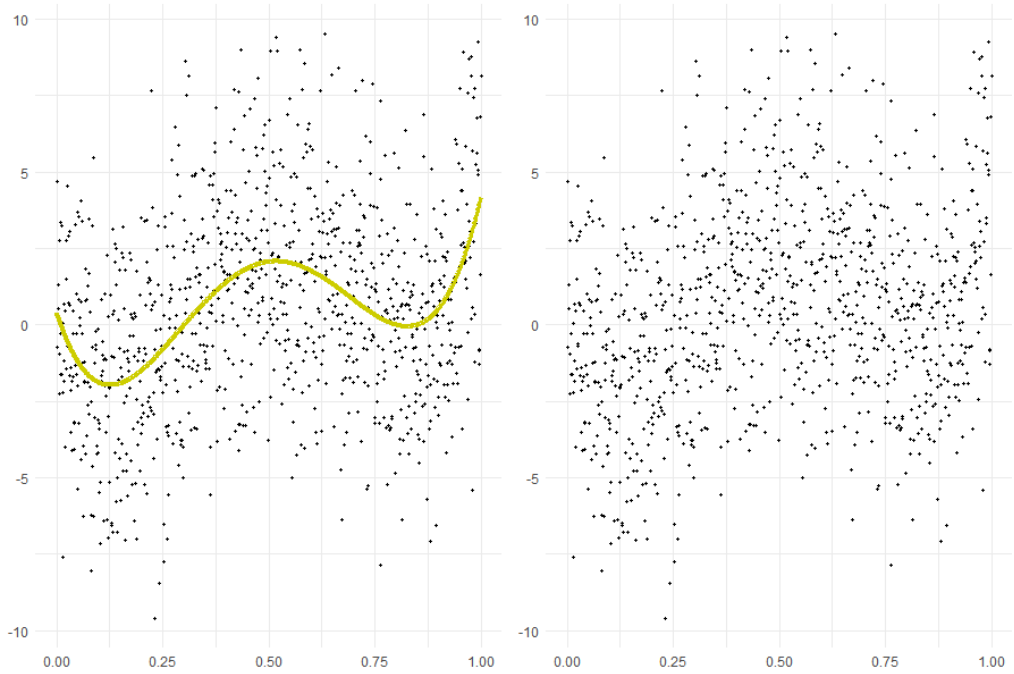
این روش زیرمجموعه‌ای از روش تغییر توزیع خطاست. از این رو، در این شرایط هر دوی آن‌ها قابل استفاده هستند. اما روش اول، زمانی که F و F' ناشناخته است، برتری دارد. در این حالت، یک مرحله از تقریب زدن به هنگام برآورد F، که خطا به همراه دارد، کمتر می‌شود.

مثال ۶.۴.

نمونه‌ای از کاربرد این روش در شکل ۵.۴ آورده شده است. در این مجموعه داده، R_X بازه‌ی بسته‌ی $[۰, ۱]$ و تابع g چندجمله‌ای از درجه ۴ بوده و به صورت زیر تعریف شده است :

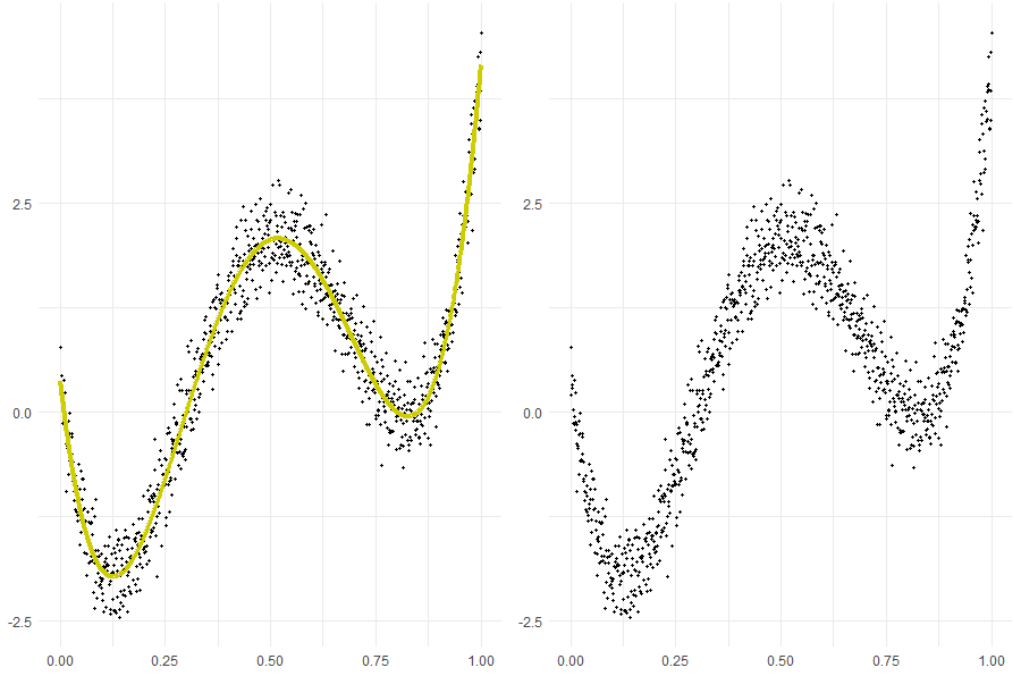
$$g(x) = ۲۰۰(x - ۰/۰۱)(x - ۰/۳)(x - ۰/۸)(x - ۰/۸۵)$$

