دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

یادگیری ماشین

تمرین شماره ۲

نام و نام خانوادگی

امیرحسین چمی

آذر ۱۴۰۴

فهرست

فهرست شکل‌ها

چکیده

این سری از تمرین‌ها ترکیبی از مباحث تئوری و تحلیلی به‌همراه استدلال ریاضی و شهودی بودند و هدف اصلی آن‌ها، درک عمیق‌تر روش‌های مختلف یادگیری، تخمین توزیع، رگرسیون، طبقه‌بندی و تحلیل رفتار الگوریتم‌ها در شرایط مختلف بود. در طول تمرین، هم با روش‌های کلاسیک مانند رگرسیون خطی و لجستیک و هم با روش‌های پیشرفته‌تری مثل کرنل‌ها، regularization و تفسیر بیزی مدل‌ها سروکار داشتیم.

حجم تمرین نسبتاً زیاد و شامل سوالات متنوعی بود که هر کدام به جنبه خاصی از یادگیری ماشین می‌پرداختند. انجام این تمرین‌ها نیازمند صرف زمان قابل توجه و دقت در جزئیات محاسباتی و مفهومی بود و در نهایت باعث شد ارتباط بین مباحث مختلف درس و تأثیر انتخاب پارامترها و مدل‌ها در عمل بهتر درک شود.

این سری از تمارین به قدری زیاد بودند که نمی توان در این چکیده به تک تک سوالات پرداخت و هدف آن ها را ریز به ریز بیان کرد اما صراحتا می توان گفت که این تمرین از تمرین قبلی عمیق تر بود و سعی می کرد مفاهیم (چه آنها که تدریس شده و چه آنها که تازه معرفی می شوند) را به صورت عمیق و پایه ای جا بیندازد.

پرسش ۱ – Parzen Windows

در این سوال داده هایی با توزیع یکنواخت $p(x) \sim U(0, a)$ و تابع کرنل زیر داریم.

$$\varphi(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases} \quad (1)$$

می خواهیم میانگین تخمین پنجره پارزن محاسبه و تاثیر پارامتر smoothing یعنی h_n را بررسی کنیم.

۱-۱. محاسبه‌ی میانگین تخمین پارزن

از جزوه ی درس می دانیم، توزیع تخمین زده شده به شکل زیر است.

$$p_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_n} \varphi\left(\frac{x - x_i}{h_n}\right) \quad (2)$$

و میانگین آن به صورت زیر بدست می آید.

$$\bar{p}_n(x) = E[p_n(x)] = \frac{1}{n} \sum_{i=1}^{n} E\left[\frac{1}{V_n} \varphi\left(\frac{x - x_i}{h_n}\right)\right] \quad (3)$$

با فرض i.i.d بودن داده های x_i تمامی expectation ها یکسان می شوند. پس داریم:

$$\bar{p}_n(x) = \frac{1}{n} . nE\left[\frac{1}{V_n} \varphi\left(\frac{x - y}{h_n}\right)\right] = \int \frac{1}{V_n} \varphi\left(\frac{x - y}{h_n}\right) p(y) dy \quad (4)$$

برای حل این انتگرال ابتدا باید تکلیف بازه های غیر صفر تابع کرنل و $p(y)$ را مشخص کنیم.

$$\varphi\left(\frac{x - y}{h_n}\right) \longrightarrow \frac{x - y}{h_n} > 0 \longrightarrow y < x \quad (5)$$

$$p(y) \longrightarrow 0 < y < a \quad (6)$$

پس در کل داده های y می توانند بین ۰ تا min(a,x) باشند تا توابع بالا مقدار غیر صفر داشته باشند. حال می توانیم انتگرال (۴) را بنویسیم.

$$\bar{p}_n(x) = \frac{1}{ah_n} \int_{0}^{\min(a,x)} e^{-\frac{(x-y)}{h_n}} dy = \frac{1}{ah_n} e^{-\frac{x}{h_n}} \int_{0}^{\min(a,x)} e^{\frac{y}{h_n}} dy \quad (7)$$

$$= \frac{1}{ah_n} e^{-\frac{x}{h_n}} h_n \left(e^{\frac{y}{h_n}}|_{0}^{\min(a,x)}\right)$$

پس در نهایت خواهیم داشت.

$$\bar{p}_n(x) = \frac{1}{a} e^{-\frac{x}{h_n}} \left(e^{\frac{\min(a,x)}{h_n}} - 1\right) \quad (8)$$

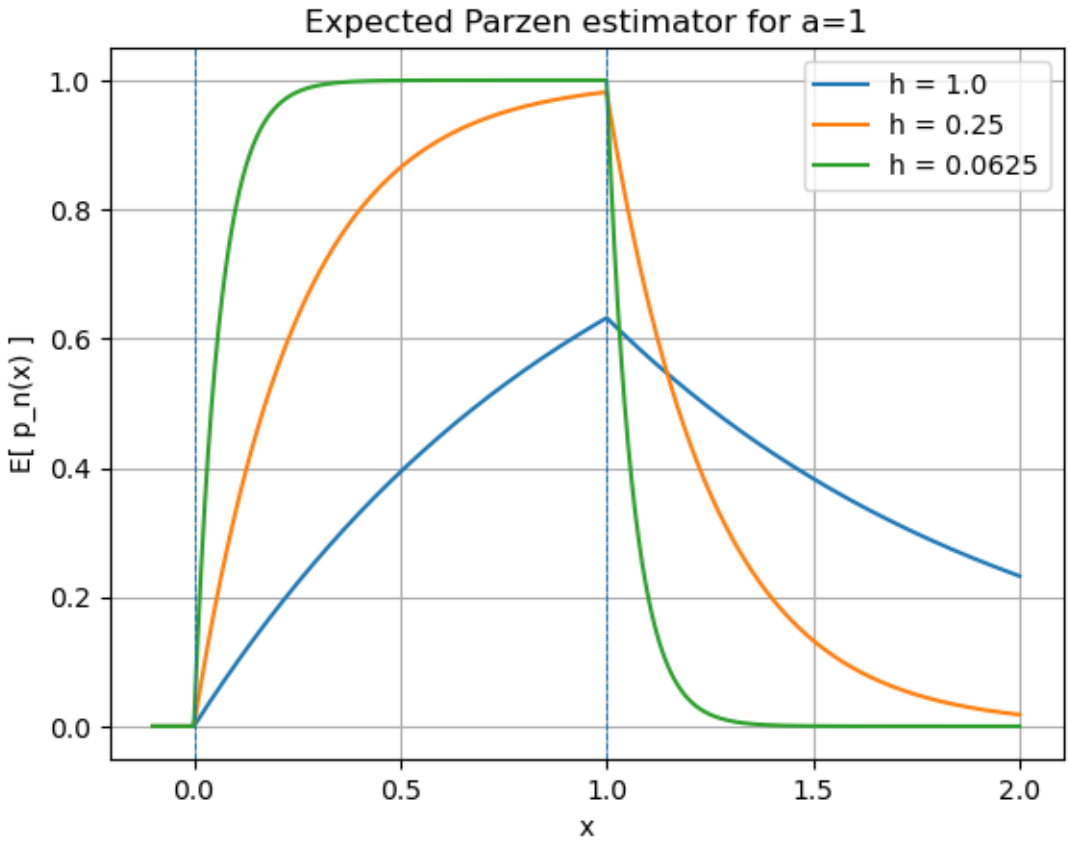
حال به ازای مقادیر مختلف x می توانیم (۸) را به یک تابع چند ضابطه ای تبدیل کنیم. اگر x<۰ آنگاه بازه‌ی [۰,x] اصلا قابل تعریف نیست. اگر $0 \le x \le a$ انگاه $\min(a,x) = x$ و اگر $x \ge a$ آنگاه $\min(a,x) = a$ می شود.

$$\bar{p}_n(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{a}\left(1 - e^{-\frac{x}{h_n}}\right) & 0 \le x \le a \\ \frac{1}{a}\left(e^{\frac{a}{h_n}} - 1\right) e^{-\frac{x}{h_n}} & x \ge a \end{cases} \qquad (9)$$

پس به این شکل میانگین مورد نظر حساب شد.

۱-۲. نمودار $\bar{p}_n(x)$ به ازای h_n های مختلف

در این بخش نمودار $\bar{p}_n(x)$ به x را با فرض ۰.۰۶۲۵ و ۰.۲۵, ۱ $= h_n$ و $a = 1$ رسم می کنم. (برای رسم خط های عمودی به ازای $x = 0$ و $x = a$ از مرجع [۱] استفاده کردم.)



شکل ۱. نمودار $\bar{p}_n(x) - x$

۱-۳. تنظیم h_n بر حسب میزان بایاس

مقدار $E[p_n(x)]$ در بازه‌ی $0 \le x \le a$ را در رابطه‌ی (۹) دیدیم. بر این اساس مقدار بایاس به صورت زیر است.

$$Bias = E[p_n(x)] - p(x) = \frac{1}{a}\left(1 - e^{-\frac{x}{h_n}}\right) - \frac{1}{a} = -\frac{1}{a}e^{-\frac{x}{h_n}} \quad (10)$$

پس در کل اندازه‌ی بایاس می شود: $|Bias| = \frac{1}{a}e^{-\frac{x}{h_n}}$

اندازه‌ی بایاس باید کمتر از یک درصد باشد. یعنی کمتر از یک درصد چگالی واقعی.

$$|Bias| < 0.01p(x) = 0.01\frac{1}{a} \quad (11)$$

$$\frac{1}{a}e^{-\frac{x}{h_n}} < 0.01\frac{1}{a} \quad (12)$$

$$e^{-\frac{x}{h_n}} < 0.01$$

پس با لگاریتم گیری خواهیم داشت.

$$-\frac{x}{h_n} < \ln(0.01) = -4.605170186 \quad (13)$$

$$h_n < \frac{x}{4.605170186} \quad (14)$$

بازه‌ی $[0, a]$ طولی به اندازه‌ی a دارد. باید این بایاس برای هر بازه‌ی $[0.01a, a]$ برقرار باشد. کمترین مقدار این بازه $x_{min} = 0.01a$ است. پس رابطه‌ی (۱۴) به صورت زیر خواهد بود.

$$h_n < \frac{0.01a}{4.605170186} \quad (15)$$

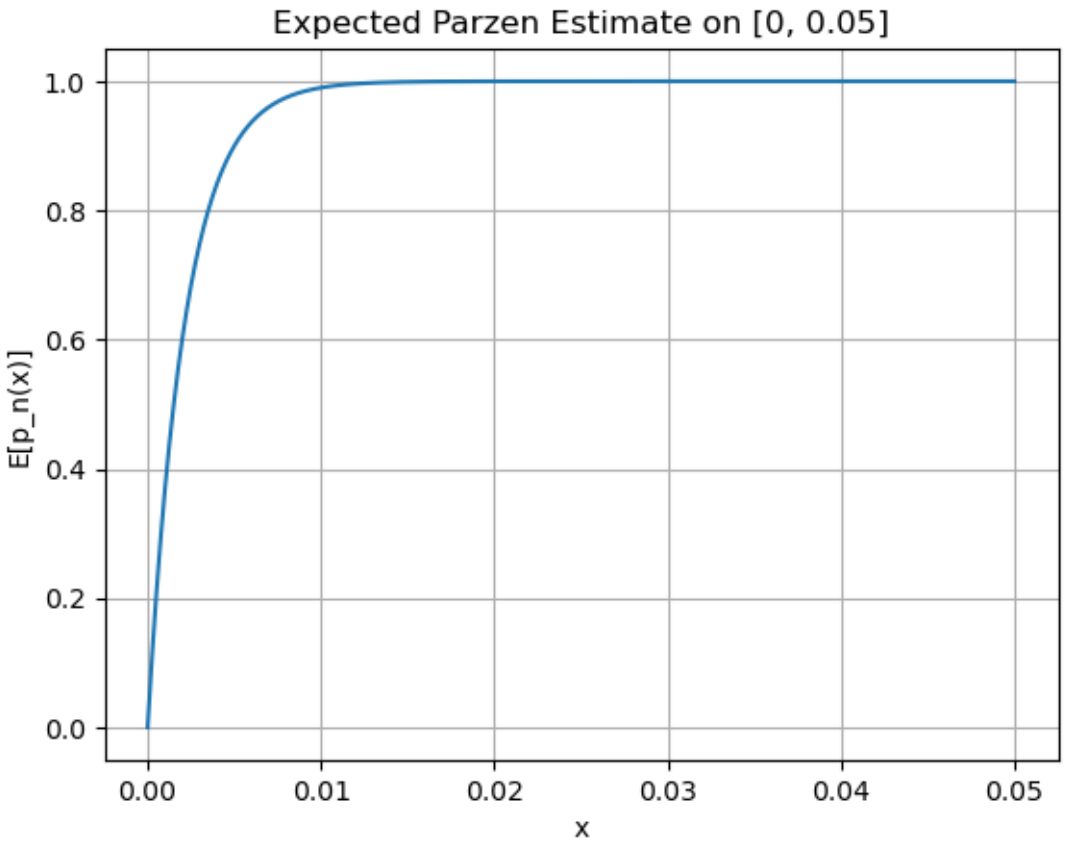
طبق رابطه‌ی (۱۵) حداکثر مقدار h_n بدست آمد. از طرفی طبق شرایط پارزن که در جزوه ی درس هم دیدیم، $\lim_{n\to\infty} h_n \to 0$ و از طرفی $\lim_{n\to\infty} nh_n \to \infty$ است. این به این معنی است که کمترین مقدار h_n صفر است ولی به صورت عملی ممکن نیست.

