رگرسیون مؤلفهٔ اصلی تابعی در مقابل رگرسیون بردار پشتیبان برای تحلیل داده‌های طیف‌سنجی

آرتا روحی[۱]، فاطمه جهادی [۲]، مهدی روزبه[۳]

تاریخ دریافت: ۱۴۰۱/۰۲/۰۲
تاریخ پذیرش: ۱۴۰۱/۱۲/۰۲

چکیده:
مشهورترین تکنیک تحلیل داده‌های تابعی رویکرد مؤلفه‌های اصلی تابعی است که ابزاری مهم برای کاهش بعد نیز است. رگرسیون بردار پشتیبان شاخه‌ای از یادگیری ماشین و ابزار قدرتمندی برای تحلیل داده است. در این مقاله با استفاده از رگرسیون مؤلفهٔ اصلی تابعی بر اساس تاوان‌های مشتق دوم، ستیغی و لاسو و با توجه به رگرسیون بردار پشتیبان با چهار هستهٔ (خطی، چندجمله‌ای، سیگموید و شعاعی) در داده‌های طیف‌سنجی به مدل‌سازی متغیر وابسته روی متغیرهای پیش‌بین پرداخته شده است. بر اساس نتایج به‌دست‌آمده طبق معیارهای نیکویی برازش پیشنهادی، مدل رگرسیون بردار پشتیبان با هستهٔ خطی و خطای بهینه‌شده ۰٫۲ مناسب‌ترین برازش را به داده‌ها داشته است.
واژه‌های کلیدی: تحلیل داده‌های تابعی، رگرسیون بردار پشتیبان، رگرسیون تابعی، رگرسیون مؤلفهٔ اصلی، یادگیری ماشین.

۱ مقدمه

تحلیل داده‌های تابعی در داده‌هایی مورداستفاده قرار می‌گیرد که دارای ماهیت تابعی و پیوسته هستند. وقتی مشاهدات تابعی باشند، امکان برآورد ضرایب رگرسیونی با استفاده از روش‌های کلاسیک امکان‌پذیر نیست و اگر بخواهیم در قالب داده‌های چند متغیره به بررسی این داده‌ها بپردازیم، بسیاری از مزایای تحلیل از بین می‌رود و به همین خاطر ضرورت تحلیل و بررسی داده‌های تابعی با در نظر گرفتن ماهیت اصلی آن‌ها بسیار احساس می‌شود. به دلیل پرکاربرد بودن مدل تابعی، شاهد استفاده بسیار زیاد از این رویکرد هستیم و روش‌های مختلفی ارائه‌شده است که هستهٔ اصلی این روش‌ها، کاهش بعدهایی از پیش‌بینی کننده‌ها یا تنظیم منظم تابع ضریب در قالب یک تاوان ناهموار است که برای اطلاعات بیشتر می‌توان به هوروث و کوکوزکا [۱۱]، هزینگ و اوبانک [۱۲] و در سال‌های اخیر به کوکوزکا و ریمهر [۷]، فبروباند و همکاران [۸] و ریس و همکاران [۱۹] مراجعه کرد. یادگیری ماشین شاخهٔ جدیدی از تحلیل آماری است که از قدرت محاسباتی گسترده‌ای توسط رایانه‌ها برای تجزیه داده‌های بزرگ استفاده می‌کند. بر این اساس، نیاز به سیستم‌هایی است که توانایی یادگیری از طریق آموزش و تشخیص الگوها را دارند تا در دسته‌بندی کردن داده‌ها عملکرد مناسبی داشته باشند. برخی از محققان از الگوریتم‌های یادگیری ماشین برای افزایش عملکرد پیش‌بینی استفاده کرده‌اند [۱۴، ۲۳]. بر اساس مطالعات ماناهوف و همکاران [۱۵] و چودهاری و همکاران [۶]، نایاک و همکاران [۱۶]، پتال و همکاران [۱۷] و راجو و همکاران [۵] معیارهایی که برای ارزیابی مدل موردبررسی قرار می‌گیرند عبارت‌اند از ریشه میانگین توان‌های دوم خطا (RMSE) و میانگین درصد خطای مطلق و یا میانگین انحراف درصد خطای مطلق (MAPE)، که فرمول‌های محاسباتی آن‌ها به‌صورت زیر است:

$$RMSE = \sqrt{\frac{۱}{T}\sum_{i=۱}^{T}(d_i - \hat{d}_i)^۲}, \quad MAPE = \frac{۱}{T}\sum_{i=۱}^{T}\left|\frac{d_i - \hat{d}_i}{d_i}\right|$$

به‌طوری‌که در فرمول بالا d_i مقدار واقعی داده‌ها و $\hat{d}_i$ مقدار برازش شده و T نمونه‌های آزمون می‌باشد. بدیهی است که روش فوق با معیار انتخابی کمتر مناسب‌تر است. در این مقاله با تعریف مختصر روش‌های رگرسیون مؤلفهٔ اصلی، ستیغی، لاسو، تابعی و بردار پشتیبان و در نهایت با مقایسه روش‌های معرفی‌شده به بررسی داده‌های طیف‌سنجی پرداخته می‌شود.

[۱] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.
[۲] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.
[۳] هیئت‌علمی گروه آمار، دانشگاه سمنان، سمنان، ایران (نویسنده مسئول:mahdi.roozbeh@semnan.ac.ir).
[۳] کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲J۰۵; ۶۲H۲۵.

۱.۱ روش مؤلفه‌های اصلی

روش مؤلفه‌های اصلی، از روش‌های کاربردی با ایده‌ی کاهش ابعاد و حفظ بیشترین اطلاعات ممکن از متغیرهای توضیحی است. محققان در این روش با استفاده از مقادیر و بردارهای ویژه ماتریس واریانس-کوواریانس یا ماتریس همبستگی، به دنبال ترکیب خطی از متغیرهای توضیحی که بیش‌ترین واریانس را توجیه کند، می‌باشند [۱۳]. از دیدگاه هندسی، این روش تبدیل خطی متعامدی است که داده‌ها را از یک دستگاه مختصات به دستگاه مختصات جدید می‌برد به‌نحوی‌که بزرگ‌ترین واریانس داده‌ها بر روی محور مختصات اول، دومین بزرگ‌ترین واریانس بر روی محور مختصات دوم قرار می‌گیرد و به همین ترتیب ادامه می‌یابد. در نتیجه با این کار متغیرها با کاهش بعد به مؤلفه‌هایی تبدیل می‌شوند که به آن‌ها مؤلفه‌های اصلی گفته می‌شود. مؤلفه‌های اصلی ضمن ناهمبستگی به نحوی سازمان‌دهی می‌شوند که تعداد کمی از مؤلفه‌ها بتوانند درصد قابل‌توجهی از تغییرات متغیرهای توضیحی اولیه را توجیه کنند. انتخاب تعداد مناسب مؤلفه‌ها موردتوجه است و روش‌های مختلفی برای تعیین تعداد مناسب این مؤلفه‌ها پیشنهادشده که در زیر به برخی از آن‌ها اشاره می‌شود. یکی از روش‌های تعیین تعداد مؤلفه‌های اصلی انتخاب شماری از مؤلفه‌هاست که بتوانند درصد قابل‌توجهی از واریانس (تغییرات کل) را توجیه کنند. روش دیگر، برگزیدن تعداد مؤلفه‌هایی است که واریانس آن‌ها بزرگ‌تر یا مساوی متوسط واریانس کل است. از روش‌های شهودی نیز برای انجام این کار، نمودار بازو است. در این نمودار مقادیر ویژه هر مؤلفه (λ_i) در برابر i رسم می‌شود. با عمود کردن ناحیه‌ای از نمودار که شیب آن به‌طور ناگهانی کم می‌شود، تعداد مؤلفه‌ها تعیین می‌گردد. توجه به این نکته ضروری است که ممکن است هرکدام از روش‌های مطرح‌شده پاسخ‌های متفاوتی ارائه دهند که محقق می‌تواند بنا بر کاربرد از روش موردنظر استفاده کند. بعد از انتخاب تعداد متغیرها نوبت به برازش مدل رگرسیون مؤلفه‌های اصلی می‌شود. در این روش، مدل رگرسیونی دیگر با مشکل هم‌خطی روبرو نیست و می‌توان از روش کمترین توان‌های دوم معمولی برای برآورد ضرایب آن استفاده کرد. از کاربردهای روش مؤلفه‌های اصلی می‌توان به تبدیل متغیرهای همبسته به متغیرهای ناهمبسته، یافتن ترکیبات خطی با واریانس نسبی بزرگ یا کوچک و کاهش در حجم داده‌ها نام برد.