ملاحظه می‌کنیم که شکل تابع g به علت بزرگ بودن واریانس در نمودار پراکنش قابل تشخیص نیست.



شکل ۵.۴: غیر قابل تشخیص بودن تابع g در نمودار

پس از بررسی‌های اولیه نمودار پراکنش، توزیع خطا را به توزیع تبدیل یافته‌ی آن با تغییر مقیاس به میزان ۰/۱ تغییر داده و نمودار داده‌های جدید را مجددا رسم می‌کنیم. در مجموعه داده‌های جدید، تشخیص فرم کلی g ساده‌تر است.



شکل ۶.۴: نمودار پراکنش برای مجموعه داده‌ی جدید

در شکل ۵.۴ محدوده‌ی مقادیر محور y از ۱۰− تا ۱۰ است. این مقدار در شکل ۶.۴ بین ۲/۵− تا تقریبا ۵ است. به این علت، کشیدگی تابع در نمودار اول و دوم کمی متفاوت است . این بازه برای بهتر دیده شدن داده‌ها در شکل در نظر گرفته شده است.

فصل ۵

الگوریتم برچسب‌گذاری

در این الگوریتم هدف آن است که دورترین نقاط نسبت به تابع g را پیدا کنیم.

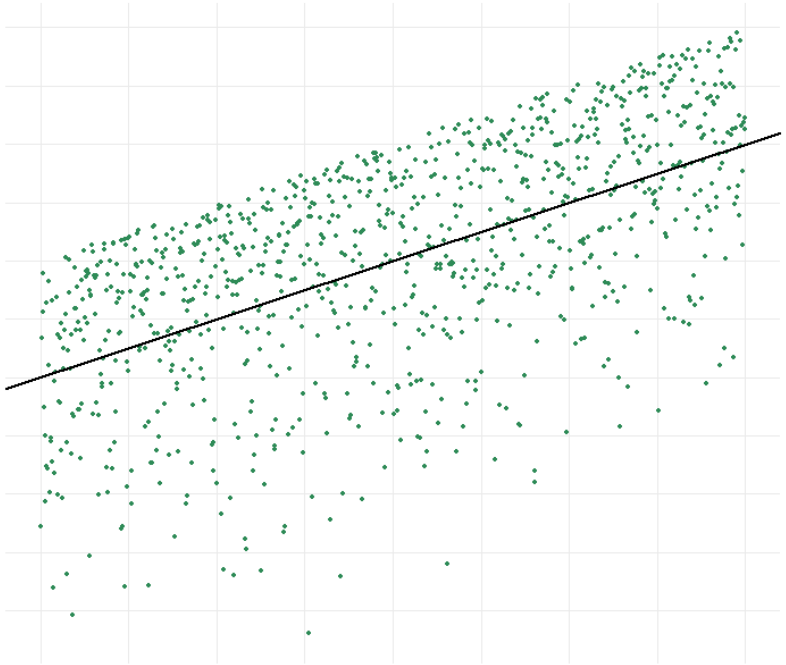
اگر خطا وجود نداشته باشد، برای هر نقطه دلخواه x از محدوده‌ی مقادیر متغیر پیشگو، دقیقا مقدار $g(x)$ را ثبت می‌کنیم. در واقعیت، خطا در قالب یک متغیر تصادفی وجود دارد و در ازای مشاهده‌ی iام، یک نمونه‌ی تصادفی از توزیع D با نمادگذاری e_i تولید می‌شود. این موضوع باعث می‌شود که مشاهده‌ی ما از متغیر پاسخ به اندازه‌ی e_i از $g(x_i)$ دور شود. این که مشاهده چقدر از تابع فاصله می‌گیرد به صورت تصادفی تعیین شده و به اندازه‌ی e_i مربوط می‌شود.