۱-۴. نمودار $\bar{p}_n(x)$ به ازای h_n مشخص در بازه‌ی خاص

در رابطه‌ی (۱۵) حداکثر مقدار h_n حساب شد. با فرض $a = 1$ خواهیم داشت.

$$h_n < \frac{0.01}{4.605170186} = 0.002171472 \quad (16)$$

با استفاده از مقدار به دست آمده، $\bar{p}_n(x)$ در بازه‌ی $[0, 0.005]$ به صورت زیر است.



شکل ۲. نمودار $\bar{p}_n(x)$ به ازای مقادیر مشخص در بازه‌ی $[0, 0.05]$

پرسش ۲ – Shift of the Eigenvalue Spectrum

هدف این سوال همان طور که صراحتا گفته شده، مروری بر جبر خطی است.

۲-۱. اثبات ویژگی بردار ویژه

می دانیم اگر ξ بردار ویژه برای ماتریس های A و B باشد باید روابط $A\xi = \lambda_A\xi$ و $B\xi = \lambda_B\xi$ برقرار باشد. پس برای اثبات باید به رابطه ای به همین شکل برسیم.

$$(A+B)\xi = A\xi + B\xi = \lambda_A\xi + \lambda_B\xi = (\lambda_A + \lambda_B)\xi \qquad (17)$$

به همین سادگی اثبات شد که ξ بردار ویژه‌ی ماتریس $A+B$ هم هست.

۲-۲. چرا استفاده از $\hat{\beta}^{ridge}$ مناسب تر است

می خواهیم نشان دهیم چرا چرا استفاده از $\hat{\beta}^{ridge}$ مناسب تر است. ابتدا با استفاده از تجزیه به مقادیر ویژه، ماتریس X^TX که قطعا متقارن است را تجزیه می کنیم.

$$X^TX = Q\Lambda Q^T \qquad (18)$$

در این رابطه Q یک ماتریس ارتوگونال است که ستون های بردارهای ویژه ماتریس X^TX هستند. و $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ یک قطری است.

حال از نتیجه‌ی بخش ۲-۱ استفاده می کنیم. فرض می کنیم $A = X^TX$ و $B = \lambda I$ است. توجه کنید که هر برداری در $\mathbb{R}^n$ یک بردار ویژه برای ماتریس I است. بر این اساس برای هر برداریژه‌ی q_i مقدار ویژه‌ی متناظر با آن در ماتریس $X^TX + \lambda I$ برابر $\lambda_i + \lambda$ است.

$$(X^TX + \lambda I)q_i = X^TXq_i + \lambda Iq_i = \lambda_i q_i + \lambda q_i = (\lambda_i + \lambda)q_i \qquad (19)$$

چه فایده ای برای ما دارد؟

- اضافه شدن λI باعث افزایش مقادیرویژه می شود در حالی برداریژه ها بدون تغییر می مانند.
- اگر ماتریس X^TX تکین یا نزدیک به تکین باشد (یعنی بعضی از مقادیرویژه صفر یا نزدیک صفر باشند.) باعث می شود که معکوس این ماتریس وجود نداشته باشد یا اینکه مقادیر نزدیک صفر باعث مقادیر بسیار بزرگ در معکوس شوند که همین باعث ناپایداری می شود. به همین دلیل باید λ مقدار مثبت و به قدر کافی بزرگ داشته باشد تا اطمینان داشته باشیم $X^TX + \lambda I$ همواره معکوس پذیر باشد.

- مورد سوم (که بی ربط به مورد دوم نیست) این است که مقدار condition number افزایش می یابد. طبق [۲] مقدار condition number معیاری است برای نشان دادن میزان حساسیت یک تابع به تغییرات در ورودی. برای ماتریس متقارن X^TX مقدار condition number به صورت $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ تعریف می شود که اگر $\lambda_{min} \approx 0$ در این صورت مقدار κ بسیار بزرگ خواهد شد یعنی مقادیر کوچک می توانند باعث خطا های بسیار بزرگ شوند و ناپایداری ایجاد کنند. در نتیجه استفاده از $X^TX + \lambda I$ مقدار condition number را کاهش داده و حساسیت به تغییر ورودی و در نتیجه خطا را کم می کند و از ناپایداری جلوگیری می کند.

۲-۳. مقایسه‌ی رگرسیون خطی و ریج

با فرض اینکه مدل به صورت $y = X\beta + \epsilon$ بوده و $E\{\epsilon\} = 0$ و $var(\epsilon) = \sigma^2 I$ باشد. رگرسیون خطی از به صورت زیر است.

$$\hat{\beta} = (X^TX)^{-1}X^Ty \tag{20}$$

امیدریاضی β به صورت زیر محاسبه می شود.

$$\begin{aligned} E\{\hat{\beta}\} &= E\{(X^TX)^{-1}X^Ty\} = E\{(X^TX)^{-1}X^T(X\beta + \epsilon)\} \\ &= E\{(X^TX)^{-1}X^TX\beta\} + E\{(X^TX)^{-1}X^T\epsilon\} \\ &= E\{\beta\} + E\{(X^TX)^{-1}X^T\}E\{\epsilon\} = E\{\beta\} = \beta \end{aligned} \tag{21}$$

پس $bias = E\{\hat{\beta}\} - \beta = 0$. این روش بایاس ندارد. واریانس آن به صورت زیر است.

$$\begin{aligned} var(\hat{\beta}) &= var((X^TX)^{-1}X^Ty) = var\big((X^TX)^{-1}X^T(X\beta + \epsilon)\big) \\ &= var((X^TX)^{-1}X^T\epsilon) \\ &= (X^TX)^{-1}X^T(\sigma^2 I)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1} \end{aligned} \tag{22}$$

اگر X^TX تکین یا نزدیک به تکین باشد، بعضی مقادیر ویژه‌ی آن خیلی کوچک بوده و در نتیجه $(X^TX)^{-1}$ مقادیری بسیار بزرگی خواهد داشت. پس این واریانس می تواند خیلی بزرگ باشد.

پس در کل با وجود اینکه رگرسیون خطی بایاس ندارد، اما می تواند واریانس خیلی زیادی داشته باشد.

اما در رگرسیون ریج داریم:

$$\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^Ty \tag{23}$$

امیدریاضی آن به صورت زیر است.

$$E\{\hat{\beta}_{ridge}\} = E\{(X^TX + \lambda I)^{-1}X^Ty\} \quad (24)$$
$$= E\{(X^TX + \lambda I)^{-1}X^T(X\beta + \epsilon)\}$$
$$= E\{(X^TX + \lambda I)^{-1}X^TX\beta\} \neq \beta$$

واضحا این امیدریاضی β نیست. پس این روش بایاس غیر صفر دارد. این ویژگی مثبتی است. واریانس آن به صورت زیر است.

$$var\left(\hat{\beta}_{ridge}\right) = var\left(\left(X^TX + \lambda I\right)^{-1}X^Ty\right) \quad (25)$$
$$= var\left(\left(X^TX + \lambda I\right)^{-1}X^T(X\beta + \epsilon)\right)$$
$$= var\left(\left(X^TX + \lambda I\right)^{-1}X^T\epsilon\right)$$
$$= \left(X^TX + \lambda I\right)^{-1}X^T(\sigma^2 I)\,X\left(X^TX + \lambda I\right)^{-1}$$
$$= \sigma^2\left(X^TX + \lambda I\right)^{-1}X^TX\left(X^TX + \lambda I\right)^{-1}$$

می توانیم این واریانس را با استفاده از تجزیه به مقادیر ویژه محاسبه کنیم. با فرض تجزیه ویژه‌ی $X^TX = U\Lambda U^T$ می توان واریانس رگرسیون ریج را به صورت زیر نوشت.

$$var(\hat{\beta}_{ridge}) = \sigma^2\frac{\lambda_i}{(\lambda_i + \lambda)^2} \quad (26)$$

در حالی که با استفاده از تجزیه ویژه برای واریانس رگرسیون خطی خواهیم داشت:

$$var(\hat{\beta}) = \frac{\sigma^2}{\lambda_i} \quad (27)$$

واضح است که واریانس رگرسیون ریج کمتر از رگرسیون خطی است. پس در کل رگرسیون ریج بایاس غیر صفر دارد اما واریانس کمتر از رگرسیون خطی است.

اگر بایاس صفر برای ما اهمیت داشته باشد به سراغ رگرسیون خطی می رویم و اگر واریانس کمتر مدل اهمیت داشته باشد باید به سراغ رگرسیون ریج برویم.

پرسش ۳ – Optimizing a Kernel

این سوال درباره‌ی رگرسیون مبتنی بر کرنل است. هدف این است که تابع هزینه را برای این سوال بدست آورده و به ازای ۲ بعد هم تحلیل کنیم.

۳-۱. بخش اول: محاسبه‌ی تابع Loss

طبق صورت سوال تابع کرنل به صورت زیر است.

$$K(x_n, x_i) = \exp(-(x_n - x_i)W(x_n - x_i)^T) \quad (۲۸)$$

خود سوال برای این بخش یک hint قرار داده که با توجه به آن باید جمع کرنل را روی تمام داده ها به جز خود داده x_i قرار دهیم. تابع پیش بینی به صورت زیر است.

$$f(x_i) = \frac{\sum_{n \neq i} K(x_n, x_i)\, y_n}{\sum_{n \neq i} K(x_n, x_i)} \quad (۲۹)$$

با توجه به اینکه می خواهیم از residual sum of squares استفاده کنیم پس به صورت زیر loss را تعریف می کنیم.

$$L(W) = \sum_{i=۱}^{N} \left(y_i - f(x_i)\right)^۲ = \sum_{i=۱}^{N} \left(y_i - \frac{\sum_{n \neq i} K(x_n, x_i)\, y_n}{\sum_{n \neq i} K(x_n, x_i)} \right)^۲ \quad (۳۰)$$

$$= \sum_{i=۱}^{N} \left(y_i - \frac{\sum_{n \neq i} \exp(-(x_n - x_i)W(x_n - x_i)^T)\, y_n}{\sum_{n \neq i} \exp(-(x_n - x_i)W(x_n - x_i)^T)} \right)^۲$$

۳-۲. بخش دوم: محاسبه به ازای ۲ بعد

با توجه به اینکه بعد مسئله ۲ است پس هر داده به صورت $x_n = (x_n^{(۱)}, x_n^{(۲)})$ تعریف می شود. عبارتی که در توان exponential داریم به صورت زیر نوشته خواهد شد.

$$(x_n - x_i)W(x_n - x_i)^T \quad (۳۱)$$
$$= W_{۱۱}\left(x_n^{(۱)} - x_i^{(۱)}\right)^۲ + ۲W_{۱۲}\left(x_n^{(۱)} - x_i^{(۱)}\right)\left(x_n^{(۲)} - x_i^{(۲)}\right)$$
$$+ W_{۲۲}\left(x_n^{(۲)} - x_i^{(۲)}\right)^۲$$

پس بر این اساس تابع loss به صورت زیر است.

$$L(W) = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 \qquad (۳۲)$$

$$= \sum_{i=1}^{N} \Bigg(y_i - \frac{\sum_{n \neq i} \exp\left(-W_{11}\left(x_n^{(1)} - x_i^{(1)}\right)^2 - 2W_{12}\left(x_n^{(1)} - x_i^{(1)}\right)\left(x_n^{(2)} - x_i^{(2)}\right) - W_{22}\left(x_n^{(2)} - x_i^{(2)}\right)^2\right) y_n}{\sum_{n \neq i} \exp\left(-W_{11}\left(x_n^{(1)} - x_i^{(1)}\right)^2 - 2W_{12}\left(x_n^{(1)} - x_i^{(1)}\right)\left(x_n^{(2)} - x_i^{(2)}\right) - W_{22}\left(x_n^{(2)} - x_i^{(2)}\right)^2\right)} \Bigg)^2$$

(برای اینکه بتوان کسر را به صورت کامل نشان داد مجبور شدم سایز فونت را کوچک تر کنم.)

برای اینکه کوتاه تر بنویسیم می توان متغیر های $\Delta_{ni}^{(1)} = x_n^{(1)} - x_i^{(1)}$ و $\Delta_{ni}^{(2)} = x_n^{(2)} - x_i^{(2)}$ تعریف کرد. بر این اساس می توان کوتاه تر نوشت.

$$L(W) = \sum_{i=1}^{N} \left(y_i - \frac{\sum_{n \neq i} \exp\left(-W_{11}\left(\Delta_{ni}^{(1)}\right)^2 - 2W_{12}\left(\Delta_{ni}^{(1)}\right)\left(\Delta_{ni}^{(2)}\right) - W_{22}\left(\Delta_{ni}^{(2)}\right)^2\right) y_n}{\sum_{n \neq i} \exp\left(-W_{11}\left(\Delta_{ni}^{(1)}\right)^2 - 2W_{12}\left(\Delta_{ni}^{(1)}\right)\left(\Delta_{ni}^{(2)}\right) - W_{22}\left(\Delta_{ni}^{(2)}\right)^2\right)} \right)^2 \qquad (۳۳)$$

(برای اینکه بتوان کسر را به صورت کامل نشان داد مجبور شدم سایز فونت را کوچک تر کنم.)

پرسش ۴ – KNN and the Curse of Dimensionality

در این سوال مفهوم نفرین (نحسی) ابعاد را می فهمیم.