۲.۱ رگرسیون لاسو

ایده اولیه رگرسیون لاسو را لئو برایمن با نام گروتی بیان کرد [۲۱]. او نخستین بار مسئله بهینه‌سازی نامنفی خود را گروتی نام نهاد و به‌صورت زیر مطرح کرد:

$$\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p}c_j\hat{\beta}_jX_{ij}\right)^2 \qquad c_j\geq 0,\quad \sum_j c_j\leq s \quad (1)$$

که در آن برآوردگر اولیه $\hat{\beta}_j$ با روش کمترین توان‌های دوم انتخاب می‌شود. پارامتر s پارامتر تاوان است که با کاهش آن گروتی محدود می‌شود. این روش در بین محققان به نام تاوان مرگ نیز معروف است. در این روش برخی از متغیرها حذف‌شده و بقیه آن‌ها منقبض می‌شود. هنگامی‌که ضرایب کوچک غیرصفر دارد، باید در استفاده از این روش احتیاط لازم انجام شود. با توجه به این مسئله، لاسو تابع هدفی ارائه می‌دهد که در آن از برآوردگرهای کمترین توان‌های دوم استفاده نمی‌کند و به‌صورت زیر نوشته می‌شود:

$$\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p}X_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\},$$
$$i=1,\ldots,n,\quad j=1,\ldots,p \qquad (2)$$

که در آن Y_i متغیر پاسخ، X_{ij} متغیر توضیحی و β_j ضرایب رگرسیون می‌باشد. از مزایای اثبات شده این روش این است که برآوردگرهای پایدار، اریب و پیوسته تولید می‌کند [۲۱]. همچنین این روش برخی از ضرایب را منقبض و بقیه را صفر می‌کند. از معایب روش لاسو می‌توان به عملکرد ضعیف آن در حضور متغیرهای توضیحی همبسته اشاره کرد زیرا که بین متغیرهای همبسته به‌ویژه متغیرهایی که به‌صورت گروهی همبسته‌اند یکی از آن‌ها را انتخاب می‌کند و متغیر انتخاب‌شده لزومی ندارد که بهترین متغیر باشد. لاسو برای یک مدل رگرسیونی خطی با p متغیر توضیحی و n مشاهده، حداکثر n متغیر را انتخاب می‌کند، بنابراین هنگامی‌که متغیرهای بیشتری (بیشتر از n) در مدل معنی‌دار باشند، در این روش شانسی برای انتخاب ندارند.

۳.۱ رگرسیون ستیغی

در مسائل رگرسیونی گاهی اوقات با هم‌خطی بین متغیرهای پیش‌بین روبرو می‌شویم که معمولاً در مدل‌هایی با تعداد پارامتر بالا اتفاق می‌افتد. آندری نیکولایویچ تیخونوف که تحقیقات مهم او در توپولوژی، تجزیه‌وتحلیل عملکردی، فیزیک، ریاضیات شناخته شده است، منظم

[3] Principal Component Analysis

[3] Lasso

سازی تیخونوف را به‌عنوان راه‌حلی برای رویارویی با این‌گونه مسائل معرفی نمود. در داده‌های با بعد بالا چون ماتریس $\boldsymbol{X}^\top \boldsymbol{X}$ وارون‌پذیر نیست، روش کمترین توان‌های دوم رگرسیون خطی نمی‌تواند مفید باشد، زیرا برآورد کمترین توان‌های دوم معمولی در مدل رگرسیون خطی زمانی که تعداد متغیرها بیشتر از داده‌ها باشد، تعریف نمی‌شود. در این شرایط با استفاده از روش ستیغی می‌توان وارون ماتریس را به دست آورد، بدین‌صورت که با افزودن مقدار مثبت k به ماتریس $\boldsymbol{X}^\top \boldsymbol{X}$ این ماتریس وارون‌پذیر خواهد شد [۱، ۳، ۱۰]. در مدل رگرسیون برآوردگر ستیغی دیگر نااریب نیست اما می‌تواند دارای واریانس کمتری می‌باشد و مقدار برآوردگر پایا است؛ یعنی تغییرات جزئی در داده‌ها، بر برآوردگر تأثیری نخواهد داشت. برآورد $\boldsymbol{\beta}$ در این روش به‌صورت زیر است:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \qquad k \geq 0 \qquad (۳)$$

که در معادله اخیر به k پارامتر ستیغی گفته می‌شود و یافتن مقدار بهینه این پارامتر بسیار مهم است. برای انتخاب k باید به‌گونه‌ای عمل شود که کاهش در واریانس برآوردگر اریب بیش از افزایش مربع اریبی باشد، در نتیجه MSE آن کمتر از واریانس برآوردگر نااریب خواهد بود.

۲ استفاده از روش هموارسازی در تحلیل داده‌های تابعی

در ابتدا به برآورد یک منحنی یا خط ساده برای برازش روی داده‌ها نیاز است و چون داده‌ها به‌صورت گسسته می‌باشند، باید داده‌های گسسته را به داده‌های پیوسته تبدیل کرد. از روش‌های درون‌یابی و هموارسازی برای ایجاد یک منحنی با توجه به نوع داده می‌توان استفاده کرد. برای تحلیل داده‌های تابعی در ابتدا باید داده‌های گسسته را به پیوسته تبدیل کرد. برای این منظور از روش‌های هموارسازی و پایه‌هایی نظیر فوریه برای داده‌های دارای دوره تناوب، اسپلاین برای سایر داده‌ها می‌توان استفاده کرد.

به‌طورکلی فرم قرارگیری داده‌ها، اگر به‌صورت ترکیب خطی باشند به آن توابع پایه (تابع اساسی) گفته می‌شود و به‌صورت زیر قابل‌نمایش است:

$$Y_i = \sum_{j=1}^{k} c_j \phi_j(t_i) + \epsilon_i = f(t_i) + \epsilon_i \qquad (۴)$$

که در آن تابع f ترکیب خطی از ضرایب c_j، تابع پایه ϕ_j، Y_i متغیرهای پاسخ و t_i متغیرهای مستقل است. همان‌طور که هر بردار در فضای برداری را می‌توان به‌عنوان یک ترکیب خطی از بردارهای پایه نشان داد، هر تابع پیوسته در فضای تابعی را می‌توان به‌صورت ترکیبی خطی از توابع پایه نوشت. اکنون برای توابع پایه ϕ_j حالت‌های زیر را در نظر می‌گیریم:

۱.۲ تقریب توابع پایه با استفاده از پایه‌های فوریه

تابع پایه فوریه برای داده‌هایی کاربرد دارد که نوسانی بوده و دارای دوره تناوب باشند، همانند داده‌های آب‌وهوا که در زمستان معمولاً سرد و در تابستان گرم هستند. پایه‌های فوریه به‌صورت زیر تعریف می‌شود:

$$\{1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \ldots, \sin(m\omega t), \cos(m\omega t)\}$$

که در آن ω را دوره نوسان نامیده و برابر با $\omega = \frac{2\pi}{p}$ است به‌طوری‌که p دوره تکرارشونده است. برای مثال دوره تکرارشونده برای داده‌های آب‌وهوا برابر با ۳۶۵ روز است.

۲.۲ تقریب توابع پایه با استفاده از پایه‌های اسپلاین

پایه‌های اسپلاین هر دو نقطهٔ مجاور را با یک تابع برآورد می‌کند. اساساً اسپلاین‌ها همانند تابع‌های چندجمله‌ای هستند که ابتدا داده‌های گسسته را به چند قسمت مساوی تقسیم می‌کنند و سپس به دنبال بهترین منحنی برای برازش به هر قسمت می‌باشد که اگر درجه آن صفر باشد با یک خط افقی به برآورد می‌پردازد و اگر درجه آن یک باشد به‌صورت خطی و درجه‌های بالاتر را به‌صورت منحنی برآورد می‌کند. در ضمن قسمت‌هایی از منحنی که در محل پیوستن به هم هستند را می‌توان هموار کرد، این نقاط گره نامیده می‌شود. برای انتخاب گره‌ها، اگر تعداد آن‌ها خیلی زیاد باشد اریبی بسیار کم و واریانس بسیار زیاد می‌شود و باعث ناهمواری نمودار می‌شود. نکته‌ای که باید در نظر گرفت این است که باید در هر گره حداقل یک داده وجود داشته باشد. برخی دیگر از توابع پایه می‌توان به توابع پایه ثابت، توانی، نمایی و... اشاره کرد.

[3] Ridge
[3] Fourier
[3] Spline

۳.۲ مدل رگرسیون تابعی

تابع رگرسیونی به‌صورت

$$Y_i = f(t_i) + \epsilon_i, \quad i = ۱, \ldots, n$$

را در نظر می‌گیریم که در آن خطاها مستقل و دارای توزیع نرمال با میانگین صفر و واریانس $\sigma^۲$ است. برای برآورد $f(t_i)$ بر اساس توابع پایه می‌توان نوشت:

$$\hat{f}(t_i) = \sum_{j=۱}^{p} c_j \phi_j(t_i) \qquad (۵)$$

به‌طوری‌که در معادلهٔ ۵، $\phi_j(t)$ تابع پایه و c_jها ضرایب می‌باشند. برای برآورد ضرایب تابعی باید عبارت زیر را حداقل کرد:

$$\boldsymbol{H(c)} = \sum_{i=۱}^{n} (Y_i - f(t_i))^۲ = \sum_{i=۱}^{n} (Y_i - \sum_{j=۱}^{p} c_j \phi_j(t_i))^۲ \qquad (۶)$$

همچنین می‌توان معادله‌ی فوق را به‌صورت ماتریس زیر نوشت:

$$\boldsymbol{H(c)} = (\boldsymbol{Y} - \boldsymbol{\Phi c})^\top (\boldsymbol{Y} - \boldsymbol{\Phi c}) \qquad (۷)$$