توزیع D نحوه‌ی تولید نمونه‌ی تصادفی از خطا را مشخص می‌کند. با بزرگ شدن σ^2 امکان مشاهده‌ی اعداد بزرگتر برای e_iها افزایش می‌یابد. برای همین، نقاط با احتمال بیشتری از g فاصله‌ی بیشتری نسبت به قبل می‌گیرند.

در صورتی که D قرینه نبوده یا چوله باشد، احتمال آن که برای یک فاصله‌ی معین، نقاط بالای تابع قرار بگیرند (علامت e_i متناظر مثبت باشد) با احتمال آن که زیر منحنی باشند، متفاوت خواهد بود.

برای مثال در شکل ۱.۵، g تابع خطی است و خطا از توزیع تبدیل یافته‌ی بتا با پارامترهای $\alpha = ۳, \beta = ۱$ و انتقال به میزان ۳/۴ و تغییر مقیاس به اندازه‌ی ۴ پیروی می‌کند. این توزیع چوله به چپ است. برای همین، میانه‌ی آن از میانگین توزیع فاصله دارد. این اتفاق باعث می‌شود که احتمال منفی بودن نمونه‌ی تولید شده از توزیع خطا، با احتمال مثبت بودن آن یکسان نباشد.

این موضوع در شکل واضح‌تر دیده می‌شود.



شکل ۱.۵: تاثیر چوله بودن توزیع D بر نمودار پراکنش

بهتر است تعریف درستی از *دور بودن* بیان کنیم. این کار به دو روش ممکن است. بسته به آن که آنالیزگر تنظیمات الگوریتم را چگونه انتخاب می‌کند، تعریف متناظر با آن را در نظر می‌گیریم.

تعریف ۱.۵. *دور بودن با معیار چندک‌های توزیع خطا :*
برای α مشخص، داده‌ی jام را نسبت به تابع در نقطه‌ی x_j دور می‌نامیم هرگاه،

$$e_j > \textit{quantile}_D(۱ - \alpha) \quad \text{یا} \quad e_j < \textit{quantile}_D(\alpha)$$

در اینجا، خروجی $\text{quantile}_D(z)$ عددی است که در رابطه‌ی زیر صدق کند :

$$F(\text{quantile}_D(z)) = P(D \leq \text{quantile}_D(z)) = z$$

تعریف ۲.۵. *دور بودن با معیار فاصله :* *برای $d > ۰$ معین، داده‌ی jام را نسبت به منحنی در نقطه‌ی $g(x_j)$ دور می‌نامیم هرگاه،*

$$|e_j| > d$$

عملکرد این روش بر پایه‌ی قضیه‌ی زیر است:

قضیه ۳.۵. *اگر X متغیر تصادفی دلخواه با تابع توزیع احتمال F باشد آنگاه،*

$$F(X) \sim U(۰, ۱)$$

در اینجا منظور از $U(۰, ۱)$ توزیع یکنواخت پیوسته بر روی بازه‌ی $[۰, ۱]$ است.

اثبات. برای اثبات هم‌توزیعی از روش تابع توزیع استفاده می‌کنیم. توجه شود که اثبات وجود معکوس CDF [۱] برای توزیع‌های پیوسته با دامنه‌ی $(۰, ۱)$ در منابع قابل ملاحظه است. قرار دهید $Y = F(X)$. تابع توزیع Y را با F_Y نشان می‌دهیم. برای $y \neq ۰, ۱$ داریم :

$$F_Y(y) = P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-۱}(y)) = F(F^{-۱}(y)) = y$$

طبق تعریف تابع توزیع، به سادگی می‌توانیم ببینیم که حکم برای مقدارهای $y = ۱$ و $y = ۰$ برقرار است.
از آنجا که تابع توزیع F_Y با تابع توزیع یکنواخت یکسان است، حکم قضیه نتیجه می‌شود.
□

۱.۵ مراحل الگوریتم

۱. انتخاب v مناسب

۲. ایجاد تقسیم بندی

دو مرحله‌ی اول طبق توضیحات فصل‌های قبل انجام می‌شوند.

۳. همگن سازی پراکندگی داده‌ها در هر بخش و پیداکردن داده‌های دور از منحنی

در صورتی که معیار ۱.۵ را برای تشخیص دور بودن نقاط انتخاب کنیم، همگن کردن نقاط سومین مرحله در اجرای الگوریتم خواهد بود. هدف آن تشکیل داده‌های همگن پس از حذف اثر g است.