۴-۱. بخش a: مقدار مشاهدات مورد استفاده برای پیش‌بینی با $p = 1$

در این بخش با توجه به مثالی که در متن سوال زده شده، در مجموع می توانیم سه حالت را برای هر داده‌ی تست x در نظر بگیریم.

$$\begin{gathered} ۰ \leq x < ۰.۰۵ \\ ۰.۰۵ \leq x \leq ۰.۹۵ \\ ۰.۹۵ < x < ۱ \end{gathered} \tag{۳۴}$$

در واقع بازه ی ۰ الی ۱ به سه بخش تقسیم شده. بازه ی $۰.۰۵ \leq x \leq ۰.۹۵$ بازه ای است که اگر داده‌ی مشاهده در آن قرار گیرد می توانیم به اندازه‌ی ۰.۰۵ کمتر یا بیشتر از آن را بررسی کنیم. اما دو بازه‌ی دیگر جایی هستند که اگر مشاهده در آن قرار گیرد، با کم یا اضافه کردن ۰.۰۵ از بازه‌ی [۰,۱] که مسئله در آن تعریف شده است خارج می شویم. هر بازه‌ی بالا را می توان به ازای مشاهده‌ی x به صورت زیر بازنویسی کرد.

$$\begin{gathered} [۰, x + ۰.۰۵] \quad with\ length\ x + ۰.۰۵ \\ [x - ۰.۰۵, x + ۰.۰۵] \quad with\ length\ ۰.۱۰ \\ [x - ۰.۰۵, ۱] \quad with\ length\ ۱.۰۵ - x \end{gathered} \tag{۳۵}$$

از آنجا که توزیع x به صورت یکنواخت روی بازه ی ۰ الی ۱ است، میانگین داده های مشاهده‌ی مورد نظر با انگرال های زیر بدست می آید.

$$\begin{aligned} E[L] &= \int_{۰}^{۰.۰۵} (x + ۰.۰۵)\, dx + \int_{۰.۰۵}^{۰.۹۵} ۰.۱۰\, dx + \int_{۰.۹۵}^{۱} (۱.۰۵ - x)\, dx \\ &= \left(\frac{x^۲}{۲} + ۰.۰۵x\right)_{۰}^{۰.۰۵} + (۰.۱x)_{۰.۰۵}^{۰.۹۵} + \left(۱.۰۵x - \frac{x^۲}{۲}\right)_{۰.۹۵}^{۱} \\ &= ۰.۰۰۳۷۵ + ۰.۰۹ + ۰.۰۰۳۷۵ = ۰.۰۹۷۵ \end{aligned} \tag{۳۶}$$

پس ۰.۰۹۷۵ یا ٪۹.۷۵ از داده های موجود برای پیش‌بینی استفاده می شوند. البته به صورت شهودی می توان گفت که با توجه به یکنواخت بودن توزیع داده ها و با توجه به اینکه بازه‌ی $۰.۰۵ \leq x \leq ۰.۹۵$ (که مقدار بیشتری داده نسبت به بازه‌ی های دیگر را شامل می شود) دقیقا ٪۹۰ فضای مسئله است، پس اگر از بازه های کوچک نزدیک *boundary* ها چشم پوشی کنیم می توان گفت که به تقریب به صورت میانگین از ٪۱۰ داده های موجود برای پیش‌بینی استفاده می کنیم. (منظورم این است که مقدار حدودی ٪۱۰ درصد قبل از محاسبات هم قابل حدس بود.)

۴-۲. بخش b: مقدار مشاهدات مورد استفاده برای پیش‌بینی با $p = 2$

در بخش قبلی به ازای یک ویژگی دیدیم که ۹.۷۵٪ داده ها مورد استفاده قرار می گیرد. حال با اضافه شدن ویژگی دوم و با فرض مستقل بودن ویژگی، میانگین داده های مورد استفاده به صورت زیر خواهد بود.

$$E[L_1 L_2] = E[L_1]E[L_2] = ۰.۰۹۷۵ * ۰.۰۹۷۵ = \ ۰.۰۰۹۵۰۶۲۵ \qquad (۳۷)$$

پس ۰.۰۰۹۵۰۶۲۵ یا ٪۰.۹۵۰۶۲۵ داده های موجود برای پیش‌بینی استفاده می شوند. خیلی عدد کمتری شد. شهود آن را به ازای یک ویژگی در بخش قبل دیدیم و بر اساس آن به صورت شهودی به ازای دو ویژگی تنها ۰.۰۱ یا ٪۱ از داده ها استفاده می شوند.

۴-۳. بخش c: مقدار مشاهدات مورد استفاده برای پیش‌بینی با $\mathbf{p = 100}$

مشابه بخش های قبل به ازای ۱۰۰ تا ویژگی مستقل میانگین داده های مورد استفاده به صورت زیر خواهد بود.

$$average\ fraction = (E[L])^{۱۰۰} = ۰.۰۹۷۵^{۱۰۰} \qquad (۳۸)$$

که برای همین حالت با استفاده از روش شهودی خواهد بود:

$$average\ fraction = ۰.۱^{۱۰۰} \qquad (۳۹)$$

پس $۰.۱^{۱۰۰}$ یا $٪۱۰^{-۱۰۰}$ داده های موجود برای پیش‌بینی استفاده می شوند. می بینیم که عدد به دست آمده مقدار بسیار کمی شده.

۴-۴. بخش d: نتیجه گیری از بخش های قبل

در بخش های قبل دیدیم که به ازای یک ویژگی ٪۱۰ از داده ها در نزدیکی نقطه‌ی تست هستند و این مقدار برای دو ویژگی تنها ٪۱ و برای ۱۰۰ ویژگی مقدار بسیار کم $٪ ۱۰^{-۱۰۰}$ است. با اضافه شدن ویژگی ها، داده های نزدیک به داده‌ی تست به صورت نمایی کاهش پیدا می کنند.

فرض روش KNN این است که ما می توانیم تعداد مناسبی داده‌ی train در نزدیکی داده‌ی تست پیدا کنیم. اما وقتی تعداد ویژگی ها زیاد می شود، تقریبا هیچ داده ای در همسایگی نزدیک داده‌ی تست وجود ندارد. به عبارت دیگر نزدیک ترین همسایه در ابعاد بالا، بسیار دور است. پس باید تعداد داده های ما بی نهایت باشد که ممکن نیست یا اینکه برای پیدا کردن نزدیک ترین داده باید منطقه ی بسیار بزرگی از فضای ویژگی ها را بگردیم که این دیگر مفهوم نزدیکی و محلی بودن (locality) که در KNN مورد نظر بود را ندارد.

پس می توان گفت KNN بر اساس locality کار می کند ولی در ابعاد بالا اصلا locality تعریف درستی ندارد. (تعریف نمی شود) و داده ای هم در نزدیکی داده ی تست وجود ندارد.

پرسش ۵ – Choosing Metrics

این سوال مستقیما در مورد انتخاب معیار محاسبه‌ی فاصله از مباحث KNN است. در این سوال با معیار های مختلف آشنا شده و در هر بخش بر حسب کاربرد بهترین را انتخاب می کنیم.

برای مرور فاصله ی اقلیدسی از [۳]، آشنایی با فاصله‌ی منهتن از [٤]، آشنایی با فاصله‌ی همینگ از [٥] و آشنایی با فاصله‌ی cosine از [٦] استفاده کردم. برای هر کدام از این معیار ها تعریفی در نظر گرفته می شود که به صورت زیر است:

- فاصله‌ی اقلیدسی: فاصله‌ی اقلیدسی کوتاه‌ترین فاصله‌ی مستقیم بین دو نقطه در فضای ویژگی‌ها است. این فاصله تفاوت دو نمونه را با در نظر گرفتن اندازه‌ی واقعی اختلاف‌ها در هر بعد اندازه‌گیری می‌کند.
- فاصله‌ی منهتن: فاصله دو نقطه را به جای کوتاه‌ترین خطی که آن‌ها را بهم متصل می‌کند، به صورت خطوط افقی و عمودی که آن دو نقطه را بهم می‌رساند نشان می‌دهد.
- فاصله‌ی همینگ: در تئوری اطلاعات فاصله‌ی همینگ برای دو رشته با طول مساوی، برابر تعداد مکان‌هایی است که سمبل‌های متناظر متفاوت هستند. به عبارت دیگر فاصله‌ی همینگ تعداد ویژگی‌هایی را می‌شمارد که بین دو نمونه متفاوت هستند. این معیار فقط روی ویژگی‌های دسته‌ای یا باینری (۱/۰، True/False، A/B/...) قابل استفاده است، نه داده‌های پیوسته.
- فاصله‌ی Cosine: فاصله‌ی کسینوسی بر پایه‌ی زاویه بین دو بردار تعریف می‌شود(ضرب داخلی) و اندازه‌ی واقعی اهمیت ندارد. این معیار وقتی مفید است که الگوی توزیع ویژگی‌ها مهم‌تر از مقدار دقیق آن‌ها باشد؛ بنابراین فقط جهت بردار مهم است.

۵-۱. بخش اول: طبقه بندی سه کلاس نوشیدنی

معیار مناسب برای این طبقه بندی استفاده از فاصله‌ی اقلیدسی است. در ادامه هر معیار را تحلیل می کنم.

گزینه‌ی ۱: ویژگی ها در این مسئله به صورت عددی و پیوسته هستند.(مقدار کافئین و قند) این داده ها در فضای دو بعدی تعریف خواهند شد.

فاصله‌ی اقلیدسی، فاصله‌ی واقعی نقاط را در فضای ویژگی ها اندازه گیری می کند. برای مثال نوشیدنی که کمی کافئین و مقدار زیادی قند دارد واقعا از نوشیدنی با هر دو مقدار بالا فاصله‌ی بیشتری دارد.

گزینه‌ی ۲: استفاده از فاصله‌ی منهتن زمانی خوب است که داده ها ساختار شبکه ای دارند. یعنی دقیقا بتوان تعریفی که از فاصله‌ی منهتن ارائه شد را روی آن پیاده سازی کرد.

ولی کافئین و قند چنین حالت شبکه ای را تشکیل نمی دهند. با این وجود، شاید با گسسته کردن مقدار کافئین و قند بتوان حالت شبکه ای را به صورت مصنوعی ایجاد کرد و از فاصله‌ی منهتن بهره برد ولی این کار بهینه نخواهد بود.

گزینه‌ی ۳: در این معیار فقط می گوییم که "این ویژگی فرق دارد یا خیر". به وضوح این مورد برای داده های Boolean تعریف می شود نه برای مسئله ی ما که ویژگی ها عددی و پیوسته هستند.

ما نیاز داریم بدانیم چقدر کافئین یا قند تفاوت دارد بنابراین همینگ در اینجا کاربردی ندارد.

گزینه‌ی ۴: از این معیار وقتی استفاده می شود که فقط نسبت ویژگی ها مهم است. همانطور که قبلا اشاره شد در این مسئله مقدار واقعی کافئین و قند مهم است نه نسبت آنها.

برای مثال اگر نوشیدنی اول ۱۰ واحد کافئین و ۵ واحد قند داشته و نوشیدنی دوم ۲۰۰ واحد کافئین و ۱۰۰ واحد قند داشته باشد، طبق معیار cosine این دو نوشیدنی یکسان هستند زیرا نسبت ها یکسان است. پس این معیار برای این مسئله مناسب نیست.

نتیجه نهایی: بهترین معیار برای این مسئله استفاده از فاصله‌ی اقلیدسی (گزینه‌ی ۱) است.

۵-۲. بخش دوم: طبقه بندی چهار کلاس نوشیدنی

معیار مناسب برای این طبقه بندی استفاده از فاصله‌ی اقلیدسی است. در ادامه هر معیار را تحلیل می کنم.

گزینه‌ی ۱: توضیحی که خود سوال برای این گزینه اراده کرده است تحلیل خوبی است. کلاس جدید آب با مقادیر (۰و۰) اضافه شده است. این مقدار باعث می شود که نسبت ها بی معنی شوند. به عبارت دیگر فاصله‌ی واقعی در فضای دو بعدی اهمیت بیشتری پیدا می کند و اقلیدسی معیاری است که مقدار واقعی ویژگی ها را درست در نظر می گیرد.

گزینه‌ی ۲: اضافه شدن آب همچنان هیچ کمکی نمی کند که کلاس های مسئله حالت شبکه ای را ایجاد کنند. همچنان تحلیلی که در بخش اول برای فاصله‌ی منهتن ارائه شد درست است یعنی؛ می توان از این معیار استفاده کرد ولی به وضوح معیار بهینه ای نیست.

گزینه‌ی ۳: همچنان با اضافه شدن کلاس آب، ویژگی ها عددی و پیوسته هستند و حالت Boolean ندارند. پس همینگ معیار خوبی نیست.

گزینه‌ی ۴: اضافه شدن کلاس آب مشکل این معیار را بزرگ تر هم می کند. آب یک بردار صفر (نقطه‌ی (۰و۰)) است. معیار cosine را با بردار صفر نمی توان تعریف کرد. نمی توان زاویه‌ی یک بردار را با نقطه‌ی (۰و۰) محاسبه کرد. پس این گزینه اصلا قابل استفاده نیست.

نتیجه نهایی: پس همچنان معیار مناسب استفاده از فاصله‌ی اقلیدسی است.(گزینه ۱)

۵-۳. بخش سوم: طبقه بندی سه کلاس خودرو

معیار مناسب برای این طبقه بندی استفاده از فاصله‌ی همینگ است. در ادامه هر معیار را تحلیل می کنم.