در معادله بالا، $\boldsymbol{Y}$ ماتریس متغیر پاسخ، $\boldsymbol{c}$ ماتریس ضرایب و $\boldsymbol{\Phi}$ ماتریس توابع پایه می‌باشد که با مشتق‌گیری و برابر صفر قرار دادن معادله ۷ داریم:

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-۱} \boldsymbol{\Phi}^\top \boldsymbol{Y} \qquad (۸)$$

حال با توجه به برآورد بالا مقدار $\hat{f}(\boldsymbol{t})$ برابر $\hat{f}(\boldsymbol{t}) = \hat{\boldsymbol{c}}^\top \boldsymbol{\Phi}$ است و همچنین $\hat{\boldsymbol{Y}}$ به‌صورت

$$\hat{\boldsymbol{Y}} = \underbrace{\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-۱} \boldsymbol{\Phi}}_{\boldsymbol{S}} \boldsymbol{Y} = \boldsymbol{SY}$$

به دست می‌آید که به $\boldsymbol{S}$ ماتریس هموارساز می‌گویند. در عبارت بالا مقدار $\hat{\boldsymbol{Y}}$ ضریبی از مقدار $\boldsymbol{Y}$ است، یعنی با هموار کردن داده‌ها توسط $\boldsymbol{S}$ مقدار $\hat{\boldsymbol{Y}}$ به دست می‌آید. انتخاب تعداد توابع پایه بسیار اهمیت دارد، اگر تعداد این توابع کم باشد نشان‌دهنده اریبی زیاد و واریانس کم است و زیاد بودن تعداد توابع نشان‌دهنده اریبی کم و واریانس زیاد است که منجر به بیش برازش می‌شود.

۴.۲ برآورد به روش تاوان کمینه کردن طول مشتق دوم

با داشتن متغیر تابعی $X(\boldsymbol{t})$ طول قوس منحنی را با توجه به فرمول زیر می‌توان به دست آورد:

$$J_۲(X(\boldsymbol{t})) = \int \left(D^۲ X(\boldsymbol{t})\right)^۲ \mathrm{d}t$$

اگر طول قوس منحنی کم باشد، بدین معنی است که به نوسان‌های داده‌ها کمتر توجه شده و اگر طول قوس منحنی زیاد باشد، پس باید تاوان افزایش یابد. به بیانی دیگر باید مشتق دوم ضرایب تابعی را هموار کرد تا منحنی این ضرایب هموار شود. با در نظر گرفتن مربع خطای تاوانیده به‌صورت زیر داریم:

$$PENSSE_\lambda(X(\boldsymbol{t})) = (\boldsymbol{Y} - X(\boldsymbol{t}))(\boldsymbol{Y} - X(\boldsymbol{t}))^\top + \lambda J_۲(X(\boldsymbol{t}))$$

به λ پارامتر هموارساز گفته می‌شود که کنترل بین مقادیر برازش شده و هموارای منحنی بر داده‌ها می‌باشد. اگر مقدار λ افزایش یابد، هموارکننده تاوانیده و $X(\boldsymbol{t})$ تبدیل به خط می‌شود و اگر λ کاهش یابد، تاوان هم کاهش‌یافته و اجازه می‌دهد $X(\boldsymbol{t})$ روی داده‌ها برازش شود. باید مربع خطای تاوانیده را حداقل کرد و این امر در پایه‌های بی اسپلاین زمانی اتفاق می‌افتد که مقدار درجه آن برابر ۴ باشد، زیرا دارای مشتق دوم بوده و می‌توان مشتق دوم را هموار کرد و منحنی‌های هموار مناسب‌تری ایجاد کرد. اکنون با تعریف $X(\boldsymbol{t}) = \boldsymbol{\Phi}^\top \boldsymbol{c}$ و با توجه به اینکه $X(\boldsymbol{t})$ یک ترکیب خطی است داریم:

$$\int [D^m X(\boldsymbol{t})]^۲ \mathrm{d}t = \boldsymbol{c}^\top \underbrace{\int D^m \boldsymbol{\Phi} D^m \boldsymbol{\Phi}^\top \mathrm{d}t}_{\boldsymbol{R}} \boldsymbol{c} = \boldsymbol{c}^\top \boldsymbol{R} \boldsymbol{c}$$

به‌طوری‌که در معادله‌ی فوق $\boldsymbol{R}$ را به‌عنوان ماتریس تاوان می‌شناسیم. با حداقل کردن مربع خطاها و اندازه طول مشتق دوم، مقدار برآورد $\boldsymbol{c}$ به‌صورت $\hat{\boldsymbol{c}} = \left[\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{R}\right]^{-۱} \boldsymbol{\Phi}^\top \boldsymbol{Y}$ خواهد بود و مقدار برآورد $\boldsymbol{Y}$ برابر است با:

$$\hat{\boldsymbol{Y}} = \underbrace{\boldsymbol{\Phi}\left[\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{R}\right]^{-۱} \boldsymbol{\Phi}^\top}_{\boldsymbol{S}} \boldsymbol{Y} = \boldsymbol{SY}$$

برای انتخاب پارامتر هموارساز می‌توان از روش‌های متداول زیر استفاده کرد:

۱) اعتبار سنجی متقابل معمولی[۴]:

$$OCV(\lambda) = \frac{۱}{n} \sum_{i=۱}^{n} \frac{(Y_i - X(t_i))^۲}{(۱ - S_{ii})^۲} \qquad (۹)$$

[4] Ordinary cross validation

[5] Generalized cross validation

۲) اعتبار سنجی متقابل تعمیم‌یافته[5]:

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - X(t_i))^2}{[n^{-1}trace(1-S)]^2} = \left(\frac{n}{n-df(\lambda)}\right)\left(\frac{SSE}{n-df(\lambda)}\right) \quad (10)$$

باید توجه داشت که GCV هموارتر از OCV است [۴، ۱۸، ۲۰].

۳ ماشین‌های بردار پشتیبان

در سال‌های اخیر یادگیری ماشین در تحلیل داده رشد چشمگیری داشته است و پژوهشگران زیادی به این شیوه متوسل شده‌اند. در میان روش‌ها و الگوریتم‌های گوناگونی که در حوزه یادگیری ماشین است، ماشین‌های بردار پشتیبان و یکی از مهم‌ترین و پرکاربردترین آن‌ها است که ابزاری قدرتمند برای طبقه‌بندی داده‌ها می‌باشد، روش مذکور توسط وپنیک در سال ۱۹۹۵ به وجود آمده است [۲۲]. مدل رگرسیون بردار پشتیبان این انعطاف‌پذیری را به ما می‌دهد تا خطای حاصل از تفاوت برآورد با مقدار واقعی را تغییر داد و بتوان مدلی با برآورد مناسب ایجاد کرد. کنترل هواپیما بدون خلبان، آنالیز کیفیت کامپیوتر، طراحی اعضای مصنوعی، سیستم‌های مسیریابی و ... از کاربردهای این مدل است [۲].

۱.۳ مسئله بهینه‌سازی در یادگیری ماشین‌های بردار پشتیبان

بدین منظور نقطه‌ای مانند x_n را به‌گونه‌ای در نظر می‌گیریم که نزدیک‌ترین نقطه به ابرصفحه، صفحه‌ای که دو نوع داده متفاوت را از هم جدا می‌کند، است. با در نظر گرفتن ابرصفحه $|\boldsymbol{W}^\top x_n + w_0| = 0$ برای بردار $\boldsymbol{W} = [w_1, \ldots, w_N]$، معادلات خطوط حاشیه‌ای به شکل $|\boldsymbol{W}^\top x_n + w_0| = 1$ می‌باشد. هدف این است که فاصله خطوط حاشیه از هم تا حدی زیاد شود که فاصله‌ی خط حاشیه تا نقطه‌ی x_n به کمترین مقدار برسد. لذا شرط را با توجه به مقادیر گسسته‌ای که متغیر پاسخ می‌گیرد، می‌توان به‌صورت متفاوت نوشت:

$$\min_{\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{W}\|^2, \qquad s.t. \quad y_n(\boldsymbol{W}^\top x_n + w_0) \geq 1$$

مسئله فوق یک مسئله برنامه‌نویسی درجه دوم می‌باشد که حل آن دشوار است و درنتیجه با استفاده از تعریف تابع لاگرانژ چند متغیره با در نظر گرفتن $\alpha_i \geq 0$ و بردار α که شامل α_i می‌باشد. مقادیر مختلف λ_i حالت‌های مختلفی از تابع لاگرانژ به دست می‌آید، که با توجه به مسئله دوگان و تعریف $p^* = \min_x \max_{\alpha_i \geq 0, \lambda} \ell(x, \alpha, \lambda)$ است و می‌توان نوشت:

$$\max_x \min_y h(x,y) \leq \min_y \max_x h(x,y),$$

$$p^* = \min_x \max_{\alpha_i \geq 0, \lambda_i} \ell(x, \alpha, \lambda), d^* = \max_{\alpha_i \geq 0, \lambda_i} \min_x \ell(x, \alpha, \lambda),$$

$$d^* \leq p^* \quad (11)$$

حال اگر به‌جای p^*، $\min_W \frac{1}{2}\|\boldsymbol{W}\|^2$ قرار گیرد، خواهیم داشت:

$$\min_{\boldsymbol{W}, w_0} \max_{\alpha_n \geq 0} \{\frac{1}{2}\|\boldsymbol{W}\|^2 + \sum_{n=1}^{n=N} \alpha_n(1 - y_n(\boldsymbol{W}^\top x_n + w_0))\} \quad (12)$$

$$\max_{\alpha_n \geq 0} \min_{\boldsymbol{W}, w_0} \{\frac{1}{2}\|\boldsymbol{W}\|^2 + \sum_{n=1}^{n=N} \alpha_n(1 - y_n(W^\top x_n + w_0))\} \quad (13)$$