ابتدا برآورد g در هر بخش، تحت فرض‌های درست مشخص می‌شود. سپس، $\widehat{e}$ها مشابه توضیحات قبلی در هر یک از v بخش به دست می‌آیند. در ادامه با بهره‌گیری از قضیه‌ی ۳.۵ برای هر یک از نقاط مجموعه‌ی $\widehat{E}$، مقدار $F(\widehat{e})$ یا $\widehat{F}(\widehat{e})$ را محاسبه می‌کنیم. این اعداد جدید،

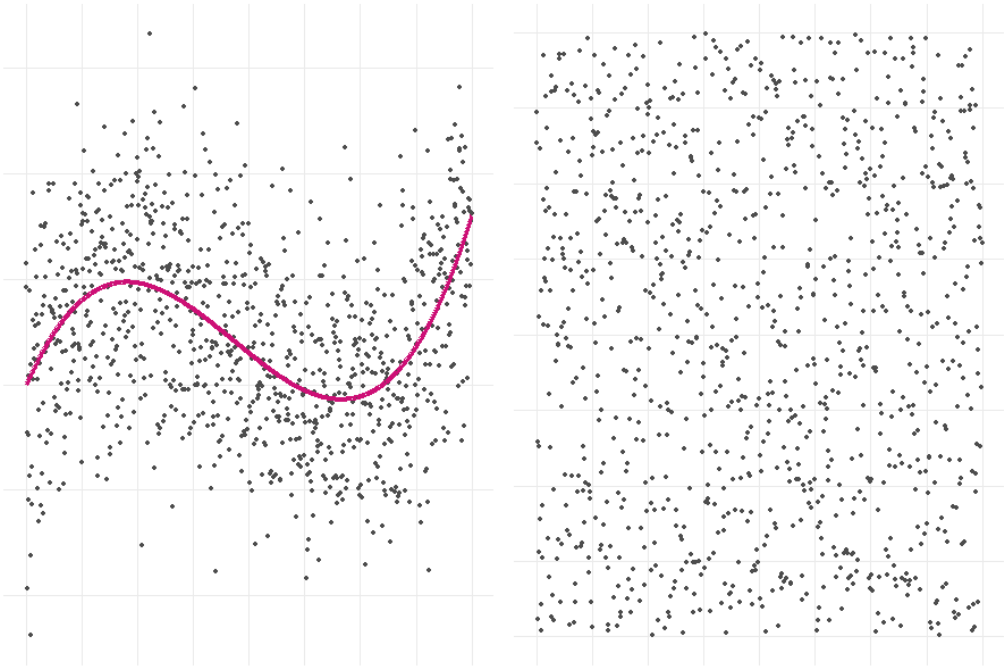
[۱] Cumulative Distribution Function

نمونه‌ای تصادفی از توزیع $U(۰, ۱)$ هستند.
اکنون بالاترین و پایین‌ترین نقاط در نمودار پراکنش، نقاط دور مورد نظر خواهند بود. به بالاترین و پایین‌ترین نقاط به ترتیب برچسب‌های "بالا" و "پایین" را منسوب می‌کنیم. به سایر نقاط برچسب "وسط" داده می‌شود.

این که نقطه تا چه اندازه بالا یا پایین باشد تا برچسبی غیر از "وسط" داشته باشد، از روی سطح α در تعریف ۱.۵ تعیین می‌شود. برای مثال در سطح $\alpha = ۰/۰۵$ نقاطی که $F(\widehat{e})$ یا $\widehat{F}(\widehat{e})$شان بیشتر از ۰/۹۵ و کمتر از ۰/۰۵ باشد، به ترتیب برچسب‌های "بالا" و "پایین" دریافت می‌کنند.

مثال ۴.۵.

در این مثال نمودار پراکنش قبل و بعد از همگن سازی رسم شده است. توزیع خطا نرمال با واریانس ۰/۰۴ و تابع $g(x) = x(x - ۰/۶)(x - ۰/۸) + ۱$ است. به علاوه، $R_X = [۰, ۱]$.
نمودار سمت راست همگن شده‌ی داده‌های سمت چپ را نشان می‌دهد.