گزینه‌ی ۱: داده های این مسئله عددی نیستند پس اصلا استفاده از فاصله‌ی اقلیدسی بی معنی است.

گزینه‌ی ۲: این معیار هم باید روی داده های عددی اعمال شود که داده های ما عددی نیستند. پس قابل استفاده نیست.

گزینه‌ی ۳: داده های لوگو، نوع خودرو و مدل خودرو همگی حالت دسته ای (category) دارند و عددی و پیوسته نیستند. فاصله‌ی همینگ دقیقا برای همین نوع داده هاست. هر بار که یک ویژگی بین دو نمونه متفاوت باشد، فاصله ۱ واحد اضافه می شود.

گزینه‌ی 4: فاصله‌ی cosine نیازمند بردار های عددی با اندازه و جهت است. لوگو و مدل و نوع خودرو مقادیر عددی نیستند. پس قابل استفاده نیست.

نتیجه نهایی: پس معیار همینگ بهترین و تنها گزینه مورد استفاده در اینجاست.(گزینه ۳)

۵-۴. بخش چهارم: استفاده از درخت تصمیم

گزینه سه جواب صحیح است. برای آشنایی با ID tree از [۷] استفاده کردم. در ادامه هر کدام از گزینه ها را بررسی می کنم.

گزینه ۱: توضیحات این گزینه کاملا غلط است. روش KNN اصلا نیازی به رسم داده ها ندارد. این روش صرفا نیازمند یک معیار فاصله سنجی است. حتی وقتی داده ها غیر عددی هستند معیار هایی مثل فاصله‌ی همینگ که در بخش قبل دیدیم قابل استفاده هستند.

گزینه ۲: این دلیل هم درست نیست. اگر چند نمونه دقیقا یکسان باشند همه‌ی آنها در یک فاصله قرار می گیرند. KNN بر اساس داده های همسایه کار می کند پس وقتی داده های تکراری در همسایگی یک نقطه داشته باشیم، احتمال اینکه کلاس غالب درست باشد افزایش می یابد. پس اگر تعداد آنها زیاد باشد باعث افزایش دقت می شود. پس وجود نمونه های تکراری اتفاقا به نفع ماست.

گزینه ۳: همان طور که از متن سوال هم مشخص است، سه کلاس با استفاده از لوگو به تنهایی قابل طبقه بندی هستند. KNN تمامی ویژگی ها را به یک اندازه مهم در نظر می گیرد. ویژگی های کم اهمیت تر مانند مدل و نوع خودرو می تواند فاصله ها را خراب کند. در منبع [۷] می بینیم که درخت تصمیم چند مرحله دارد

که دومین مرحله مربوط به انتخاب ویژگی است که با استفاده از Attribute Selection Measure (ASM) بهترین ویژگی در مجموعه داده انتخاب می شود. همین مرحله باعث می شود ویژگی های نامرتبط کنار گذاشته شوند و اولین split را روی لوگو در گره ریشه اعمال می کنیم.

بنابراین درخت می تواند خیلی راحت تر و دقیق تر تاکسی ها را فقط بر اساس لوگو جدا کند.

گزینه‌ی ۴: ویژگی ها ثابت هستند (لوگو، مدل و نوع خودرو) و به حرکت تاکسی ها وابسته نیستند. این گزینه به وضوح غلط است.

نتیجه نهایی: پس بر این اساس می بینیم که ID tree با توجه به اینکه ویژگی های نامرتبط را کنار می گذارد، پس می تواند از KNN بهتر عمل کند.(گزینه ۳)

پرسش ۶ – Kernel Ridge Regression

۶-۱. مینیمم تابع هزینه رگرسیون ریج

برای نشان دادن بردارها از حروف کوچک و bold استفاده می کنم. تابع هزینه‌ی رگرسیون ریج به صورت زیر است.

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 + \frac{\lambda}{2}\|\theta\|^2 \quad (40)$$

نمایش برداری آن به صورت زیر است.

$$J(\theta) = \frac{1}{2}\|\boldsymbol{x\theta} - \boldsymbol{y}\|^2 + \frac{\lambda}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} \quad (41)$$

ابتدا تابع هزینه را باز می کنیم و ریز تر می نویسیم.

$$J(\theta) = \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{x}^T\boldsymbol{x\theta} - \boldsymbol{y}^T\boldsymbol{x\theta} + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} + \frac{\lambda}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} \quad (42)$$

حالا از این رابطه نسبت به θ مشتق می گیریم و مساوی صفر قرار می دهیم.

$$\nabla_\theta J(\theta) = \boldsymbol{x}^T\boldsymbol{x\theta} - \boldsymbol{x}^T\boldsymbol{y} + \lambda\boldsymbol{\theta} = 0 \quad (43)$$

و

$$(\boldsymbol{x}^T\boldsymbol{x} + \lambda I)\boldsymbol{\theta} = \boldsymbol{x}^T\boldsymbol{y} \quad (44)$$

پس با استفاده از θ زیر تابع هزینه مینیمم می شود.

$$\boldsymbol{\theta} = (\boldsymbol{x}^T\boldsymbol{x} + \lambda I)^{-1}\boldsymbol{x}^T\boldsymbol{y} \quad (45)$$

۶-۲. استفاده از ترفند کرنل در رگرسیون ریج

با استفاده از یک تبدیل برای ویژگی ها، تابع هزینه به صورت زیر تبدیل شد.

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(\theta^T\phi\left(x^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\|\theta\|^2 \quad (46)$$

در این رابطه ماتریس ویژگی ها به صورت زیر است.

$$\Phi = \begin{bmatrix} \phi\left(x^{(1)}\right)^T \\ \phi\left(x^{(2)}\right)^T \\ \vdots \\ \phi\left(x^{(n)}\right)^T \end{bmatrix} \quad (47)$$

پس تابع هزینه را به این صورت می توان بازنویسی کرد.

$$J(\theta) = \frac{1}{2}\|\boldsymbol{\Phi\theta} - \boldsymbol{y}\|^2 + \frac{\lambda}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} \quad (48)$$

متناظرا از بخش اول می دانیم که با مساوی صفر قرار دادن گرادیان تابع هزینه به رابطه‌ی زیر خواهیم رسید.

$$\boldsymbol{\theta} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda I)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} \quad (49)$$

حال از رابطه‌ی معرفی شده در صورت سوال استفاده می کنیم. با فرض $A = \Phi$ و $B = \Phi^T$

$$(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda I)^{-1}\boldsymbol{\Phi}^T = \boldsymbol{\Phi}^T(\boldsymbol{\Phi\Phi}^T + \lambda I)^{-1} \quad (50)$$

پس می توانیم (۴۹) را به صورت زیر بازنویسی کنیم.

$$\boldsymbol{\theta} = \boldsymbol{\Phi}^T(\boldsymbol{\Phi\Phi}^T + \lambda I)^{-1}\boldsymbol{y} \quad (51)$$

از طرفی سوال گفته بردار $\boldsymbol{\theta}$ را می توان به صورت ترکیب خطی از بردار ویژگی ها در نظر گرفت.

$$\boldsymbol{\theta} = \sum_{i=1}^{m} \alpha_i \phi\left(x^{(i)}\right) = \Phi^T\alpha \quad (52)$$

با مقایسه ی (۵۱) و (۵۲) می توان دریافت:

$$\alpha = (\boldsymbol{\Phi\Phi}^T + \lambda I)^{-1}\boldsymbol{y} \quad (53)$$

از تعرفی ماتریس کرنل به صورت $K = \Phi\Phi^T$ تعریف شده و هر درایه آن به صورت زیر است.

$$K_{ij} = \phi\left(x^{(i)}\right)^T\phi\left(x^{(j)}\right) = k\left(x^{(i)}, x^{(j)}\right) \quad (54)$$

پس (۵۳) را می توان به صورت زیر بازنویسی کرد.

$$\alpha = (K + \lambda I)^{-1}\boldsymbol{y} \quad (55)$$

پس پیش بینی به ازای داده‌ی جدید x_{new} به صورت زیر است.

$$\hat{y}(x_{new}) = \boldsymbol{\theta}^T\phi(x_{new}) = (\Phi^T\alpha)^T\phi(x_{new}) = \alpha^T\Phi\phi(x_{new}) \quad (56)$$

که در این رابطه:

$$\Phi\phi(x_{new}) = \begin{bmatrix} \phi\left(x^{(1)}\right)^T \phi(x_{new}) \\ \vdots \\ \phi\left(x^{(m)}\right)^T \phi(x_{new}) \end{bmatrix} = \begin{bmatrix} k\left(x^{(1)}, x_{new}\right) \\ \vdots \\ k\left(x^{(m)}, x_{new}\right) \end{bmatrix} \quad (۵۷)$$

پس با توجه به (۵۷) می توان رابطه‌ی پیش بینی را به صورت زیر هم نوشت.

$$\hat{y}(x_{new}) = \sum_{i=1}^{m} \alpha_i k\left(x^{(i)}, x_{new}\right) \quad (۵۸)$$

پس به این ترتیب می توانیم در یک فضای ویژگی بزرگ (با ابعاد بالا) پیش بینی انجام دهیم بدون اینکه اصلا نیازی به محاسبه‌ی $\phi(x)$ داشته باشیم.

پرسش ۷ – Deriving Linear Regression

در این مسئله می خواهیم فرمول بهینه w^* که خطای مربعات متوسط را کمینه می کند پیدا کنیم و نشان دهیم که این فرمول به شکل نسبتی بین کوواریانس و واریانس است.

۷-۱. محاسبه‌ی $\Sigma_{yx} = E_{x,y}[yx]$

مدل ما در این مسئله به صورت زیر است.

$$y = w^T x + \epsilon \quad (۵۹)$$

امیدریاضی مورد نظر به صورت زیر است. توجه کنید که x مستقل از ϵ بوده و w اسکالر است.

$$\Sigma_{yx} = E_{x,y}[yx] = E_{x,y}[(w^T x + \epsilon)x] = E[w^T x^۲] + E[\epsilon x] \quad (۶۰)$$
$$= w^T E[x^۲] + E[\epsilon]E[x]$$

گفتیم w یک اسکالر است پس ترانهاده آن با خودش برابر است. با توجه به توزیع x و ϵ خواهیم داشت.

$$\Sigma_{yx} = w\Sigma_x + ۰ = w\Sigma_x \quad (۶۱)$$

۷-۲. محاسبه‌ی $\Sigma_{yx} = E_{x,y}[yx]$ با استفاده از نمونه ها

با فرض داده های $\{(x_n, y_n)\}_{n=۱}^N$ مقدار امیدریاضی به صورت زیر است.

$$E_{x,y}[yx] = \frac{۱}{N}\sum_{i=۱}^{N} x_n y_n \quad (۶۲)$$

۷-۳. محاسبه‌ی w^* بهینه

می خواهیم w را به گونه ای پیدا کنیم که امیدریاضی مربع خطا (residual loss) مینیمم شود. (توجه شود تمامی متغیر ها اسکالر است.)

$$A = E[(y - wx)^۲] = E[y^۲] - ۲wE[yx] + w^۲E[x^۲] \quad (۶۳)$$

حال از A نسبت به w مشتق گرفته و مساوی صفر قرار می دهیم.

$$\frac{dA}{dw} = -۲E[yx] + ۲wE[x^۲] = ۰ \quad (۶۴)$$

پس مقدار بهینه w به صورت زیر است.

$$w^* = \frac{E[yx]}{E[x^۲]} \quad (۶۵)$$

می دانیم کواریانس هر دو متغیر تصادفی X و Y به صورت زیر تعریف می شود.

$$\begin{aligned} cov(X,Y) = \Sigma_{yx} &= E[(X - E[X])(Y - E[Y])] \\ &= E\big[XY - XE[Y] - E[X]Y + E[X]E[Y]\big] \\ &= E[XY] - E[Y]E[x] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned} \qquad (٦٦)$$

پس خواهیم داشت:

$$E[XY] = \Sigma_{yx} + E[X]E[Y] = \Sigma_{yx} + \mu_x\mu_y \qquad (٦٧)$$

همچنین طبق تعریف کوواریانس برای $E[x^2]$ به طور مشابه خواهیم داشت:

$$E[x^2] = \Sigma_x + \mu_x^2 \qquad (٦٨)$$

حال با توجه به (۶۷) و (۶۸) می توانیم (۶۵) را به صورت زیر بازنویسی کنیم. توجه داشته باشید که میانگین ها صفر هستند.

$$w^* = \frac{E[yx]}{E[x^2]} = \frac{\Sigma_{yx} + \mu_x\mu_y}{\Sigma_x + \mu_x^2} = \frac{\Sigma_{yx}}{\Sigma_x} \qquad (٦٩)$$

این بخش با رابطه‌ی (۶۹) پایان می یابد. اما برای اطمینان از محدب بودن (۶۳) مشتق دوم را هم حساب می کنیم تا مطمئن باشیم نقطه‌ی بدست آمده مینیمم تابع است.

$$\frac{d^2A}{dw^2} = 2E[x^2] \qquad (٧٠)$$

از آنجا که $E[x^2] > 0$ پس مشتق دوم مثبت بوده و مقدار بدست آمده مینیمم است.