اکنون با توجه به معادلات بالا و مسئله دوگان معادله (۱۲) بزرگ‌تر یا مساوی معادله (۱۳) است، برای کمینه کردن قسمت داخلی معادله (۱۳) باید از آن مشتق گرفت و برابر صفر قرار داد. با مشتق گرفتن نسبت به $\boldsymbol{W}$ و w_0 می‌توان نوشت:

$$\nabla_{\boldsymbol{W}} \ell(\boldsymbol{W}, w_0, \alpha) = \boldsymbol{W} - \sum \alpha_n y_n x_n = 0 \Rightarrow \boldsymbol{W} = \sum \alpha_n y_n x_n$$

$$\frac{\partial \ell(\boldsymbol{W}, w_0, \alpha)}{\partial w_0} = 0 \Rightarrow -\sum \alpha_n y_n = 0$$

و خواهیم داشت:

$$\min_{\boldsymbol{W}, w_0} \ell(\boldsymbol{W}, w_0, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n x_m,$$
$$\sum_{n=1}^{N} \alpha_n y_n = 0, \alpha_n \geq 0 \quad (14)$$

اکنون باید تابع بالا را بیشینه کنیم و چون α درجه دوم می‌باشد دوباره مسئله برنامه‌نویسی درجه دوم مطرح است، با توجه به شروط $y^\top \alpha = 0, \alpha > 0$ با جایگذاری مقادیر در معادله برنامه‌نویسی درجه دوم و یافتن α هایی که تابع (۱۴) را بیشینه می‌کنند و با جایگذاری در $\boldsymbol{W} = \sum \alpha_n y_n x_n$، مقدار $\boldsymbol{W}$ به دست می‌آید. برای به دست آوردن مقدار w_0 با استفاده از شرایط کاروش-کان-تاکر برای تابع‌های محدب، می‌توان به شکل زیر عمل نمود:

$$\nabla \ell(x, \alpha)|_{x^*, \alpha^*} = 0, s.t. \alpha_i^* \geq 0, g_i(x^*) \leq 0, \alpha_i^* g_i(x^*) = 0,$$

$$i = 1, \ldots, m$$

با توجه به مشابه بودن شرایط مسئله بهینه‌سازی ماشین‌های بردار پشتیبان با شرایط کاروش-کان-تاکر و با در نظر گرفتن شرایط جدید

[5] Support vector machine

[5] Support vector regression

می‌بایست از تابع لاگرانژ مشتق گرفت:

$$\nabla_{\boldsymbol{W}}\ell(\boldsymbol{W},w_o,\alpha)|_{\boldsymbol{W}^*,w_\circ^*,\alpha^*}=\circ \Rightarrow \frac{\partial \ell(\boldsymbol{W},w_\circ,\alpha)}{\partial \boldsymbol{W}_\circ}|_{\boldsymbol{W}^*,w_\circ^*,\alpha^*}=\circ, \alpha_n^*>\circ$$

که با توجه به شروط

$$y_n(\boldsymbol{W}^{*T}y+w_\circ^*)\geq 1, n=1\ldots,N, \alpha_i^*\underbrace{(1-y_n(\boldsymbol{W}^{*\top}y+w_\circ^*))}_{g_i(x^*)}=\circ$$

برابری $\alpha_i^*\underbrace{(1-y_n(\boldsymbol{W}^{*\top}y+w_\circ^*))}_{g_i(x^*)}=\circ$ در دو زمان اتفاق می‌افتد، زمانی که α_i^* برابر صفر باشد که در این صورت با توجه به شرایط $g_i(x^*)\leq\circ$، داده‌ها در دو طرف حاشیه قرار می‌گیرد، و لذا اگر $g_i(x^*)=\circ$ و $\alpha_i>\circ$، آنگاه داده‌ها روی حاشیه قرار گرفته و لذا می‌توان $\boldsymbol{W}$ را به صورتی نوشت که فقط α_iهایی را شامل شود که مثبت هستند، زیرا در بقیه قسمت‌ها صفر می‌شود، یعنی دسته‌بندی به‌درستی انجام شده است:

$$\boldsymbol{W}=\sum_{\alpha_n>\circ}\alpha_n y_n x_n$$

به نقاطی که مقدار α_i آن‌ها بزرگ‌تر از صفر است، ماشین‌های بردار پشتیبان (SVM) می‌گوییم. اکنون داده‌هایی که α_i آن‌ها بزرگ‌تر از صفر هستند را جدا کرده و مقدار $W=\sum\alpha_n y_n x_n$ محاسبه می‌شود. مقدار $w_\circ=y_n-\boldsymbol{W}^\top x_n$ است و همچنین برای انتخاب مرز طبقه‌بندی‌ها داریم:

$$\hat{y}=sign(w_\circ+\boldsymbol{W}^\top x)=sign(y_i-\sum_{\alpha_n\geq\circ}\alpha_n y_n x_n^\top x+\sum_{\alpha_n\geq\circ}\alpha_n y_n x_n^\top x_n)$$

تمام مطالبی که گفته شد، مربوط به زمانی است که داده‌ها در خارج حاشیه و یا روی آن قرار بگیرند. با انعطاف‌پذیری در مدل می‌توان تعدادی از داده‌ها را درون حاشیه قرار داد تا دسته‌بندی مناسب‌تر باشد. هنگامی‌که داده‌ها درون حاشیه قرار بگیرند، می‌توان نوشت:

$$y_n(\boldsymbol{W}^\top x_n+w_\circ)\geq 1-\xi_n,\qquad \xi_n>\circ \qquad (۱۵)$$

به‌طوری‌که در آن ξ_n فاصله هر داده تا خط حاشیه است، پس اگر داده‌ها درست دسته‌بندی شوند آنگاه $\xi_n=\circ$ است. اکنون یک حالت دیگر ممکن است پیش آید که داده‌ها داخل حاشیه قرار گیرند. در این حالت علاوه بر مراحل گفته‌شده، می‌توان مجموع ξ_nها را هم به حداقل رساند. پس برای انجام بهینه‌سازی داریم:

$$\min \tfrac{1}{2}\|\boldsymbol{W}\|+c\sum_{n=1}^{n=N}\xi_n,\quad s.t.\quad y_n(\boldsymbol{W}^T x_n+w_\circ)\geq 1-\xi_n,$$

$$n=1,\ldots,N,\quad \xi_n\geq\circ$$

با توجه به تابع لاگرانژ داریم:

$$\ell(\boldsymbol{W},w_\circ,\xi,\alpha,\beta)=\tfrac{1}{2}\|\boldsymbol{W}\|^2+C\sum_{n=1}^{N}\xi_n+\sum_{n=1}^{N}\alpha_n(1-\xi_n-y_n(\boldsymbol{W}^T x_n+w_\circ))-\sum_{n=1}^{N}\beta_n\xi_n \qquad (۱۶)$$

با توجه به آنچه گفته شد باید مقادیر $\boldsymbol{W},w_\circ,\xi$ را کمینه و ضرایب لاگرانژ را بیشینه کرد.

۴ رگرسیون بردار پشتیبان

ماشین‌های بردار پشتیبان در طبقه‌بندی بسیار قوی هستند اما در رگرسیون شناخته شده نیستند. رگرسیون بردار پشتیبان یک حالت از ماشین‌های بردار پشتیبان است که به جای گرفتن مقادیر گسسته ۱− و ۱ در متغیرهای پاسخ، مقادیر پیوسته می‌گیرد. در ماشین‌های بردار پشتیبان هر چه تعداد داده کمتری درون حاشیه قرار گیرد، خط جداکننده مناسب‌تر است، اما در رگرسیون بردار پشتیبان با در نظر گرفتن حاشیه‌ها، هرچه تعداد داده بیشتری درون حاشیه قرار بگیرد، مدل مناسب‌تر می‌باشد. در این بخش هدف ساخت مدل روی داده‌های $\{x_k,y_k\}_{k=1}^N$ با استفاده از رگرسیون بردار پشتیبان است که متغیر پاسخ آن پیوسته است.