شکل ۲.۵: همگن سازی نمودار پراکنش

اگر می‌خواهیم بر اساس تعریف ۲.۵ عمل کنیم، مجموعه‌ی $\widehat{E}$ را ساخته و اعضایی که قدر مطلقشان از $d > ۰$ تعیین شده، بیشتر است را شناسایی می‌کنیم. در صورتی که علامت $\widehat{e}$ متناظر

با آن‌ها مثبت یود، برچسب "بالا" و در غیر این صورت برچسب "پایین" را برای نقاط متناظرشان در نظر می‌گیریم. برچسب سایر نقاط "وسط" خواهد بود.

۲.۵ استفاده از اطلاعات برچسب‌ها

۱.۲.۵ شناسایی و حذف داده‌های پرت

اگر اندازه‌ی α یا d طوری باشد که برچسب‌های بالا و پایین داده‌های پرت را مشخص کنند، آنالیزگر می‌تواند این داده‌ها را حذف کند. این کار زمانی که حجم داده بزرگ است توصیه می‌شود اما زمانی که n کوچک باشد با حذف داده‌ها اطلاعات زیادی از دست می‌رود.
نقاط پرت خطا در آنالیزهای آماری و برآوردها را افزایش می‌دهد. برخی از روش‌های آماری به این دست از داده‌ها حساسیت زیادی نشان می‌دهند و بودن و نبودن آن‌ها تغییرات بزرگی در نتایج ایجاد می‌کند.

۲.۲.۵ تعیین قرینگی یا میزان چولگی D

اگر با معیار ۲.۵ برای برچسب‌های "بالا" و "پایین" تنظیمات یکسان در نظر بگیریم، مشابه بودن تعداد نقاط در هر دسته ، تا حدی قرینه بودن توزیع D را نشان می‌دهد. اگر تعداد نقاط با برچسب بالا با گروه دیگر بسیار متفاوت باشد، احتمالا توزیع خطا چوله یا نامتقارن است.

تذکر ۵.۵. *اگر معیار چندک‌ها برای میزن دوربودن نقاط از g انتخاب شود و α را برای نقاط بالا و پایین منحنی یکسان بگیریم، بدون توجه به نوع توزیع خطا انتظار می‌رود که تعداد نقاط در دسته‌های "بالا" و "پایین" تقریبا برابر باشند.*

علت این موضع آن است که در این معیار با تنظیمات یکسان، احتمال آن که نقطه‌ای برچسبی غیر از "وسط" بگیرد، برابر αاست. تفاوت‌های معنی‌دار در تعداد این برچسب‌ها در هر بخش و در کل داده، نشان از آن است که در جایی از تنظیمات و یا اجرای الگوریتم اشتباه عمل کرده‌ایم.

۳.۲.۵ جابه‌جایی داده‌های پرت

آنالیزگر می‌تواند به جای حذف داده‌ها، نقاط با برچسب‌های "بالا" و "پایین" را طوری جابه‌جا کند که به مقدار واقعی g نزدیک‌تر شوند. برای مثال می‌تواند تابعی از برآورد واریانس و نوع برچسب

طوری تعریف کند که کسر مشخصی از $\widehat{\sigma}$ را از نقاط "بالا" کم و کسر دیگری از آن را به نقاط "پایین" اضافه کند.

این که چه تغییراتی بر هر دسته اعمال شود و ضابطه‌ی h بهتر است چه باشد، جای بحث دارد.

۳.۵ تنظیمات

۱.۳.۵ تعداد بخش‌ها

همانطور که پیشتر نیز اشاره شد، مشخص کردن v بخشی از تنظیمات هر سه الگوریتم است. هر چه v بیشتر باشد، دقت الگوریتم افزایش می‌یابد اما حجم نمونه در هر بخش کم می‌شود. کم بودن حجم نمونه در استنباط‌ها و برآوردها ایجاد خطا می‌کند. پیشنهاد می‌شود تعداد بخش‌ها با رسم نمودارها پراکنش تعیین شوند. سپس با آزمون فرض‌های مشخص شده در فصل‌های قبل به بررسی گزینه‌ها پرداخته شود.

۲.۳.۵ اندازه‌ی α یا d

در هر یک از دو معیار معرفی شده برای دور بودن نقاط، نیاز است تا مقدار α یا d مشخص شود. الگوریتم در ازای مقادیر بزرگ آن‌ها، نقاطی که فاصله‌ی زیادی از منحنی دارند را پیدا می‌کند. هر چه این مقادیر را کمتر در نظر بگیریم، تعداد نقاط برچسب‌دار بیشتر و به تابع g نزدیک‌ترند. آنالیزگر می‌تواند به دلخواه و بر حسب نیاز مسئله مدل سازی، این اعداد را مشخص نماید. خوب است که مقادیر مختلف مورد آزمایش قرار بگیرند و نقاط برچسب‌دار در نمودار بررسی شوند.

