بسمه تعالی

پاسخ سری سوم تمرینات درس یادگیری ماشین

امیرحسین رمضانی بناب (۹۹۲۱۰۲۹۴)

۱ سوال اول

۱.۱ آ

فرض کنیم دادگان آموزش ما به شکل زیر باشند:

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

اگر روی این دادگان E_{out} را به شکل صریح محاسبه کنیم داریم:

$$E_{out}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}\big[(g^{(\mathcal{D})}(x) - f(x))^2\big]$$

حال اگر روی تمام $\mathcal{D}$های موجود مقدار متوسط بگیریم به عبارت زیر خواهیم رسید:

$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] = \mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_{\mathbf{x}}\big[(g^{(\mathcal{D})}(x) - f(x))^2\big]\Big] \quad (۱)$$

$$= \mathbb{E}_{\mathbf{x}}\Big[\mathbb{E}_{\mathcal{D}}\big[\underbrace{(g^{(\mathcal{D})}(x) - f(x))^2}_{\text{error measure}}\big]\Big] \quad (۲)$$

که به دلیل مثبت و کراندار بودن عبارت داخل براکت توانستیم از ۱ به ۲ برسیم. حال تعریف می‌کنیم:

$$\bar{g}(x) = \mathbb{E}\big[g^{(\mathcal{D})}(x)\big] \quad (۳)$$

حال روی عبارت $\mathbb{E}_{\mathcal{D}}\big[(g^{(\mathcal{D})}(x) - f(x))^2\big]$ متمرکز می‌شویم. داریم:

$$\mathbb{E}_{\mathcal{D}}\big[(g^{(\mathcal{D})}(x) - f(x))^2\big] \quad (۴)$$

$$= \mathbb{E}_{\mathcal{D}}\big[(g^{(\mathcal{D})}(x) - \bar{g}(x) + \bar{g}(x) - f(x))^2\big] \quad (۵)$$

$$= \mathbb{E}_{\mathcal{D}}\big[(g^{(\mathcal{D})}(x) - \bar{g}(x))^2 + (\bar{g}(x) - f(x))^2 + 2(g^{(\mathcal{D})}(x) - \bar{g}(x))(\bar{g}(x) - f(x))\big] \quad (۶)$$

$$= \mathbb{E}_{\mathcal{D}}\big[(g^{(\mathcal{D})}(x) - \bar{g}(x))^2\big] + \mathbb{E}_{\mathcal{D}}\big[(\bar{g}(x) - f(x))^2\big] + \mathbb{E}_{\mathcal{D}}\big[2(g^{(\mathcal{D})}(x) - \bar{g}(x))(\bar{g}(x) - f(x))\big] \quad (۷)$$

حال داریم :

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\big[2(g^{(\mathcal{D})}(x) - \bar{g}(x))(\bar{g}(x) - f(x))\big] &= 2(\bar{g}(x) - f(x))\mathbb{E}_{\mathcal{D}}\big[(g^{(\mathcal{D})}(x) - \bar{g}(x))\big] \\
&= 2(\bar{g}(x) - f(x))\big(\mathbb{E}_{\mathcal{D}}\big[g^{(\mathcal{D})}(x)\big] - \bar{g}(x)\big) \\
(\text{using eq.3}) &= 2(\bar{g}(x) - f(x))\big(\bar{g}(x) - \bar{g}(x)\big) \\
&= 0
\end{aligned}$$

در نهایت با استفاده از رابطه‌ی بالا، رابطه‌ی ۷ به شکل زیر در می‌آید:

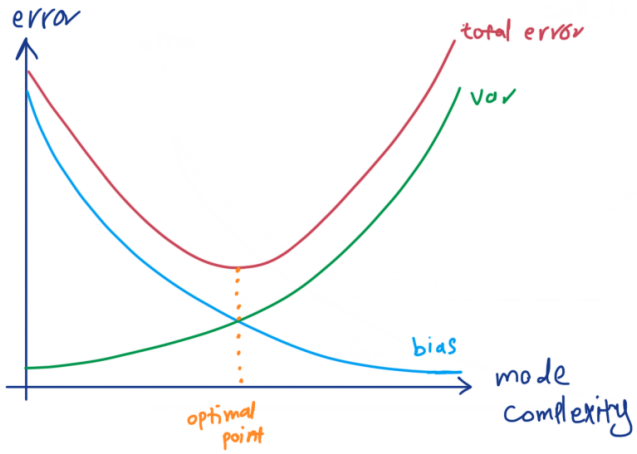
$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(x) - f(x))^2\right] \\
&\quad = \mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(x) - \bar{g}(x))^2\right] + \mathbb{E}_{\mathcal{D}}\left[(\bar{g}(x) - f(x))^2\right] \\
&\quad = \underbrace{\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(x) - \bar{g}(x))^2\right]}_{var(x)} + \underbrace{(\bar{g}(x) - f(x))^2}_{bias(x)}
\end{aligned}$$

حال با جایگذاری این عبارت در ۲ داریم:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathbf{x}}\Big[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(x) - f(x))^2\right]\Big] \\
&= \mathbb{E}_{\mathbf{x}}\Big[bias(x) + var(x)\Big] \\
&= bias + var
\end{aligned}$$

۲.۱ ب

با فرض تعداد نمونه‌های ثابت،نمودار خطای کل و همچنین اریبی و واریانس به شکل زیر خواهد بود.



در ابتدا وقتی مدل ساده‌است فضای فرضیه کوچک بوده و فاصله‌ی $\bar{g}$ از f زیاد است. در نتیجه مقدار بایاس بسیار بزرگ است. از طرف دیگر به دلیل کوچک بودن فضای فرضیه g به $\bar{g}$ نزدیک است و مقدار واریانس کوچک است. هر چه مدل پیچیده‌تر می‌شود فضای فرضیه بزرگتر شده و $\bar{g}$ به f نزدیک‌تر می‌شود. در نتیجه بایاس کاهش می‌یابد ولی به دلیل اینکه $\bar{g}$ از g دور می‌شود، واریانس افزایش پیدا می‌کند.

۳.۱ پ

در صورتی که فضای فرضیه کوچک باشد بهترین مدلی که می‌توانیم به آن برسیم($\bar{g}$) با تابع هدف(f) فاصله‌ی زیادی خواهد داشت و در نتیجه اریبی مقدار بالایی به خود می‌گیرد.

راه حل اول: در نتیجه یک راه حل پیشنهادی این است که از فضای فرضیه‌ای پیچیده تر و با پارامترهای بیشتر استفاده کنیم تا بدین ترتیب احتمال نزدیک بودن $\bar{g}$ به f و درنتیجه کاهشش اریبی را بیشتر کنیم. به عنوان مثال به جای رگرسیون خطی از شبکه‌های عصبی استفاده کنیم.

راه حل دوم: راه حل دوم می‌تواند بیشتر کردن تعداد ویژگی‌های غیر همبسته باشد. به این شکل که هرچه تعداد ویژگی‌های غیرهمبسته موجود در دیتاست را بیشتر کنیم، تعداد پارامترهای مدل افزایش پیدا می‌کند و فضای فرضیه بزرگتر می‌شود. در نتیجه اریبی کاهش می‌یابد.

۴.۱ ت

هر چقدر تعداد ویژگی‌ها کاهش یابد، تعداد پارامترهای مدل استفاده شده نیز کاهش یافته و درنتیجه فضای فرضیه کوچک و کوچکتر می‌شود و در نتیجه احتمال فاصله گرفتن $\bar{g}$ و f افزایش می‌یابد و منجر به افزایش بایاس می‌شود . همچنین با کوچک شدن فضای فرضیه g به $\bar{g}$ نزدیک شده و واریانس کاهش می‌یابد.

۵.۱ ث

راه حل اول: برای کاهش واریانس یک راه افزایش تعداد نمونه‌هاست. با این کار باند تعمیم‌پذیری(Generalization) در نامساوی هافدینگ کوچکتر می‌شود پس Generalization بهبود خواهد یافت که در نتیجه واریانس کاهش می‌یابد.
راه حل دوم: راه حل دوم کاهش پیچیدگی مدل است. هرچه تعداد ویژگی‌ها کمتر شود تعداد پارامترها کاهش می‌یابد و بدین ترتیب پیچیدگی مدل کم می‌شود. با کاهش پیچیدگی مدل فضای فرضیه کوچکتر می‌شود و بدین ترتیب واریانس کاهش می‌یابد. پس می‌توان به عنوان مثال برخی ویژگی‌ها که دارای اطلاعات کمتری هستند را از مدل کنار گذاشت.