رگرسیون‌های ستیغی و لاسو و شبکه‌ی الاستیک با اضافه کردن یک پارامتر تاوان اضافی باهدف به حداقل رساندن پیچیدگی و یا کاهش تعداد ویژگی‌های مورداستفاده در مدل نهایی، مدل‌سازی را انجام می‌دهند، اما در روش رگرسیون بردار پشتیبان خطا قرار نیست کمینه شود بلکه این انعطاف‌پذیری وجود دارد که بتوان خطا را تغییر داد تا مدل نهایی کاراتر شود. انتخاب خطای بهینه توسط اعتبارسنجی متقابل به دست می‌آید. مدل رگرسیون بردار پشتیبان به‌صورت زیر تعریف می‌شود:

$$f(x,\boldsymbol{W})=\boldsymbol{W}^\top x+b \qquad (۱۷)$$

برای دستیابی به مدل نهایی، یک خطای آستانه ϵ تعریف می‌شود تا در معادله‌ی زیر به کمترین مقدار برسد:

$$|y-f(x,\boldsymbol{W})|_\epsilon=\begin{cases}\circ, & |y-f(x,\boldsymbol{W})|\leq\epsilon\\ |y-f(x,\boldsymbol{W})|-\epsilon, & o.w\end{cases} \qquad (۱۸)$$

اکنون با تعریف R و با توجه به $\|\boldsymbol{W}\|$ داریم:

$$R = \frac{۱}{۲}\|\boldsymbol{W}\|^۲ + c\left(\sum_{i=۱}^{N}|y_i - f(x_i, \boldsymbol{W})|_\epsilon\right)$$

حال با تعریف متغیر ξ و ξ^* معادله‌ی بالا تبدیل به معادله (۱۹) با شروط زیر می‌شود:

$$(\boldsymbol{W}^\top x_i + b) - y_i \leq \epsilon + \xi_i, y_i - (\boldsymbol{W}^\top x_i + b) \leq \epsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \geq ۰,$$

بنابراین

$$R = \frac{۱}{۲}\|\boldsymbol{W}\|^۲ + c\sum_{i=۱}^{N}(\xi_i + \xi_i^*) \qquad (۱۹)$$

۱.۴ عدم وجود مرز خطی در مدل رگرسیون بردار پشتیبان و حل آن

اگر مرز خطی بین داده‌ها وجود نداشته باشد، باید داده‌ها را به فضایی جدید برده و در آن فضا برای داده‌ها مرز خطی پیدا کرد. در همه‌ی مسائل فوق باید x را به $\Phi(x)$ تبدیل کرد، اما چون داده‌ها همه وارد فضای جدید می‌شوند، لذا محاسبه‌ی ضرب داخلی $\Phi(x)\Phi(x)^\top$ بسیار طولانی است، بنابراین راهی معرفی می‌شود تا داده‌ها را بدون اینکه به فضای جدید تغییر داد، ضرب داخلی را بتوان حساب کرد. یکی از این راه‌ها استفاده از ترفند هسته است که در ادامه به معرفی آن خواهیم پرداخت.

۱.۱.۴ هسته‌های معروف در یادگیری ماشین

چهار هسته معروف ماشین‌های یادگیری پشتیبان عبارت‌اند از:

- هسته خطی، ساده‌ترین تابع هسته است که حاصل‌ضرب داخلی $< x, y >$ به‌علاوه یک مقدار ثابت اختیاری c می‌باشد.

$$k(x, y) = x^\top y + c$$

اگر داده‌ها را بتوان با یک خط از هم جدا کرد، استفاده از این روش سودمند می‌باشد.

- هسته چندجمله‌ای، این هسته زمانی که کلیه داده‌های آموزش نرمال شده‌اند، مناسب‌تر عمل می‌کند. شکل تابعی آن به‌صورت زیر می‌باشد:

$$k(x, y) = (\alpha x^\top y + c)^d$$

که در آن پارامترهای c و α و درجه‌ی چندجمله‌ای d قابل انتخاب و تنظیم است.

- هسته گوسی، نمونه‌ای از تابع شعاعی است که شکل تابع هسته آن به‌صورت زیر است:

$$k(x, y) = exp\left(\frac{\|x - y\|^۲}{۲\sigma^۲}\right)$$

که در آن پارامتر قابل تنظیم σ نقش عمده‌ای در تابع هسته دارد، به‌طوری‌که اگر بیش‌ازحد بزرگ تعیین شود، رفتاری تقریباً خطی نشان داده و پیش‌بینی بر اساس آن شروع به از دست دادن رفتار غیرخطی خود خواهد کرد. اگر بیش‌ازحد کوچک تعیین گردد، تابع فاقد تنظیم و مرز تصمیم‌گیری در مورد اختلال در داده‌های آموزش بسیار حساس خواهد بود.

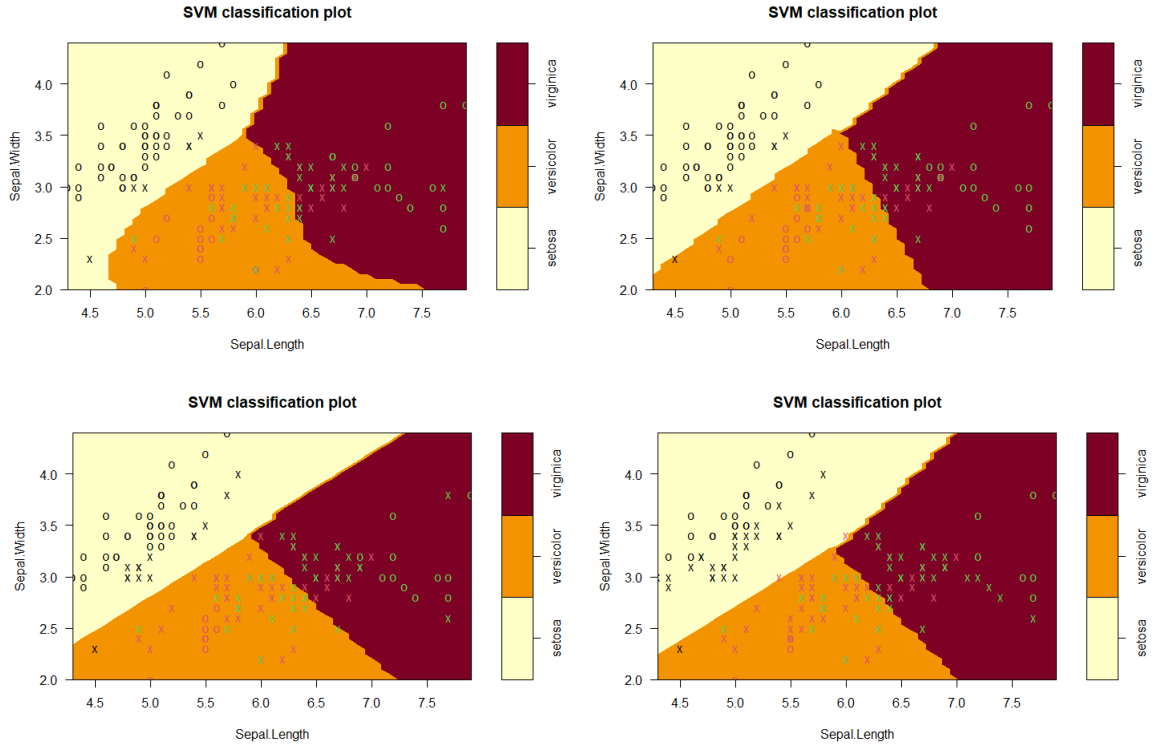
- هسته سیگموید، به‌عنوان هسته چندلایه پرسپترون شناخته می‌شود. هسته سیگموید از قسمت شبکه‌ی عصبی می‌آید و جایی که تابع سیگموئید دوقطبی است، اغلب به‌عنوان تابعی از فعال‌سازی نورون‌های مصنوعی استفاده می‌شود. هسته آن به‌صورت زیر می‌باشد:

$$k(x, y) = \tanh(\alpha x^\top y + c)$$

دو پارامتر قابل تنظیم در هسته سیگموید وجود دارد که عبارت‌اند از شیب α و عرض از مبدأ c. بنا بر آنچه گفته شد، تکنیک رگرسیون بردار پشتیبان برای ساخت مدل به توابع هسته متکی است. از طرفی انتخاب اینکه کدام هسته با توجه به کدام داده مناسب‌تر بوده و عملکرد بهتری دارد، دشوار است و نیاز به حل مسائل بهینه‌سازی دارد.

مثال ۱.۴. *داده‌های گل زنبق[6] شامل متغیرهای طول و عرض کاسبرگ و گلبرگ گل زنبق است که شامل سه کلاس ۵۰ نمونه‌ای است که در آن هر کلاس به یک نوع گیاه زنبق اشاره دارد. داده‌های گل زنبق توسط رانلد فیشر در سال ۱۹۳۶ معرفی شد. نتایج تحلیل و مدل‌سازی این داده‌ها توسط تکنیک رگرسیون بردار پشتیبان با هسته‌های متفاوت را می‌توان در شکل ۱ مشاهده کرد.*

[6] Iris Flower Data Set



شکل ۱: طبقه‌بندی داده‌های گل زنبق با چهار هسته رگرسیون بردار پشتیبان، نمودار بالا سمت چپ، نمودار طبقه‌بندی داده‌های مذکور با هسته چندجمله‌ای، نمودار بالا سمت راست، با هسته خطی، نمودار پایین سمت چپ با هسته سیگموید و نمودار پایین سمت راست با هسته شعاعی می‌باشد.