آنالیزگر می‌تواند از مقادیر متفاوت α یا d برای برچسب‌های "بالا" و "پایین" استفاده کند. کاربرد آن بیشتر در توزیع‌های چوله است.

۳.۳.۵ تعداد اجراها

اگر تغییری روی داده‌ها انجام نگیرد، هر بار اجرای این الگوریتم نتایج یکسانی را ایجاد می‌کند. در صورتی که داده‌های پرت حذف شوند یا به وسیله‌ی تابعی جابه‌جا شوند، اجرای مجدد الگوریتم برچسب‌های جدیدی تولید می‌کند.

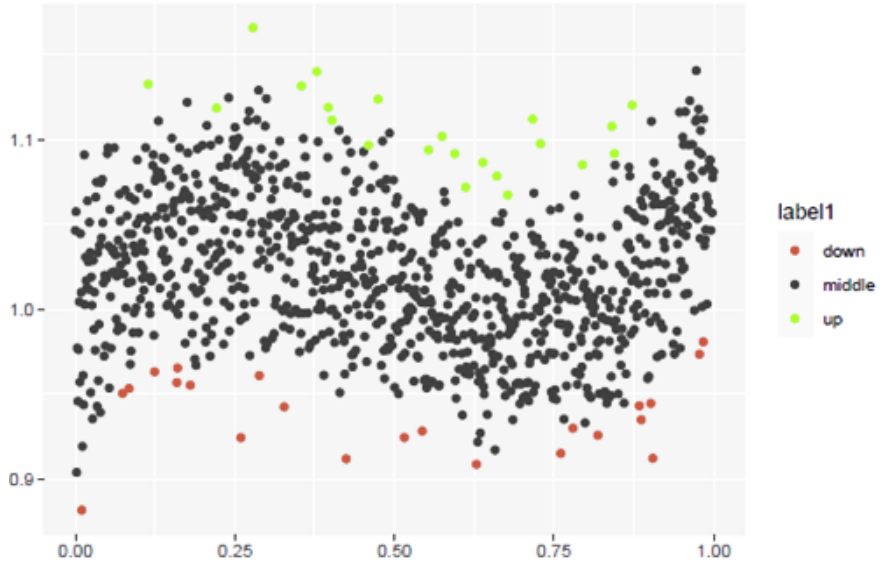
برای مثال، فرض کنید در هر مرحله داده‌ها با برچسب‌های "بالا" و "پایین" را حذف می‌کنیم. در اجرای اول، داده‌هایی که از چندک $\alpha - ۱$ توزیع خطا بیشتر بودند برچسب "بالا" دریافت کرده

و سپس حذف می‌شوند. این اتفاق برای داده‌های پایین منحنی نیز برای چندک α می‌افتد. داده‌های جدید پس از حذف مشاهدات برچسب‌دار، بین چندک‌های α و $1-\alpha$ توزیع خطا قرار دارند. این موضوع باعث می‌شود که $\widehat{F}$ جدید تغییر کند (چون توزیع خطا در داده‌ی جدید با داده‌های اول کمی متفاوت است). بنابراین در مرحله‌ی بعد، با تنظیمات یکسان، چندک‌های متناظر با α عوض می‌شوند. از این رو ممکن است در اجرای دوم نقاط جدیدی برچسب‌گذاری شوند. این اتفاق در اجراهای سوم، چهارم و الی آخر تکرار می‌گردد.

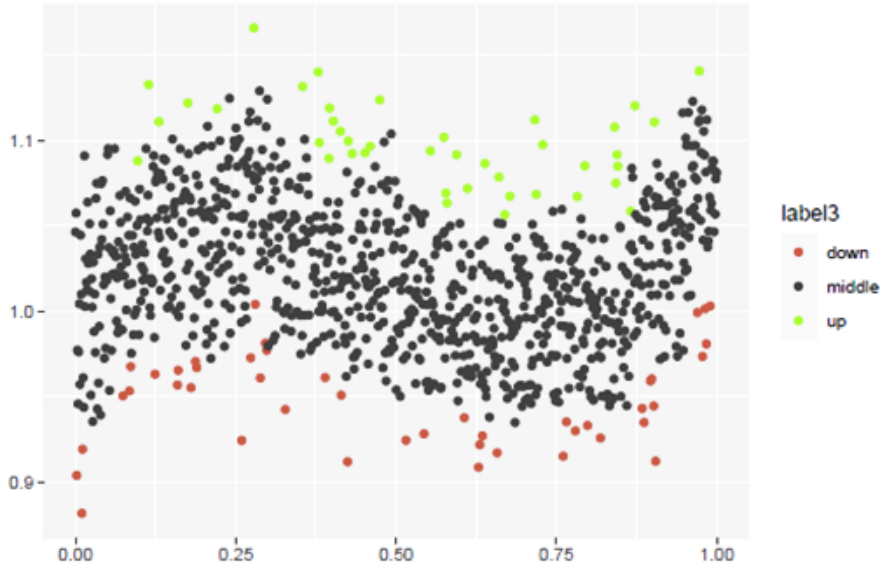
شبیه‌سازی‌های انجام شده نشان می‌دهد که پس از تکرار الگوریتم، نتایج پس از چند اجرا ثابت می‌ماند. به علاوه، تعداد نقاط شناسایی شده پس از حذف داده‌ها برای تنظیمات یکسان، در هر اجرا کمتر از قبل است.