۷-۴. محاسبه‌ی w^* بهینه با استفاده از توزیع جدید

در این بخش باید با فرض توزیع $N(\mu_x, \Sigma_x)$ برای داده های x مقدار بهینه ی $\boldsymbol{w}^*$ را پیدا کنیم. این مورد را در بخش قبل در روابط (۶۷)، (۶۸) و (۶۹) دیدیم. پس با توجه به آن ها مقدار بهینه به صورت زیر خواهد بود.

$$w^* = \frac{E[yx]}{E[x^2]} = \frac{\Sigma_{yx} + \mu_x\mu_y}{\Sigma_x + \mu_x^2} \qquad (٧١)$$

۷-۵. اهمیت توزیع داده های x

در تمامی مراحل بخش ۳ و ۴ از گوسی بودن داده ها هیچ استفاده ای نکردیم. در تمامی روابط (۶۳) الی (۶۹) نهایتا به ممان های اول و دوم x و y نیاز داریم. یعنی صرفا به شرط داشتن مقادیر میانگین و کوواریانس، می توانیم به رابطه ی (۷۱) برسیم. در واقع اصلا هیچ فرضی در مورد ساختار و توزیع داده ها نیاز نداریم. تنها چیزی

که باید بر قرار باشد (علاوه بر داشتن ممان های اول و دوم) این است که $E[yx]$ و $E[x^{۲}]$ مقادیر محدود و $E[x^{۲}]$ مقدار غیر صفر داشته باشد تا تقسیم در (۷۱) معتبر باشد.

پرسش ۸ – Logistic Regression with Newton's Method

در این سوال تابع هزینه‌ی لجیستیک را در حالت ماتریسی بررسی می کنیم. قدم به قدم روابط را محاسبه می کنیم تا به فرمول آپدیت روش نیوتن برسیم.

توجه کنید همان طور که سوال هم اشاره کرده از حروف انگلیسی bold برای نشان دادن بردار استفاده کردم.

۸-۱. محاسبه‌ی $\nabla_w J(w)$

تابع هزینه به فرم برداری به صورت زیر نوشته می شود.

$$J(w) = -\boldsymbol{y}.\ln \boldsymbol{s} - (۱ - \boldsymbol{y}).\ln(۱ - \boldsymbol{s}) \quad (۷۲)$$

در این رابطه به شکل مستقیم از W استفاده نشده بلکه W درون S تعریف شده است. به همین دلیل برای محاسبه‌ی گرادیان نسبت به W باید از قاعده ی زنجیره ای استفاده کنیم.

بردار S و هر المان آن به صورت زیر تعریف می شود.

$$\boldsymbol{s} = \begin{bmatrix} s(x_۱^T w) \\ s(x_۲^T w) \\ \vdots \\ s(x_n^T w) \end{bmatrix} \quad , \quad s(y) \triangleq \frac{۱}{۱ + e^{-y}}$$

ابتدا مشتق هر المان از بردار **s** را مشخص می کنیم.

$$\frac{d}{dy} s(y) = \frac{e^{-y}}{(۱ + e^{-y})^۲} \quad (۷۳)$$

می توانیم با استفاده از خود $s(y)$ مشتق آن را به صورت زیر نشان دهیم.

$$\frac{d}{dy} s(y) = s(y)\big(۱ - s(y)\big) \quad (۷۴)$$

پس مشتق نسبت به W به صورت زیر تعریف می شود.

$$\left[\frac{\partial s_i}{\partial w}\right]^T = s_i(۱ - s_i)x_i^T \quad (۷۵)$$

و گرادیان به صورت زیر خواهد بود.

$$\nabla_w \boldsymbol{s} = \begin{bmatrix} s_۱(۱ - s_۱)x_۱^T \\ s_۲(۱ - s_۲)x_۲^T \\ \vdots \\ s_n(۱ - s_n)x_n^T \end{bmatrix} = diag(\boldsymbol{s} \odot (۱ - \boldsymbol{s}))X \quad (۷۶)$$

منظور از $\odot$ ضرب المان به المان (ضرب Hadamard) است.

حالا باید از J(w) نسبت به s گرادیان بگیریم. برای هر المان مشتق به صورت زیر است.

$$\frac{\partial J}{\partial s_i} = -\frac{y_i}{s_i} + \frac{1 - y_i}{1 - s_i} \tag{۷۷}$$

حالا همین مشتق های المانی را در قالب بردار می نویسیم و بردار بدست آمده را باز می کنیم تا بتوانیم بر حسب diag() بازنویسی کنیم. به صورت زیر:

$$\nabla_s J = \begin{bmatrix} \frac{\partial J}{\partial s_1} \\ \frac{\partial J}{\partial s_2} \\ \vdots \\ \frac{\partial J}{\partial s_n} \end{bmatrix} = \begin{bmatrix} -\frac{y_1}{s_1} + \frac{1 - y_1}{1 - s_1} \\ -\frac{y_2}{s_2} + \frac{1 - y_2}{1 - s_2} \\ \vdots \\ -\frac{y_n}{s_n} + \frac{1 - y_n}{1 - s_n} \end{bmatrix} = \begin{bmatrix} -\frac{y_1}{s_1} \\ -\frac{y_2}{s_2} \\ \vdots \\ -\frac{y_n}{s_n} \end{bmatrix} + \begin{bmatrix} \frac{1 - y_1}{1 - s_1} \\ \frac{1 - y_2}{1 - s_2} \\ \vdots \\ \frac{1 - y_n}{1 - s_n} \end{bmatrix} \tag{۷۸}$$

$$= -diag\left(\frac{1}{\boldsymbol{s}}\right)\boldsymbol{y} + diag\left(\frac{1}{1 - \boldsymbol{s}}\right)(1 - \boldsymbol{y})$$

$$= \frac{\boldsymbol{s} - \boldsymbol{y}}{\boldsymbol{s} \odot (1 - \boldsymbol{s})}$$

حال با استفاده از قاعده ی زنجیره ای و روابط (۷۶) و (۷۸) می توانیم گرادیان خواسته شده را محاسبه کنیم.

$$\nabla_w J(w) = (\nabla_w \boldsymbol{s})^T (\nabla_s J) \tag{۷۹}$$

$$= X^T diag\big(\boldsymbol{s} \odot (1 - \boldsymbol{s})\big) \frac{\boldsymbol{s} - \boldsymbol{y}}{\boldsymbol{s} \odot (1 - \boldsymbol{s})}$$

$$= X^T (\boldsymbol{s} - \boldsymbol{y})$$

به این شکل گرادیان تابع هزینه نسبت به w بده آمد.

۸-۲. محاسبه‌ی ماتریس هسین

مقدار $\nabla_w J$ و $\nabla_w \boldsymbol{s}$ را در بخش ۸-۱ حساب کردیم، با استفاده از آن می توانیم ماتریس هسین را هم بدست آوریم.

$$\nabla_w^2 J(w) = \frac{d}{dw}[X^T(\boldsymbol{s} - \boldsymbol{y})] = X^T \frac{ds}{dw} = X^T diag\big(\boldsymbol{s} \odot (1 - \boldsymbol{s})\big) X \tag{۸۰}$$

۸-۳. رابطه‌ی آپدیت روش نیوتن

رابطه‌ی آپدیت روش نیوتن برای مینیمم کردن $J(w)$ به صورت زیر است.

$$w_{new} = w_{old} - \left[\nabla_w^2 J(w_{old})\right]^{-1} \nabla_w J(w_{old}) \quad (۸۱)$$

گرادیان و هسین $J(w)$ در بخش های قبلی بدست آمد. صرفا آن ها را جاگذاری می کنیم.

$$w_{new} = w_{old} - \left[X^T diag\big(\boldsymbol{s} \odot (۱ - \boldsymbol{s})\big) X\right]^{-1} X^T(\boldsymbol{s} - \boldsymbol{y}) \quad (۸۲)$$

۸-۴. بررسی مثال عددی

داد ه ها به صورت زیر تعریف می شوند.

$$X = \begin{bmatrix} ۰.۲ & ۳.۱ & ۱ \\ ۱.۰ & ۳.۰ & ۱ \\ -۰.۲ & ۱.۲ & ۱ \\ 1.0 & 1.1 & 1 \end{bmatrix} \quad , \quad y = \begin{bmatrix} ۱ \\ ۱ \\ ۰ \\ 0 \end{bmatrix} \quad , \quad w^{(۰)} = \begin{bmatrix} -۱ \\ ۱ \\ ۰ \end{bmatrix} \quad (۸۳)$$

با استفاده از این مقادیر و نوشتن کد (کد نباید ارائه شود) optimization را تاجایی ادامه می دهیم که تغییرات w کمتر از $۱۰^{-۵}$ باشد. در این حالت به مقادیر زیر می رسیم.

$$w = \begin{bmatrix} ۵.۲۱۴۶ \\ ۷.۸۴۲۱ \\ -۱۸.۶۱۹۳ \end{bmatrix} \quad (۸٤)$$

پرسش ۹ – A Bayesian Interpretation of Lasso

این سوال در کل در مورد تخمین MAP می باشد.

۹-۱. بسط احتمال پسین با استفاده از رابطه‌ی بیز

با استفاده از قانون بیز می توانیم احتمال پسین را به صورت زیر بازنویسی کنیم. (در تساوی دوم از رابطه‌ی chain rule استفاده شده است.)

$$f\big(w\big|(x_i,y_i)\big)=\frac{f\big((x_i,y_i)\big|w\big)f(w)}{f(x_i,y_i)}=\frac{f(y_i|x_i,w)f(x_i|w)f(w)}{f(x_i,y_i)} \qquad (۸۵)$$

مخرج در این رابطه مستقل از w است پس کنار گذاشته می شود. از طرفی با فرض مستقل بودن x_i از w ها می توان نوشت $f(x_i|w)=f(x_i)$ و از طرفی این احتمال هم مستقل از w است پس کنار گذاشته می شود. پس بر این اساس این رابطه به صورت زیر تبدیل می شود.

$$f\big(w\big|(x_i,y_i)\big)\propto f(y_i|x_i,w)f(w) \qquad (۸۶)$$

حال باید هر احتمال را محاسبه کنیم. با *likelihood* شروع می کنیم.

$$f(y_i|x_i,w)=\frac{۱}{\sigma\sqrt{۲\pi}}\exp\left(-\frac{(y_i-w.x_i)^۲}{۲\sigma^۲}\right) \qquad (۸۹)$$

با توجه به استقلال داده ها، برای احتمال کل داده ها می توان نوشت:

$$f(y|x,w)=\prod_{i=1}^{n}f(y_i|x_i,w) \qquad (۹۰)$$

$$=\left(\frac{۱}{\sigma\sqrt{۲\pi}}\right)^۲\exp\left(-\frac{۱}{۲\sigma^۲}\sum_{i=۱}^{n}(y_i-w.x_i)^۲\right)$$

با توجه به استقلال تمامی المان های w از هم، احتمال پیشین هم به صورت زیر محاسبه می شود.

$$f(w_i)=\frac{۱}{۲b}\exp\left(-\frac{|w_i|}{b}\right) \qquad (۹۱)$$

$$f(w)=\prod_{i=۱}^{d}f(w_i)=\left(\frac{۱}{۲b}\right)^d\exp\left(-\frac{۱}{b}\sum_{i=۱}^{d}|w_i|\right) \qquad (۹۲)$$

حال با استفاده از (۹۰) و (۹۲) می توانیم (۸۶) را به صورت زیر بازنویسی کنیم.

$$f\left(w\middle|(x_i,y_i)\right) \propto \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w.x_i)^2\right)\left(\frac{1}{2b}\right)^d \exp\left(-\frac{1}{b}\sum_{i=1}^{d}|w_i|\right) \quad (93)$$

برای اینکه رابطه در یک خطر قرار گیرد مجبور شدم سایز فونت را کوچک تر کنم.

خواسته‌ی سوال با رابطه‌ی (۹۳) برآورده شد. اما همچنان در این رابطه ضرایبی وجود دارند که صرفا نقش نرمالیزاسیون ایفا می کنند و می توان آنها را هم کنار گذاشته و به رابطه‌ی زیر رسید.

$$f\left(w\middle|(x_i,y_i)\right) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w.x_i)^2 - \frac{1}{b}\sum_{i=1}^{d}|w_i|\right) \quad (94)$$

۹-۲. بررسی تخمین MAP

طبق صورت سوال log-likelihood را به صورت زیر تعریف می کنیم.

$$\ell(w) \triangleq \ln f\left(w\middle|(x_i,y_i)\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w.x_i)^2 - \frac{1}{b}\sum_{i=1}^{d}|w_i| \quad (95)$$

تخمینگر MAP با استفاده از log-likelihood به صورت زیر محاسبه می شود.

$$\widehat{w}_{MAP} = arg\max_{w}\ell(w) \quad (96)$$
$$= arg\max_{w}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w.x_i)^2 - \frac{1}{b}\sum_{i=1}^{d}|w_i|\right)$$

ماکزیمم کردن یک تابع، برابر با مینیمم کردن قرینه‌ی آن تابع است. اگر این تابع را در ۱- ضرب کنیم تابع قرینه می شود و حالا در این تابع باید به جای ماکزیمم به دنبال مینییمم باشیم.

$$\widehat{w}_{MAP} = arg\min_{w}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w.x_i)^2 + \frac{1}{b}\sum_{i=1}^{d}|w_i|\right) \quad (97)$$

ضرب کردن یک عدد مثبت ثابت در کل یک تابع هیچ تاثیری در نقطه‌ی ماکزیمم یا مینیمم نمی گذارد. پس می توانیم در $2\sigma^2$ ضرب کنیم و مطمئن باشیم که تخمینگر دقیقا مساوی قبل عمل می کند.

$$\widehat{w}_{MAP} = arg\min_{w}\left(\sum_{i=1}^{n}(y_i - w.x_i)^2 + \frac{2\sigma^2}{b}\sum_{i=1}^{d}|w_i|\right) \quad (98)$$

توجه کنید که $\sum_{i=1}^{d}|w_i| = \|w\|_1$.