۵ مدل‌سازی داده‌های طیف‌سنجی

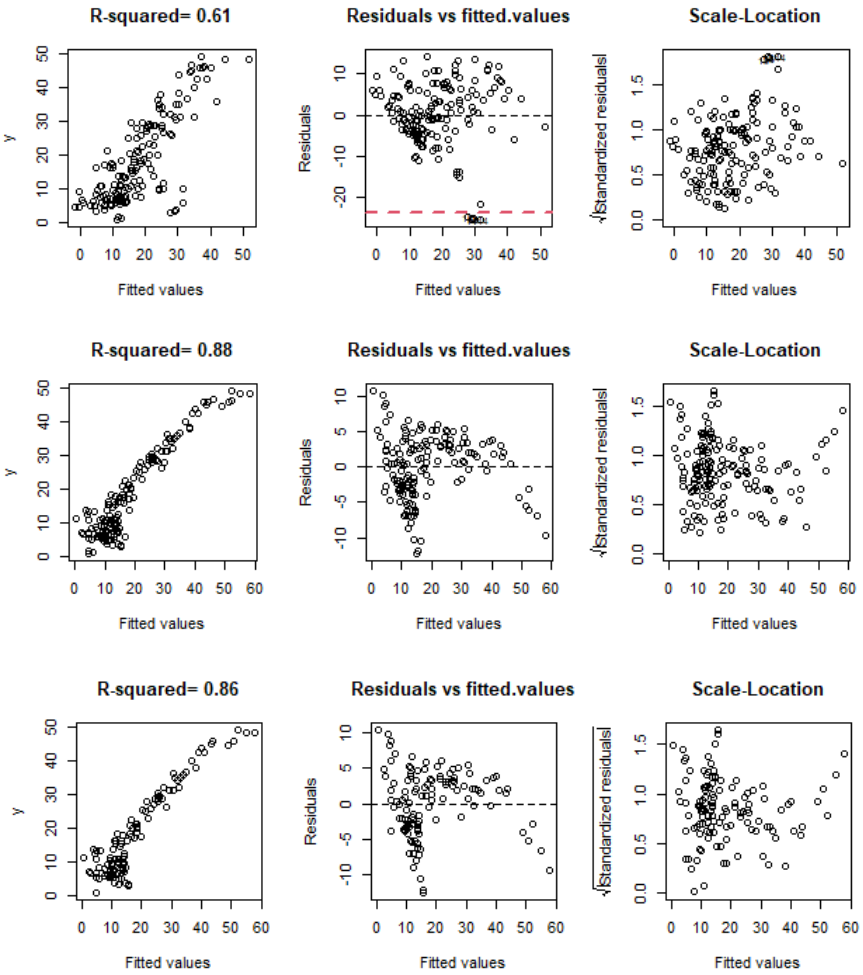
در این بخش به بررسی رگرسیون مؤلفه‌های اصلی با تاوان‌های لاسو، ستیغی و مشتق دوم و مقایسه آن با رگرسیون بردار پشتیبان با چهار هسته متفاوت پرداخته می‌شود. در این قسمت کلیه تحلیل‌ها در نرم‌افزار R انجام شده است. داده‌های طیف‌سنجی از بسته fda.usc در نرم‌افزار R اقتباس شده است. شامل ۱۰۰ طیف جذب مادون‌قرمز نزدیک است که برای پیش‌بینی مقدار رطوبت، چربی و پروتئین برای تیکه‌های گوشت خرد شده استفاده می‌شود. این داده‌ها بر اساس آنالیز غذا و خوراک Tecator Infratec می‌باشد، که در محدوده طول‌موج ۸۵۰ − ۱۰۵۰ نانومتر کار می‌کند. هدف پیش‌بینی محتوای چربی است. (برای اطلاعات بیشتر می‌توان به فراتی و همکاران مراجعه کرد [۹]). مدل تابعی برای داده‌های مذکور به‌صورت زیر نوشته می‌شود:

$$Fat = \beta_{۰}(t) + \sum_{i=۱}^{۱۰۰} \beta_i(t) X_i(t) \qquad (۲۰)$$

در مدل فوق Fat میزان چربی موجود در هر قسمت گوشت خالص و $X_i(t)$ اطلاعات ثبت‌شده توسط امواج مادون‌قرمز با طول‌موج ۸۵۰-۱۰۵۰ می‌باشد و t برابر ۱۰۰ طیف کانال می‌باشد. با استفاده از اعتبارسنجی متقابل ۵ تاخورده، داده‌ها به دو گروه آموزش و آزمون تقسیم‌شده‌اند. اکنون می‌توان مدل رگرسیون مؤلفهٔ اصلی تابعی را که از اعتبار سنجی متقابل به‌دست‌آمده، به‌صورت زیر نوشت:

$$Fat = \sum_{j} \alpha_j(t) PC_j(t) \qquad (۲۱)$$

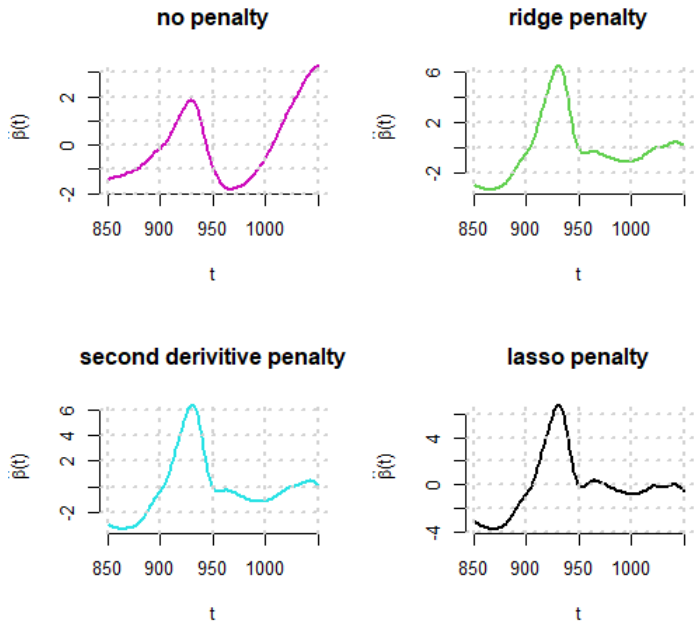
به‌طوری‌که در مدل فوق $PC_j(t)$ مؤلفه اصلی تابعی متغیرهای توضیحی می‌باشد که با استفاده از دستور fregre.pc در بسته fda.usc نرم‌افزار R قابل‌محاسبه است و $\alpha_j(t)$ها ضرایب مربوطه هستند. با استفاده از سه مؤلفهٔ اصلی اول با توجه به تأثیر بیشتر در مدل، می‌توان مدل‌سازی را انجام داد.



شکل ۲: نمودار ردیف اول: مدل رگرسیون مؤلفهٔ اصلی بدون تاوان ، نمودار ردیف دوم: مدل رگرسیون مؤلفهٔ اصلی با تاوان ستیغی، نمودار ردیف سوم: مدل رگرسیون مؤلفهٔ اصلی با تاوان مشتق دوم می‌باشد که نمودار سمت چپ: پراکنش مقادیر واقعی در مقابل مقادیر برازش شده، نمودار وسط: باقیمانده‌ها در مقابل مقادیر برازش شده و نمودار سمت راست: ریشه قدرمطلق باقیمانده‌ها در مقابل مقادیر برازش شده.

باقیمانده و باقیمانده استاندارد با مقادیر برازش شده برای سه مدل (بدون تاوان ، تاوان مشتق دوم و ستیغی) را می‌توان در شکل ۲ مشاهده کرد، عدد مجذور همبستگی در مدل با تاوان ستیغی مناسب بوده و بازه تغییرات باقیمانده داده‌ها بین منفی ده تا مثبت ده است که نسبت به دو مدل دیگر مناسب‌تر است.

اکنون به بررسی برآورد ضرایب تابعی در مدل‌های فوق پرداخته می‌شود. با توجه به شکل ۳ نمودار بالا سمت چپ برآورد ضرایب تابعی مدل بدون تاوان است که نمودار آن نسبت به بقیه هموار نیست و طول منحنی آن نسبت به بقیه بیشتر است و در شکل ۳ نمودار بالا سمت راست، نمودار برآورد ضرایب تابعی با تاوان ستیغی بر اساس سه مؤلفهٔ اصلی است که منحنی آن هموارتر از بدون تاوان می‌باشد. نمودار پایین سمت چپ شکل ۳ برآورد ضرایب تابعی با تاوان مشتق دوم، همانند تاوان ستیغی می‌باشد فقط در ابتدا و انتهای بازه عملکرد مناسب‌تری دارد و نمودار پایین سمت راست شکل ۳ برآورد ضرایب تابعی برای مدل رگرسیون مؤلفه اصلی با تاوان لاسو است که منحنی هموارتر و طول منحنی آن از بقیه کمتر می‌باشد و نتایج آن مناسب‌تر است. حال داده‌های آزمون را در مدل‌ها بررسی و سپس به انتخاب یکی از چهار مدل یعنی مدل بدون تاوان ، مدل با تاوان مشتق دوم و مدل با تاوان ستیغی و لاسو پرداخته می‌شود. می‌توان در شکل ۴ نمودارهای رسم شده مقادیر برازش شده در مقابل مقادیر واقعی برای داده‌های آزمون را مشاهده کرد. نتایج