در مثال ۴.۵ نقاط برچسب‌دار برای $\alpha = ۰/۰۵$ در ۱، ۳ و ۲۰ تکرار الگوریتم را در شکل‌های زیر آورده شده‌اند.

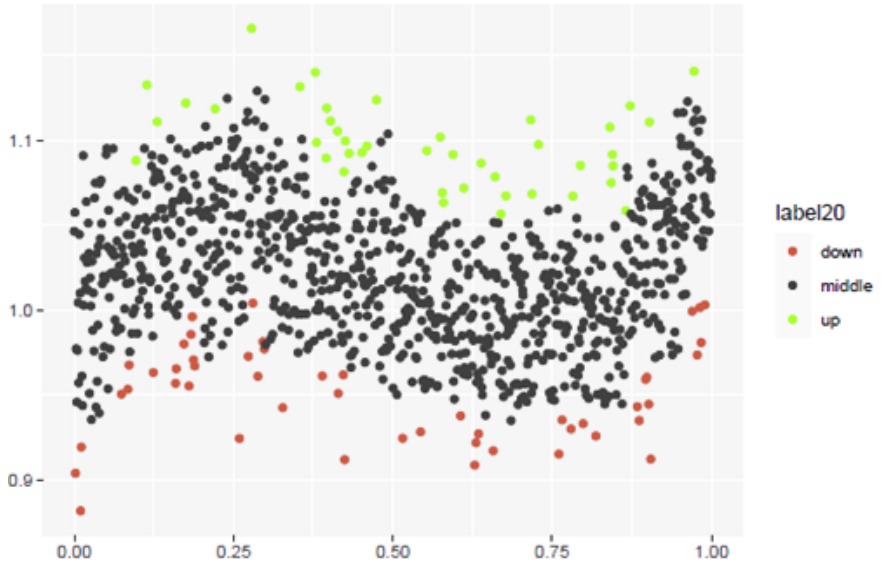
مشاهده می‌کنیم که تعداد برچسب‌های جدید از اجرای اول به سوم کمتر از تعداد برچسب‌های اجرای اول است. تفاوت بسیار کمی میان نقاط برچسب‌دار اجرای سوم و بیستم وجود دارد.



شکل ۳.۵: اجرای اول



شکل ۴.۵: اجرای سوم

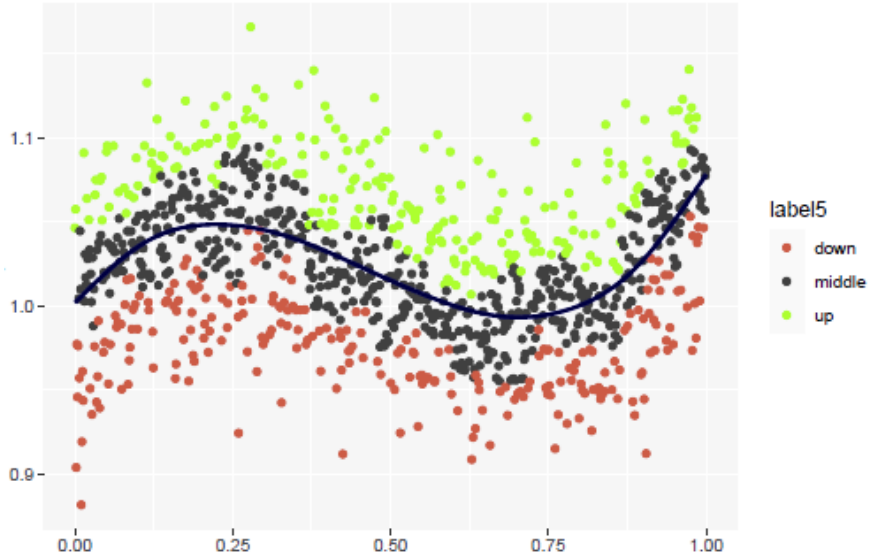


شکل ۵.۵: اجرای بیستم

آنالیزگر می‌تواند در هر اجرا مقدار d را کمتر و یا α را بیشتر در نظر بگیرد تا بتواند با از تکرار