به همین سادگی نشان دادیم مینیمم کردن (۹۸) معادل ماکزیمم کردن (۹۶) است. پارامتر λ هم به صورت زیر تعریف می شود.

$$\lambda = \frac{۲\sigma^{۲}}{b} \qquad (۹۹)$$

پرسش ۱۰– Parzen Window Classification with Gaussian Kernel

در این سوال یک مسئله‌ی طبقه بندی داریم که در آن توزیع شرطی داده ها به شرط کلاس ها را با استفاده روش پارزن محاسبه می کنیم. تابع کرنل در اینجا به صورت زیر است.

$$\varphi\left(\frac{x-x_i}{h}\right) \propto \exp\left[-\frac{(x-x_i)^T(x-x_i)}{۲h^۲}\right] \quad (۱۰۰)$$

این توزیع یک ضریب هم پشت exp در رابطه ی (۱۰۰) دارد ولی چون صرفا یک ضریب ثابت است که در تمامی کلاس ها تکرار می شود، می توان صرف نظر کرد.

هدف این سوال بررسی اثر مقدار پارامتر smoothing روی توزیع تخمین زده شده و طبقه بند است.

۱۰-۱. بخش a: طبقه‌بندی به ازای $h = 1$

مطابق آنچه در مورد تخمین توزیع با استفاده از روش پارزن در سوال ۱ دیدیم، توزیع شرطی به صورت زیر خوهد بود.

$$p_n(x|\omega_i) = \frac{۱}{n_i}\sum_{j=۱}^{n_i}\frac{۱}{h^d}\,\varphi\left(\frac{x-x_i}{h}\right) \quad (۱۰۱)$$
$$= \frac{۱}{n_i h^d}\,A\sum_{j=۱}^{n_i}\exp\left(-\frac{(x-x_i)^T(x-x_i)}{۲h^۲}\right)$$

در رابطه‌ی (۱۰۱) منظور از A ضریب نرمالیزاسیونی است که در (۱۰۰) با ذکر دلیل صرف نظر شده بود (حذف این ضریب ها ممکن است باعث شود مقادیر بدست آمده برای احتمال بیشتر از ۱ شود. از آنجا که ماکزیمم این احتمال ها مورد نظر ماست پس مشکلی نخواهد بود.) و d ابعاد مسئله است که در اینجا برابر با ۳ می باشد. با استفاده از قانون بیز و رابطه‌ی (۱۰۱) احتمال پسین را محاسبه کرده و کلاس متناظر با بیشترین مقدار را به عنوان کلاس برای داده‌ی تست، انتخاب می کنیم.

$$p(\omega_i|x) = \frac{p_n(x|\omega_i)p(\omega_i)}{\sum_j p_n(x|\omega_j)p(\omega_j)} \quad (۱۰۲)$$

صورت مسئله چیزی درباره ی احتمال های پیشین نگفته است. از طرفی اگر بخواهیم با استفاده از نمونه های هر کلاس برآوردی از احتمال پیشین داشته باشیم به علت کم بودن تعداد نمونه ها این برآورد دقیق نخواهد بود. پس فرض می شود احتمال های پیشین برابر هستند.

از طرفی مخرج در (۱۰۲) صرفا ضریبی است که برای همه ی کلاس ها تکرار می شود پس قابل صرف نظر است. پس تصمیم گیری به صورت زیر خواهد بود.

$$\omega_i = \arg\max_i p(\omega_i|x) = \arg\max_i p_n(x|\omega_i) \qquad (۱۰۳)$$

خروجی کد به ازای $h = ۱$ به صورت زیر است.

Test point ۱: [۰٫۵ ۱. ۰.]

ω۱: density = ۱٫۲۵۹۱۹۰e-۰۱

ω۲: density = ٤٫۷۱۰۷۸۵e-۰۱

ω۳: density = ۳٫۷۷٤٦٥٦e-۰۱

Predicted class: ω۲

Test point ۲: [۰٫۳۱ ۱٫۵۱ -۰٫۵]

ω۱: density = ۱٫۵۳٤۳۸٦e-۰۱

ω۲: density = ٤٫۸۲۷۹٦۳e-۰۱

ω۳: density = ۲٫۰۵۰٤۰۳e-۰۱

Predicted class: ω۲

Test point ۳: [-۰٫۳ ۰٫٤٤ -۰٫۱]

ω۱: density = ۱٫۳۹۹۰۸۱e-۰۱

ω۲: density = ۳٫۷۸۲۵٤۳e-۰۱

ω۳: density = ۱٫۷۷۰٦٦٤e-۰۱

Predicted class: ω۲

می بینیم که طبقه بند برای هر سه داده‌ی تست کلاس ۲ را مشخص کرده است. به ازای هر داده‌ی تست مقدار احتمال شرطی به شرط هر کلاس حساب شده و کلاس متناظر با بیشترین مقدار انتخاب شده است. در اینجا بیش از این توضیح نمی دهم زیرا تحلیل و نتیجه گیری بهتر در انتهای بخش بعدی خواهد بود.

۱۰-۲. بخش b: طبقه‌بندی به ازای $h = 1$

خروجی کد به ازای $h = ۰.۱$ به صورت زیر است.

Test point ۱: [۰٫۵ ۱. ۰.]

ω۱: density = ۸٫۷۷٤٦۹۵e-۲۰

ω۲: density = ٦٫۷٦۹۰۲۰e-۰۵

ω۳: density = ۸٫۰۰۰۱۸۹e-۱۷

Predicted class: ω۲

Test point ۲: [۰,۳۱ ۱,۵۱ -۰,۵]

ω۱: density = ۲,۸۷۱۱۱٦e-۲۰

ω۲: density = ۱,۲۰۰۱۸٤e-۰۵

ω۳: density = ۲,۱۵۹۳۲۹e-۲۵

Predicted class: ω۲

Test point ۳: [-۰,۳ ۰,٤٤ -۰,۱]

ω۱: density = ۳,٦۳۷۱۲۳e-۱۲

ω۲: density = ۳,۷۸۳۰۹٤e-۰٤

ω۳: density = ۱,۰٦٤۰۲٦e-۳۹

Predicted class: ω۲

باز هم می بینیم که طبقه بند در هر سه نقطه‌ی تست کلاس دوم را انتخاب کرده است.

در بخش a که $h = ۱$ است توزیع های گوسی تخمین زده شده پهن تر هستند. یعنی توزیع ها نرم (smooth) و کم نوسان هستند. در این حالت چون دامنه ی گوسی ها بزرگ تر است، مقدار احتمال برای داده های تست از مرتبه‌ی ۰.۱ هستند.

اما در بخش b که $h = ۰.۱$ است توزیع های گوسی تخمین زده شده تیز تر و باریک تر می شوند. به این معنی که مقدار توزیع در دامنه ها بسیار کمتر خواهد بود. در واقع مقدار چگالی در اکثر نقاط به طور واضحی کوچک می شود. همین باعث می شود که مقدار احتمال برای داده های تست خیلی کمتر باشد.

این رفتار دو پیامد دارد. با کاهش h به نظر می رسد که مدل کلاس غالب را با اطمینان بیشتری تشخیص می دهد اما این یک مزیت نیست (پیامد اول). این تمایز بیش از حد ناشی از حساسیت شدید مدل به نمونه های آموزشی است و عملا منجر به overfitting کاهش توانایی generalization می شود(پیامد دوم).

پرسش ۱۱ – Maximum likelihood estimation

۱۱-۱. بخش a: محاسبه‌ی عددی MLE و مقایسه با حل فرم بسته

برای داده های نرمال likelihood به صورت زیر است.

$$L(\mu,\sigma^{۲})=\prod_{i=۱}^{n}\frac{۱}{\sqrt{۲\pi\sigma^{۲}}}\exp\left(-\frac{(x_i-\mu)^{۲}}{۲\sigma^{۲}}\right) \quad (۱۰٤)$$

لگاریتم آن به صورت زیر است.

$$\ell(\mu,\sigma^{۲})=-\frac{n}{۲}\log(۲\pi)-\frac{n}{۲}\log(\sigma^{۲})-\frac{۱}{۲\sigma^{۲}}\sum_{i=۱}^{n}(x_i-\mu)^{۲} \quad (۱۰٥)$$

با مشتق گرفتن نسبت به $\sigma^{۲}$ و μ و مساوی صفر قرار دادن می توان تخمین MLE آن ها را به دست آورد.

$$\frac{\partial\ell}{\partial\mu}=\frac{۱}{\sigma^{۲}}\sum_{i=۱}^{n}(x_i-\mu)=۰ \quad \Longrightarrow \quad \hat{\mu}_{MLE}=\bar{X}=\frac{۱}{n}\sum_{i=۱}^{n}x_i \quad (۱۰٦)$$

$$\frac{\partial\ell}{\partial\sigma^{۲}}=-\frac{n}{۲\sigma^{۲}}+\frac{۱}{۲(\sigma^{۲})^{۲}}\sum_{i=۱}^{n}(x_i-\mu)^{۲}=۰ \quad \Rightarrow \quad \widehat{\sigma^{۲}}=\frac{۱}{n}\sum_{i=۱}^{n}(x_i-\bar{X})^{۲} \quad (۱۰۷)$$

برای پیاده سازی این بخش از [۸] و [۹] استفاده کردم. برای مینیمم کردن تابع log likelihood باید از یک نقطه شروع می کردیم من از (۰و۰) استفاده کردم. مقادیر عددی و تئوری به صورت زیر هستند.

Numerical MLE:

mu_hat numerical=۱۰۰,٤۸۰٥۱۳٦۸۳۱٤٤۳٥ , sigma۲_hat numerical=۲۱٦,۳۱۸٦۷۰۳٦٥۹۱۲٤٥

Closed-form MLE:

mu_hat closed form=۱۰۰,٤۸۰٤۹۹۹۹۹۹۹۹۹ , sigma۲_hat closed form=۲۱٦,۳۱۸٤٦۹۷٥

می بینیم که مقادیر حل عددی با تئوری تا ۴ رقم اعشار کاملا مشابه هستند.

۱۱-۲. بخش b: بررسی مینیمم شدن منفی log likelihood

این مورد را هم از دید تئوری و هم عددی می توان بررسی کرد. ابتدا با بررسی تئوری شروع می کنم. برای بررسی مینیمم بودن منفی log likelihood در نقاط بدست آمده در تحلیل تئوری می توان مشتق دوم این عبارت

را نسبت به هر کدام از متغیر ها محاسبه کرد. اگر این مشتق دوم ها مثبت باشند یعنی نقاط بدست آمده همان مینیمم ها هستند.

$$\frac{\partial^2(-\ell)}{\partial\mu^2} = \frac{n}{\sigma^2} > 0 \qquad (۱۰۸)$$

پارامتر های n و σ^2 به ذاته مثبت هستند. پس مشتق دوم مثبت است و نقطه‌ی بدست آمده مینیمم است.

$$\frac{\partial^2(-\ell)}{\partial(\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}(x_i - \mu)^2 > 0 \qquad (۱۰۹)$$

پارامتر های این مشتق دوم هم یا توان دو دارند یا به ذاته مثبت هستند پس نقطه بدست آمده مینیمم است.

برای اینکه به صورت عددی نشان دهیم، مقادیر بدست آمده مینیمم هستند ابتدا مقدار تابع negative log likelihood را در نقطه‌ی بدست آمده از بخش قبل حساب می کنیم و سپس مقدار تابع را به ازای افزایش یا کاهش به اندازه ی خیلی کم (epsilon) نقطه حساب می کنیم. اگر نقاط بدست آمد در بخش قبل واقعا مینیمم تابع باشند باید مقدار تابع در نقاط کمتر از بقیه‌ی نقاط باشد.

NLL at MLE: ۸۲۱,٤٦۲۸۷۸٥۲۲۸۱۹۹

NLL mu + eps: ۸۲۱,٤٦۲۹۲٤۷٥۰۹٥۷٥

NLL mu - eps: ۸۲۱,٤٦۲۹۲٤۷٥۰۹٥۷٥

NLL sigma۲ + eps: ۸۲۱,٤٦۳۱٤٤۸۳۱۷۸۲٦

NLL sigma۲ - eps: ۸۲۱,٤٦۳۱٤٦٤۷۸۳۱٦۷

خوب می بینیم که به ازای تغییر اندک، مقدار تابع بیشتر از مقدار تابع در نقطه‌ی مینیمم بدست آمده شده. پس به این شکل نشان دادیم نقطه‌ی بدست آمده مینیمم تابع است.

۱۱-۳. بخش c: محاسبه‌ی ماتریس کوواریانس تخمینگر در حالت حدی

در واقع این بخش به زبان ساده می گوید اگر ما بخواهیم آزمایش را بار ها و بار ها تکرار کنیم، تخمین های $\hat{\mu}$ و $\widehat{\sigma^2}$ چگونه تغییر می کنند.

برای محاسبه‌ی این بخش نیاز به استفاده از ماتریس اطلاعات فیشر داریم. من با ماتریس اطلاعات فیشر برای محاسبه یک باند برای تخمینگر در درس تئوری آشکارسازی و تخمین آشنا شدم. در مورد ماتریس اطلاعات فیشر می توانید در [۱۰] بخوانید.