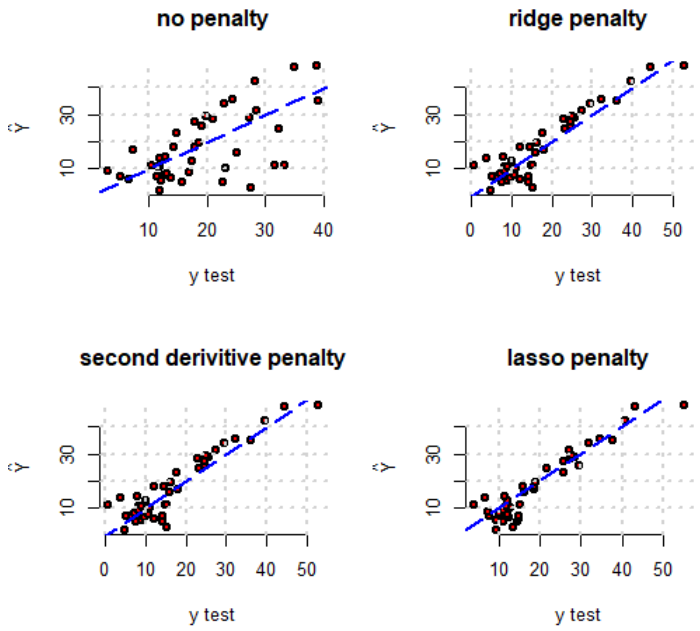
شکل ۳: نمودار برآورد ضرایب تابعی با تاوان‌های متفاوت

به‌دست‌آمده در جدول ۱ گویای آن است که با اختلاف خیلی کم مدل رگرسیون مؤلفهٔ اصلی با تاوان لاسو با توجه به عدد مجذور همبستگی ۰٫۹۴۴۳۰۹۶ که بالاتر و میانگین توان دوم خطا ۴٫۵۷۹۱۵۷ و میانگین انحراف درصد خطای مطلق ۰٫۲۷۹۴۲۸۸ که عددی پایین‌تر نسبت به نتایج سه مدل رگرسیون تابعی مؤلفهٔ اصلی بدون تاوان ، تاوان ستیغی و مشتق دوم می‌باشد مدل رگرسیون مؤلفهٔ اصلی تابعی با تاوان لاسو مناسب‌تر نسبت به سایر تاوان‌ها در این مدل عمل کرده است. اکنون با استفاده از رگرسیون بردار پشتیبان با چهار هسته متفاوت به برآورد داده‌های چربی پرداخته می‌شود. مدل رگرسیون بردار پشتیبان برای داده طیف‌سنجی به‌صورت زیر می‌توان نوشت:

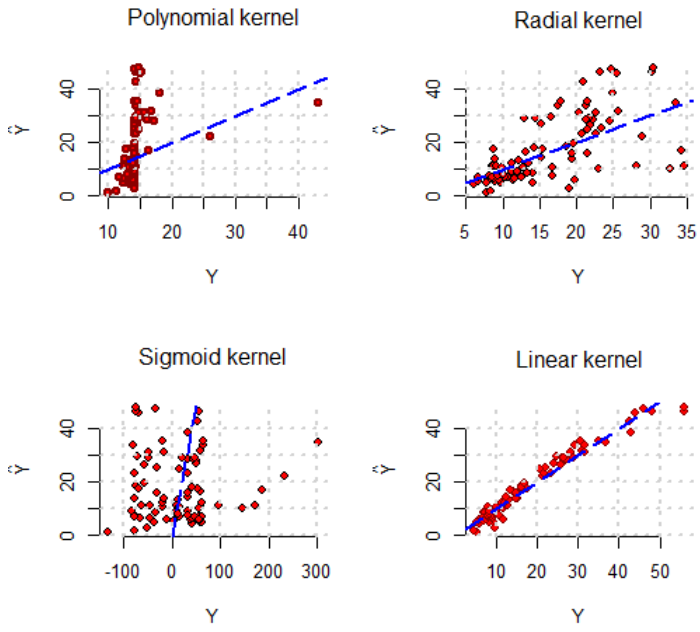
$$\boldsymbol{Fat} = \boldsymbol{W}\boldsymbol{X} + \epsilon \qquad (۲۲)$$

در مدل بالا $\boldsymbol{Fat}$ داده چربی می‌باشد، ماتریس $\boldsymbol{X}$ تعداد ۱۰۰ داده در ۲۱۵ طیف جذب مادون‌قرمز می‌باشد و $\boldsymbol{W}$ ضرایب مدل رگرسیون بردار پشتیبان می‌باشد. نتایج به‌دست‌آمده با استفاده از رگرسیون بردار پشتیبان، هسته خطی با پارامترهای $\gamma = ۰٫۱$ و $\epsilon = ۰٫۱$ و تعداد ۹۳ بردار پشتیبان برآوردی مناسب برای داده‌های آزمون از خود نشان داده است. شکل ۵ بیانگر رسم مقادیر واقعی در مقابل مقادیر برازش شده می‌باشد. هسته خطی در مقایسه با هسته‌های دیگر پراکندگی داده‌ها حوالی خط‌چین آبی‌رنگ بیشتر است.

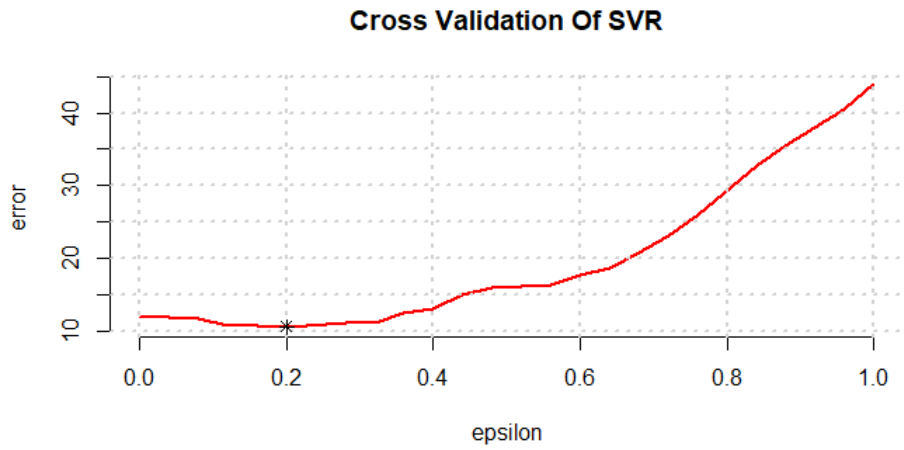
با تغییر میزان خطا و به دست آوردن خطای بهینه توسط اعتبار سنجی متقابل که در شکل ۶ نمایش داده شده است. عدد به‌دست‌آمده برای خطا ۰٫۲ می‌باشد. آزمون مقدار همبستگی به عدد ۰٫۹۷۸۱۰۰۴ تغییر کرد. نتایج مدل انتخاب‌شده را می‌توان در شکل ۷ مشاهده کرد که بیانگر مقادیر برازش شده در مقابل مقادیر واقعی برای داده‌های آزمون، با هسته خطی و خطای ۰٫۲ و با مشخص شدن متغیرهای بردار پشتیبان که تعداد آن‌ها ۴۴ می‌باشد. مقایسه بین مدل‌های رگرسیون بردار پشتیبان با هسته‌های متفاوت، همان‌طور که در جدول ۱ مشخص است، نتایج با هسته خطی تعمیم‌یافته مناسب‌تر نسبت به بقیه هسته‌ها می‌باشد.



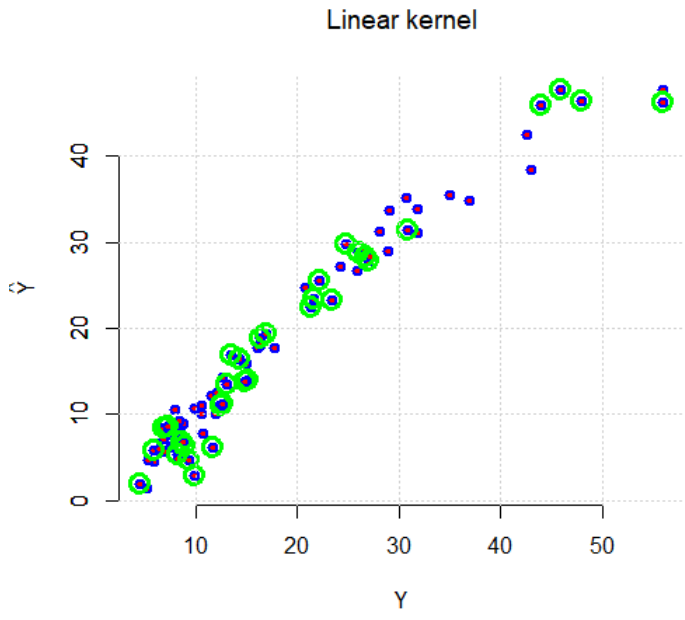
شکل ۴: نمودار مقادیر برازش شده در مقابل مقادیر واقعی برای داده‌های آزمون که نمودار بالا سمت چپ بدون تاوان ، نمودار بالا سمت راست با تاوان ستیغی، نمودار پایین سمت چپ با تاوان مشتق دوم و نمودار پایین سمت راست با تاوان لاسو می‌باشد.