الگوریتم، پس از شناسایی و حذف نقاط برچسب‌دار، شکل کلی تابع g را از روی نمودار پراکنش تشخیص دهد.

در شکل بعدی، الگوریتم با این تنظیمات بر داده‌های مثال ۴.۵ پیاده شده است. ملاحظه می‌شود که تشخیص شکل تابع g ساده‌تر شده است.



شکل ۶.۵: تشخیص شکل تابع با استفاده از الگوریتم برچسب‌گذاری

واژه‌نامه فارسی به انگلیسی

مقیاس ..... Scale
داده‌های پرت ..... Outliers
چندجمله‌ای ..... Polynomial
مشاهدات ..... Observations
آزمایش تصادفی ..... Random Experiment
نمودار پراکنش ..... Scatter Plot
فضای نمونه ..... Sample Space
متغیر تصادفی ..... Random Variable
پیشامد ..... Event
پیشگو ..... Predictor
متغیر کمکی ..... Covariate
متغیر مستقل ..... Independent Random Variable
بردار ..... Vector
پیشبینی ..... Prediction
مدل آماری ..... Statistical Model
متغیر پاسخ ..... Response Variable
امید ریاضی ..... Expectation
تابع چگالی ..... Density Function
تابع توزیع ..... Cumulative Function
آماره ..... Statistic
برآورد ..... Estimate
برآورد فاصله‌ای ..... Confidence Interval

نااریب ..... Unbiased
آزمون فرض ..... Hypothesis Testing
توزیع‌های تبدیل یافته ..... Transmuted Distributions
تابع مولد گشتاور ..... Moment Generating Function
پیوسته یکنواخت ..... Uniformly Continuous
فشرده ..... Compact
چولگی ..... Skewness
بیش برآورد ..... Overestimation
داده‌ی پرت ..... Outliers
نمودار پراکنش ..... Scatter Plot
نمودار تراکم ..... Density Plot
پراکندگی ..... Dispersion
افراز ..... Partition
بخش ..... Section
رگرسیون خطی ..... Linear Regression
طرح آزمایش ..... Design Of Experiment
استوار ..... Robust
خطای قابل اغماض ..... Negligible Error
استنباط ..... Inference
پارامتر ..... Parameter
ناپارامتری ..... Nonparametric

کتاب‌نامه

[1] Ross, S. (2012). *A first course in probability*, 9th ed, Pearson.

[2] Hogg, R. V., Tanis, E. and Zimmerman, D. (2013). *Probability and statistical inference*, 9th ed, Pearson.

[3] Montgomery, D. C. (2013). *Design and alnalysis of experiments*, 8th ed, New York.

[4] Montgomery, D. C., et. al., (2012). *Introduction to linear regression analysis*, 5th ed, John Wiley, New York.

[5] Hogg, R. V., et. al. (2019). *Introduction to mathematical statistics*. 8th ed, Pearson.

[6] Conover, R. J. (1999). *Practical nonparametric Statistics*. 3rd ed, John Wiley, New York.

[7] Wayne, W. D. (1990). *Applied Nonparametric Statistics*. 2nd ed, PWS-KENT Pub.

[8] Rudin, W. M. .*Principle Of Mathematical Analysis*. 3rd ed.

Abstract

The three algorithms introduced in this paper help make the process of data analysis and model developing easier. Application of said algorithms enables us to guess the distribution of error and estimate its CDF. The results allow us to study important characteristics of error distribution such as spread, symmetry and skewness. Several methods for reduction of error variance are suggested which can improve analyst's ability to detect the true relationship between predictors and response variable.

College of Science
School of Mathematics, Statistics, and Computer Science

Presenting and Evaluating the Effectiveness of an Algorithm on the Reduction of the Undesirable Effect of Error Variance in Statistical Modeling

Roxana Darvishi

Supervisor: Dr. Ali Kamalinejad

A thesis submitted in partial fulfillment of the requirements for the degree of B.Sc. in Statistics

Summer 2022