برای تعداد زیادی نمونه، توزیع به صورت زیر خواهد شد.

$$\hat{\theta} = \left(\hat{\mu}, \widehat{\sigma^2}\right) \sim N(\theta, I(\theta)^{-1}) \qquad (۱۱۰)$$

که در این رابطه، $I(\theta)$ همان ماتریس اطلاعات فیشر است. برای یک نمونه $x_i \sim N(\mu, \sigma^2)$ ماتریس اطلاعات فیشر به صورت زیر است.

$$I_1(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \quad (111)$$

برای n تا نمونه خواهیم داشت.

$$I_n(\mu, \sigma^2) = n.I_1(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \quad (112)$$

بر این اساس ماتریس کوواریانس MLE برابر معکوس ماتریس اطلاعات فیشر بوده و به صورت زیر بدست می آید.

$$asymtotic\ covriance\ matrix = \Sigma = I_n^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \quad (113)$$

با استفاده از کد برای این دیتاست به صورت زیر بدست می آید.

```
array([[ ۱,۰۸۱۵۹۲۳۵,  ۰.        ],
       [ ۰.        , ٤٦٧,٩٣٦٨٠٣٥٥]])
```

۱۱-۴. بخش d: بهبود ماتریس کوواریانس تخمینگر در حالت حدی

در بخش قبل ماتریس کوواریانس را برای تخمینگر MLE حساب کردیم ولی یک مشکل وجود داشت و آن هم این است که σ^2 بایاس دارد یعنی می توان بعد استفاده از یک ضریب، بایاس آن را از بین برد و این ماتریس کوواریانس را بهبود بخشید.

$$E[\hat{\sigma}^2] = \frac{1}{n} E\left[\sum_{i=1}^{n} (x_i - \bar{X})^2\right] \quad (114)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} E[(x_i - \mu)^2] - nE[(\bar{X} - \mu)^2]\right)$$

$$= \frac{1}{n}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \frac{1}{n}(n\sigma^2 - \sigma^2) = \frac{n-1}{n}\sigma^2$$

حال می توانیم با تعریف متغیر $s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$ بایاس را صفر کنیم. حالا ماتریس کوواریانس در حالت حدی (asymptotic) به صورت زیر می شود.

$$asymtotic\ covriance\ matrix = \Sigma = \begin{pmatrix} \frac{s^2}{n} & 0 \\ 0 & \frac{2s^4}{n} \end{pmatrix} \quad (115)$$

با استفاده از کد برای این دیتاست به صورت زیر بدست می آید.

Better estimated asymptotic covariance matrix:

[[۱,۰۸۷۰۲۷۴۹ ۰.]

[۰. ۴۷۲,۶۵۱۵۰۲۲۹]]

۱۱-۵. بخش e: فاصله اطمینان ٪۹۵

با استفاده از بخش c می توان گفت $\hat{\sigma}^2 \xrightarrow{d} N\left(\sigma^2, \frac{2\sigma^4}{n}\right)$ و خطای استاندارد (standard error-SE) به صورت زیر است.

$$\widehat{SE}(\hat{\sigma}^2) = \sqrt{\frac{2\hat{\sigma}^4}{n}} \quad (116)$$

برای داشتن فاصله اطمینان ٪۹۵ باید داشته باشیم $\hat{\sigma}^2 \pm 1.96\sqrt{\frac{2\hat{\sigma}^4}{n}}$.

حال با استفاده از کد می توان این مقادیر را محاسبه کرد.

۹۵٪ large-sample CI for sigma^۲: (np.float۶۴(۱۷۳.۹۲۰۰۴۹۶۷۹),

np.float۶۴(۲۵۸,۷۱۶۸۸۹۸۲۱))

۱۱-۶. بخش f: تست فرض

در این بخش یک مسئله آشکار سازی داریم که با استفاده از روش های مختلف تست فرض آن را حل می کنیم.

$$H_0: \mu = 103 \ , \ H_1: \mu \neq 103 \quad (117)$$

Z-test .۱-۶-۱۱

برای آشنایی با این روش تست فرض، از [۱۱] استفاده کردم. در کل این روش وقتی استفاده می شود که توزیع آماره‌ی تحت فرضیه‌ی صفر می تواند به صورت یک توزیع نرمال تخمین زده شود. یعنی تحت فرضیه‌ی صفر داریم: $Z \xrightarrow{d} N(۰,۱)$

مقدار Z-statistic به صورت زیر محاسبه می شود. $(\mu_{.} = ۱۰۳)$

$$Z = \frac{\hat{\mu} - \mu_{.}}{\sqrt{\frac{\hat{\sigma}^{۲}}{n}}} \quad (۱۱۸)$$

با استفاده از کد، مقادیر p-value و Z-statistic به صورت هستند.

Z-test statistic: -۲,٤۲۲٦۰٤۷۲۲٦۲۷۲۱۰۷

Z-test p-value: ۰,۰۱٥٤۰۹٦۸۳٥۹۱٦٥٤٤

مقدار بزرگ Z-statistic و مقدار اندک p-value، فرضیه‌ی صفر را رد می کنند.

Likelihood ratio chi-squared test .۲-۶-۱۱

برای آشنایی با این روش تست فرض از [۱۲] استفاده کردم. این روش وقتی مورد استفاده قرار میگیرد که بین درستی یک فرضیه و متمم آن تصمیم می گیریم. آماره‌ی آن به این صورت محاسبه می شود.

$$LR = ۲\left(\ell(\hat{\mu}, \hat{\sigma}^{۲}) - \ell(\mu_{.}, \tilde{\sigma}^{۲})\right) \quad (۱۱۹)$$

که $(\hat{\mu}, \hat{\sigma}^{۲})$ بدون هیچ محدودیتی و برای تمامی تخمین های MLE و $(\mu_{.}, \tilde{\sigma}^{۲})$ برای تخمینگر های MLE محدود به فرضیه‌ی صفر است.

$$\tilde{\sigma}^{۲} = \frac{۱}{n}\sum_{i=۱}^{n}(x_i - \mu_{.})^{۲} \quad (۱۲۰)$$

به صورت حدی خواهیم داشت: $LR \xrightarrow{d} \chi^{۲}(۱)$ و حل فرم بسته آن به صورت زیر است.

$$LR = n \log\left(\frac{\tilde{\sigma}^{۲}}{\hat{\sigma}^{۲}}\right) \quad (۱۲۱)$$

با استفاده از کد مقادیر به صورت زیر محاسبه شدند.

LR statistic: ٥,۷۸٤٥٤۸۷۷۸٥۹٦٥۷۲

LR test p-value: ۰,۰۱٦۱٦۷٦٤۷٥۱٦۰۷۷٥۷۷

۱۱-۶-۳. Wald chi-squared test

برای آشنایی با این روش تست فرض از [۱۳] استفاده کردم. این تست فرض برای وقتی است که می خواهیم بدانیم پارامتر تخمین زده شده توسط نمونه برابر با پارامتر مورد نظر است یا خیر. آماره‌ی آن به صورت زیر تعریف می شود.

$$W = \frac{(\hat{\mu} - \mu_.)^2}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \qquad (۱۲۲)$$

و به صورت حدی داریم: $W \xrightarrow{d} \chi^2(1)$

با توجه به Z-test به راحتی می توان دریافت که $W = Z^2$. با استفاده از کد مقادیر به صورت زیر محاسبه شدند.

Wald statistic: ۵٫۸۶۹۰۱۳۶۴۲۰۹۵۶۶۳

Wald test p-value: ۰٫۰۱۵۴۰۹۶۸۳۵۹۱۶۵۴۲۸۸

با توجه به هر سه تست فرض می توان گفت، برای تعداد زیادی نمونه، هر سه تست فرض p-value های مشابهی می دهند ولی هر کدام از رابطه و مفهوم مختص خودشان می آیند.

۱۱-۷. بخش g: coefficient of variation

پارامتر coefficient of variation به صورت زیر محاسبه می شود.

$$CV = \frac{\sigma}{\mu} \qquad (۱۲۳)$$

۱۱-۷-۱. خاصیت invariance نسبت به یک ضریب ثابت مثبت

اگر X یک متغیر تصادفی با میانگین μ و واریانس σ^2 باشد. آنگاه متغیر تصادفی $Y = cX$ دارای میانگین $c\mu$ و انحراف معیار $c\sigma$ است.

بر این اساس coefficient of variation برای متغیر Y به صورت زیر است.

$$CV(Y) = \frac{c\sigma}{c\mu} = \frac{\sigma}{\mu} = CV(X) \qquad (۱۲۴)$$

۱۱-۷-۲. تخمین coefficient of variation

خاصیت invariance می گوید؛ اگر $\hat{\theta}$ تخمین MLE برای θ باشد آنگاه $g(\hat{\theta})$ تخمینگر MLE برای $g(\theta)$ است. پس بر این اساس اگر $\hat{\mu}$ تخمین MLE برای μ و $\hat{\sigma}$ تخمین MLE برای σ باشد، انگاه تخمین MLE برای coefficient of variation به صورت زیر است.

$$\widehat{CV} = \frac{\hat{\sigma}}{\hat{\mu}} \qquad (۱۲۵)$$

با استفاده از کد می توان این مقدار را برای داده های مشخص شده محاسبه کرد.

MLE of coefficient of variation: ۰,۱٤٦۳۷٤۳٦۱٤٦۲٦۳٤۷٤

۱۱-۷-۳. ۹۵٪ Confidence Interval via the Delta Method

بردار پارامتر $\theta = \begin{pmatrix} \mu \\ \sigma^{۲} \end{pmatrix}$ و تابع $g(\theta) = \frac{\sqrt{\sigma^{۲}}}{\mu}$ است. مشتق این تابع نسبت به μ و $\sigma^{۲}$ به صورت زیر است.

$$\frac{\partial g}{\partial \mu} = -\frac{\sqrt{\sigma^{۲}}}{\mu^{۲}} \quad , \quad \frac{\partial g}{\partial \sigma^{۲}} = \frac{۱}{۲\mu\sqrt{\sigma^{۲}}}$$

پس گرادیان تابع به صورت زیر است.

$$\nabla_{\theta} g(\theta) = \begin{pmatrix} -\frac{\sqrt{\sigma^{۲}}}{\mu^{۲}} \\ \frac{۱}{۲\mu\sqrt{\sigma^{۲}}} \end{pmatrix} \qquad (۱۲٦)$$

در بخش های قبلی ماتریس کوواریانس حالت حدی را حساب کردیم و با Σ نمایش دادیم. با استفاده از روش دلتا داریم:

$$var(\widehat{CV}) = \nabla_{\theta} g(\theta)^{T} \, \Sigma \, \nabla_{\theta} g(\theta) = \frac{۱}{n}\left(\frac{\sigma^{۴}}{\mu^{۴}} + \frac{\sigma^{۲}}{۲\mu^{۲}}\right) \qquad (۱۲۷)$$

با استفاده از این مقادیر، مقدار خطای استاندارد (standard error-SE) به صورت زیر است.

$$\widehat{SE}(\widehat{CV}) = \sqrt{\frac{۱}{n}\left(\frac{\hat{\sigma}^{۴}}{\hat{\mu}^{۴}} + \frac{\hat{\sigma}^{۲}}{۲\hat{\mu}^{۲}}\right)} \qquad (۱۲۸)$$

حال فاصله اطمینان ٪۹۵ به صورت زیر بدست می آید.

$$\widehat{CV} \pm ۱.۹۶ \times \widehat{SE}(\widehat{CV}) \qquad (۱۲۹)$$

با استفاده از کد و داده ها می توان به صورت زیر محاسبه کرد.

٩٥٪ delta-method CI for CV: (np.float٦٤(٠.١٣١٧٢٥٥٥٦٣٥٩٨٠٩٨),
np.float٦٤(٠,١٦١٠٢٣١٦٦٥٦٥٤٥٩٧))

پرسش ۱۲ – Non-Linear Regression using basis functions

این سوال در مورد پیاده سازی رگرسیون غیر خطی با استفاده از توابع پایه است که با استفاده از یک تبدیل هر داده را به یک بردار تبدیل می کند. تمامی مراحل مطابق آنچه سوال خواسته پیش می برم و در اینجا مستقیما نتایج را بررسی خواهیم کرد.

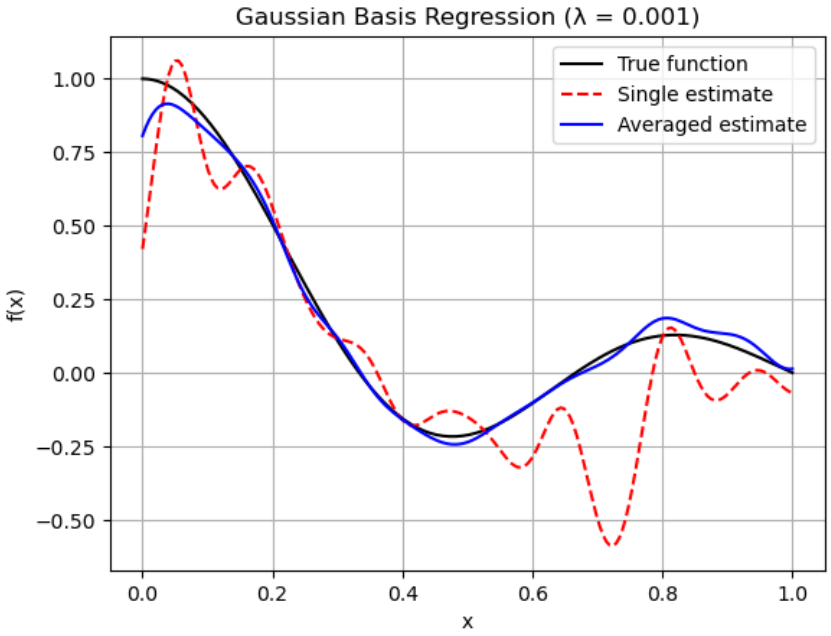
۱۲-۳. نمودار های gaussian basis regression به ازای لاندار های مختلف

در این بخش به ازای لاندار های مختلف تابع f(x) رسم شده و در ادامه تحلیل می شود.