شکل ۵: نمودار مقادیر برازش شده در مقابل مقادیر واقعی برای داده‌های آزمون که نمودار بالا سمت چپ رگرسیون بردار پشتیبان با هسته چندجمله‌ای، نمودار بالا سمت راست با هسته شعاعی، نمودار پایین سمت چپ با هسته سیگموئید و نمودار پایین سمت راست با هسته خطی می‌باشد.



شکل ۶: مقدار کمینه مشخص‌شده نشان‌دهنده مقدار خطای بهینه است.



شکل ۷: رسم مقادیر واقعی در مقابل مقادیر برازش شده برای مدل رگرسیون بردار پشتیبان با هسته خطی و خطای ۰٫۲، دایره‌های سبزرنگ نشان‌دهنده متغیرهای بردار پشتیبان است.

جدول ۱: جدول مقایسه بین مدل‌های رگرسیون مؤلفه اصلی تابعی و رگرسیون بردار پشتیبان

معیار	مجذور همبستگی	میانگین توان دوم خطا	میانگین انحراف درصد خطای مطلق
رگرسیون مؤلفهٔ اصلی تابعی بدون تاوان	۰٫۶۸۱	۹٫۵۸۰	۰٫۴۳۰۹
رگرسیون مؤلفهٔ اصلی تابعی با تاوان مشتق دوم	۰٫۹۳۲	۴٫۷۳۲	۰٫۷۸۵
رگرسیون مؤلفهٔ اصلی تابعی با تاوان ستیغی	۰٫۹۳۲	۴٫۷۳۲	۰٫۷۷۸
رگرسیون مؤلفهٔ اصلی تابعی با تاوان لاسو	۰٫۹۴۴	۴٫۵۷۹	۰٫۲۷۹
رگرسیون بردار پشتیبان با هسته خطی	۰٫۹۷۸	۲٫۸۰۷	۰٫۱۵۲
رگرسیون بردار پشتیبان با هسته چندجمله‌ای	۰٫۳۳۰	۱۲٫۶۵۸	۰٫۶۶۲
رگرسیون بردار پشتیبان با هسته سیگموید	۰٫۰۱۳	۷۶٫۲۸۱	۰٫۹۷۸
رگرسیون بردار پشتیبان با هسته گوسی	۰٫۶۳۷	۹٫۹۷۶	۰٫۴۱۹

۶ بحث و نتیجه‌گیری

در این مقاله، در ابتدا با تعریف دو مدل رگرسیون مؤلفه اصلی تابعی و رگرسیون بردار پشتیبان سپس به بررسی و تحلیل داده‌های طیف‌سنجی پرداخته شد. همان‌طور که در جدول ۱ مشخص است، با استفاده از معیارهای مجذور همبستگی، میانگین توان دوم خطا و میانگین انحراف درصد خطای مطلق به مقایسه بین مدل‌های موجود پرداخته شد؛ که در بین مدل‌های رگرسیون مؤلفهٔ اصلی تابعی بدون تاوان و با تاوان‌های مشتق دوم، ستیغی و لاسو و رگرسیون بردار پشتیبان با هسته‌های خطی، چندجمله‌ای، سیگموید و گوسی، مدل رگرسیون بردار پشتیبان با هسته خطی که با مقدار مجذور همبستگی ۰٫۹۷۸ که بالاترین مقدار بین بقیه مدل‌ها است. میانگین توان دوم خطا و میانگین درصد خطای مطلق در این مدل به ترتیب برابر با ۲٫۸۰۷ و ۰٫۱۵۲ که کمترین مقدار بین مابقی مدل‌ها است. در نتیجه بهترین عملکرد را در مقایسه با مدل‌های مذکور از خود نشان داده است.

تقدیر و تشکر

نویسندگان مقاله کمال قدردانی و تشکر را از پیشنهادهای ارزنده داوران، سردبیر و ویراستار محترم مجله که باعث ارائه بهتر و افزایش سطح کیفی مقاله شده است، دارند.

مراجع

[۱] روزبه، م؛ و امینی، م. (۱۳۹۸)، برآوردگر استوار مرزبندی شده تعمیم‌یافته محتمل در مدل رگرسیون نیمه‌پارامتری، *مجله علوم آماری ایران*، ۱۳(۲)، ۴۴۱-۴۶۰.

[۲] روزبه، م.، روحی، آ.، جهادی، ف؛ و زال زاده، س. (۱۴۰۰)، مدل رگرسیون بردار تکیه‌گاه و مقایسه آن با رگرسیون نیم‌پارامتری، *اندیشه آماری*، ۲۶(۲)، ۲۱-۳۲.

[۳] روزبه، م؛ و معنوی، م. (۱۳۹۹)، مدل‌سازی سن تقویمی به روش رگرسیون ستیغی کمترین توان های دوم پیراسته، *مجله علوم آماری ایران*، ۱۴(۲)، ۴۰۹-۴۲۸.

[4] Amini, M., and Roozbeh, M. (2015), Optimal Partial Ridge Estimation in Restricted Semiparametric Regression Models, *Journal of Multivariate Analysis*, **136**, 26-40.

[5] Araújo, R. D. A., Oliveira, A. L. , and Meira, S., (2015), A hybrid model for high-frequency stock market forecasting, *Expert Systems with Applications*, **42** , 4081-4096.

[6] Choudhury, S., Ghosh, S., Bhattacharya, A., Fernandes, K. J., and Tiwari, M. K., (2014), A real time clustering and SVM based price-volatility prediction for optimal trading strategy, *Neurocomputing*, **131**, 419-426.

[7] Dette, H., Kokot, K., and Aue, A., (2017), Functional data analysis in the Banach space of continuous functions, *arXiv*, preprint arXiv,1710.07781.

[8] Febrero-Band, M., Galeano, P., and Gonzalez-Manteiga, W., (2017), Functional principal component regression and functional partial least-squares regression: An overview and a comparative study, *International Statistical Review* , **85(1)**, 61-83.

[9] Ferraty, F., and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*, New York, Springer.

[10] Hoerl, A.E., and Kennard, R.W., (1975), Ridge regression: some simulation, *Communication in Statistics*, **4** , 4105–4123.

[11] Horveźth, L., and Kokoszka, P., (2012),*Inference for functional data with applications*, Springer Science and Business Media, New York.

[12] Hsing, T., and Eubank, R., (2015), *Theoretical foundations of functional data analysis, with an introduction to linear operators*, John Wiley and Sons, Hoboken.

[13] Jolliffe, I.T., (2002), *Principal Component Analysis*, Springer series in statistics, Aberdeen.

[14] Kao, L. J., Chiu, C. C., Lu, C. J., and Yang J. L., (2013), Integration of nonlinear independent component analysis and support vector regression for stock price forecasting, *Neurocomputing*, **99** , 534-542.

[15] Manahov, V., Hudson, R., and Gebka, B., (2014), Does high frequency trading affect technical analysis and marketefficiency and if so how, *Journal of International Financial Markets, Institutions and Money*, **28** , 131-157.

[16] Nayak, R. K., Mishra, D., and Rath, A. K., (2015), A naïve svm-knn based stock market trend reversal analysis for indian benchmark indices, *Applied Soft Computing*, **35** , 670-680.

[17] Patel, J., Shah, S., Thakkar, P., and Kotecha, K., (2015), Predicting stock market index using fusion of machine learning techniques, *Expert Systems with Applications*, **42** , 2162-2172.

[18] Ramsay, J. O., and Silverman, B. W., (2005),*Functional Data Analysis*, Springer-Verlag, New York.

[19] Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T., (2017), Methods for scalar-on-function regression, *International Statistical Review*, **85(2)** , 228-249.

[20] Roozbeh, M. (2018), Optimal QR-Based Estimation in Partially Linear Regression Models with Correlated Errors Using GCV Criterion, *Computational Statistics & Data Analysis*, **117**, 45-61.

[21] Tibshirani, R., (1996), *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological), **58(1)** , 267-288.

[22] Vapnik, V. N., (1995), *The Nature of Statistical Learning Theory*, New York.

[23] Xiao, Y., Xiao, J., Lu, F., and Wang, S., (2014), Ensemble anns-pso-ga approach for day-ahead stock e-exchange prices forecasting, *International Journal of Computational Intelligence Systems*, **7** , 272-290.

Functional principal component regression versus support vector regression for the analysis of spectroscopic data

Arta Rouhi[1], Fatemeh Jahadi[2] and Mahdi Roozbeh[3]

Abstract:

The most popular technique for functional data analysis is the functional principal component approach, which is also an important tool for dimension reduction. Support vector regression is branch of machine learning and strong tool for data analysis. In this paper by using the method of functional principal component regression based on the second derivative penalty, ridge and lasso and support vector regression with four kernels (linear, polynomial, sigmoid and radial) in spectroscopic data, the dependent variable on the predictor variables was modeled. According to the obtained results, based on the proposed criteria for evaluating the goodness of fit, support vector regression with linear kernel and error equal to 0.2 has had the most appropriate fit to the data set.

Keywords: Functional data analysis, Functional regression, Machine learning, Principal component regression, Support vector regression.

[1]**Master's degree graduate, statistics and Computer science, Semnan university, Semnan, Iran**

[2]**Master's degree graduate, statistics and Computer science, Semnan university, Semnan, Iran**

[3]**Faculty of mathematics, Semnan university, Semnan, Iran**