۱۲-۳-۱. به ازای $\lambda = 0.001$:

در این حالت مدل تقریبا بدون regularization است. ضرایب β آزادی زیادی دارند. در تخمینی که به ازای یک مجموعه داده زدیم (نمودار خط چین قرمز) می بینیم که مدل بسیار نوسان دار است و به نویز حساس است. و در بعضی نواحی به شدت از تابع واقعی فاصله می گیرد. علت این اتفاق این است که مدل دچار overfitting شده است. تخمین میانگین گیری شده (نمودار آبی رنگ) نسبت به تخمین تک بار (یک مجموعه داده) صاف تر است. عملا نمودار خیلی بهتر شده و به تابع سینک اصلی بسیار نزدیک تر شده.

در کل در این حالت مدل وازیانس بالایی دارد . راه حل ارائه شده عملکرد چندان راضی کننده ای ندار

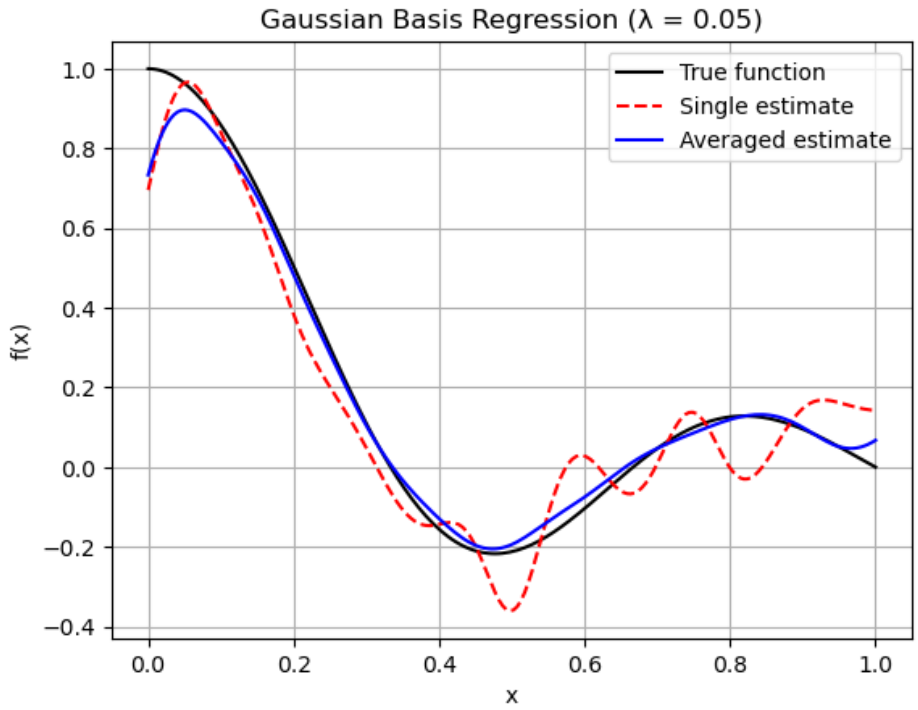


شکل ۳. نمودار رگرسیون به ازای $\lambda = 0.001$

۱۲-۳-۲. به ازای $\lambda = 0.05$:

در این حالت مدل تعادل خوبی بین انعطاف پذیری و محدود سازی ضرایب ایجاد کرده. در تخمینی که به ازای یک مجموعه داده زدیم (نمودار خط چین قرمز) می بینیم که بسیار به نمودار واقعی نزدیک تر و نوسان آن کم تر شده (هر چند همچنان نوسان را می توان به وضوح دید). تخمین میانگین گیری شده (نمودار آبی رنگ) نسبت به تخمین یک مجموعه خیلی بهتر شده. تقریبا بر تابع سینک منطبق شده و صاف (smooth) شده است.

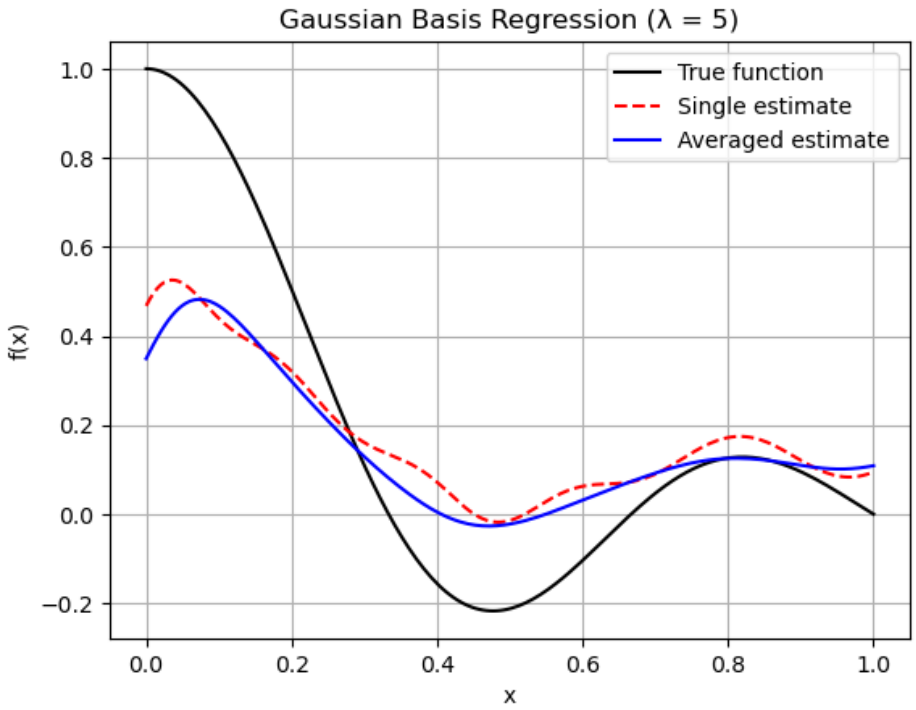
بر حسب این نمودار می بینیم که $\lambda = ۰.۰۵$ بهترین انتخاب است. (مقدار $\lambda = ۵$ را هم در ادامه خواهیم دید که وضع آن بدتر است.)



شکل ۴. نمودار رگرسیون به ازای $\lambda = 0.05$

۱۲-۳-۳. به ازای $\lambda = 5$:

در این حالت ضرایب به شدت به صفر میل داده می شوند. یعنی مدل بیش از حد ساده می شود. در تخمین به ازای یک مجموعه داده (نمودار خط چین قرمز) می بینیم که دامنه تابع به شدت کاهش یافته و مدل نمی تواند ساختار تابع سینک را بیاموزد. در این حالت مدل دچار underfitting شده است. بایاس بالا دارد و کیفیت تخمین پایین است. به حدی که حتی تخمین میانگین گیری شده هم که در نمودار های قبل خوب عمل می کرد، نمی تواند به خوبی عمل کند.



شکل ۵. نمودار رگرسیون به ازای $\lambda = 5$

پس همان طور که گفته شد $\lambda = ۰.۰۵$ بهترین انتخاب است. علت اینکه تابع میانگین گیری شده عملکرد بهتری از خود نشان می دهد این است که واریانس مدل در ۱۰۰ بار تکرار و جمع شده و با میانگین گیری، واریانس آن بسیار کاهش می یابد. به همین دلیل نمودار تابع میانگین گیری شده بسیار به تابع واقعی نزدیک می شود.

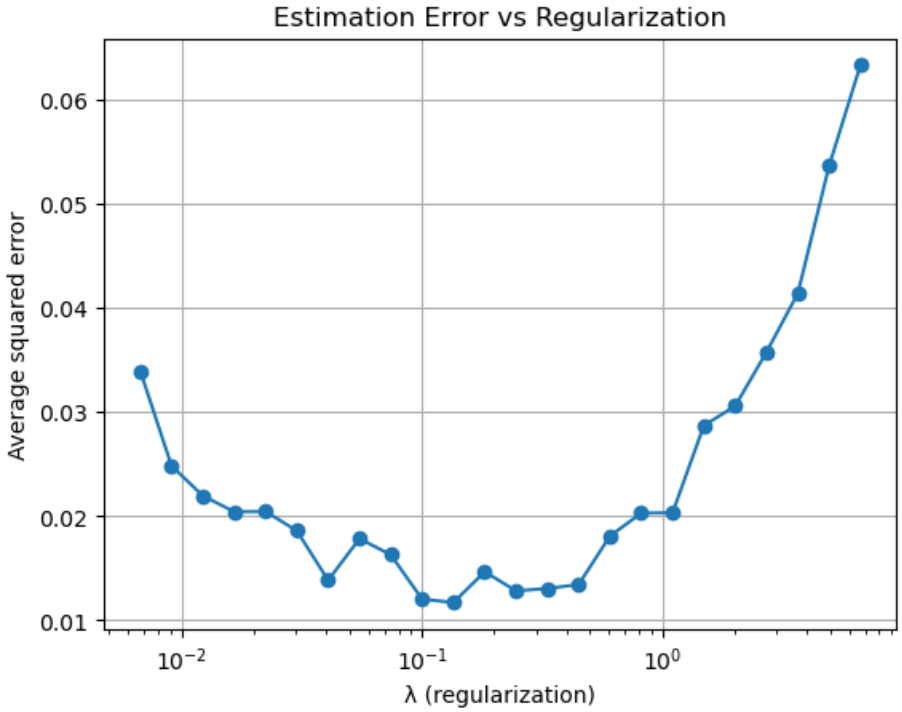
۱۲-۴. نمودار میانگین مربع خطاها بر حسب λ

در این بخش خواسته شده که دو سمت نمودار میانگین مربع خطا را تحلیل کنیم.

سمت چپ نمودار که مقادیر لاندا بسیار کم است در واقع regularization ناچیز است. مدل ها همگی به نویز حساس هستند و ضرایب β بزرگ و ناپایدار می شوند. مدل ها به ازای این لاندا های کوچک واریانس زیادی داشته و دچار overfitting شده اند.

سمت راست نمودار که مقادیر لاندا بسیار بزرگ هستند، شاهد regularization زیادی هستیم. ضرایب به شدت به سمت صفر میل داده می شوند. مدل بیش از حد ساده می شود و خطا افزایش می یابد. در این حالت مدل بایاس زیادی پیدا می کند و دچار underfitting می شود.

در مورد نقطه‌ی میانی این نمودار هم می توان گفت که بهترین trade-off را بین بایاس و واریانس مدل برقرار کرده.



شکل ۶. نمودار میانگین مربع خطا به ازای مقادیر مختلف λ

مراجع

[۱] "axvline," [Online]. Available: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.axvline.html.

[۲] "wikipedia, Condition number," [درون خطی]. Available: https://en.wikipedia.org/wiki/Condition_number.

[۳] "wikipedia, Euclidean distance," [درون خطی]. Available: https://en.wikipedia.org/wiki/Euclidean_distance.

[۴] "wikipedia, Taxicab geometry," [درون خطی]. Available: https://en.wikipedia.org/wiki/Taxicab_geometry.

[۵] "wikipedia, Hamming distance," [درون خطی]. Available: https://en.wikipedia.org/wiki/Hamming_distance.

[۶] "wikipedia, Cosine similarity," [درون خطی]. Available: https://en.wikipedia.org/wiki/Cosine_similarity.

[۷] "درخت تصمیم در یادگیری ماشین – به زبان ساده + نمونه کد," Faradars, [درون خطی]. Available: https://blog.faradars.org/%D۸٪AF%D۸٪B۱٪D۸٪AE%D۸٪AA-%D۸٪□□٪□۸٪□۵٪□۹٪۸۵٪□□٪۸□٪□۹٪۸۵-٪□۸٪□□٪□۸٪□۱-%DB%۸C%D۸٪A۷٪D۸٪AF%DA%AF%DB%۸C%D۸٪B۱٪DB%۸C-%D۹٪۸۵٪D۸٪A۷٪D۸٪B۴٪DB٪۸C٪D۹٪۸۶/#٪d۸٪af٪d۸٪b۱٪d۸٪ae٪d۸٪aa-%d۸٪aa٪d۸٪b۵٪d۹٪۸۵٪db٪۸c٪d۹٪۸۵-٪d۸٪af٪d۸٪b۱-٪db٪۸c٪d۸٪a۷٪d۸٪.

[۸] "SciPy, minimize," [درون خطی]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html.

[۹] "youtube, Intro to Scipy Optimization: Minimize Method," [درون خطی]. Available: https://www.youtube.com/watch?v=G۰yP_TM-oag.

[۱۰] "wikipedia, Fisher information," [درون خطی]. Available: https://en.wikipedia.org/wiki/Fisher_information.

[۱۱] "Z-test," wikipedia, [درون خطی]. Available: https://en.wikipedia.org/wiki/Z-test.

[۱۲] "Likelihood-ratio test," wikipedia, [درون خطی]. Available: https://en.wikipedia.org/wiki/Likelihood-ratio_test.

[۱۳] "Wald test," wikipedia, [درون خطی]. Available: https://en.wikipedia.org/wiki/Wald_test.