۶.۱ ج

ویژگی همبسته ویژگی هایی هستند که عملا اطلاعات خاصی به مدل اضافه نمی‌کنند. وقتی ما یک ویژگی که با ویژگی دیگری همبسته است را از دیتاست حذف می‌کنیم، فضای فرضیه تغییری نمی‌کند. زیرا آن ویژگی جزو ویژگی‌های مفید مدل که حجم فضای فرضیه را مشخص می‌کنند، نبوده است. با عدم تغییر فضای فرضیه بایاس ثابت باقی می‌ماند. زیرا فاصله‌ی $\bar{g}$ از f تغییر نمی‌کند. اما به دلیل اینکه آن ویژگی ممکن است حاوی نویز بوده باشد و عملا به دلیل وجود نویز ما را در رسیدن به تابع هدف گمراه کرده باشد، پس حذف کردن آن باعث کاهش واریانس می‌شود. زیرا حساسیت مدل به نویز را کاهش می‌دهد و فرضیه‌ی بهتری را می‌یابد.

۲ سوال دوم

فرض کنید روش $MaximumLiklihood$ را انتخاب می‌کنیم. در آن‌صورت درواقع قصد داریم مقدار عبارت زیر را بشینه کنیم:

$$argmax_{w\in\mathbb{R}^d}\mathbb{P}(D|w)$$

چنین پارامتری که عبارت را بیشینه می‌کند $\hat{w}_{ML}$ می‌نامیم. از آنجایی که داده‌های ما از توزیع $i.i.d$ آمده‌اند،پس عبارت را به شکل زیر می‌توان نوشت:

$$\Pi_{x_i\in\mathcal{X}}\mathbb{P}(x_i|w)$$

که اگر از عبارت بالا لگارتیم بگیریم داریم :

$$\mathcal{L}=\sum_{x_i\in\mathcal{X}} log(\mathbb{P}(x_i|w))$$

به عبارتی :

$$\hat{w}_{ML}=argmax_w\mathcal{L}$$

برای اینکه به چنین عبارتی برسیم از L مشتق می‌گیریم:

$$\frac{\partial\mathcal{L}}{\partial w_i}=0,\forall w_i\in w$$

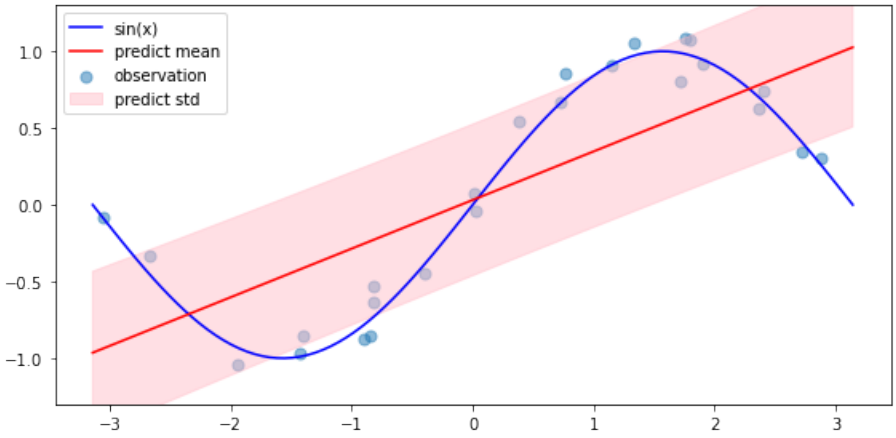
و بدین شکل حاصل را به دست می‌آوریم و w بهینه حاصل می‌شود. مزیت روش MAP بر ML این است که در آن می‌توانیم دانسته‌های خودمان در مورد فضای مساله را نیز لحاظ کنیم و از آن استفاده نماییم. اما در روش تخمین بیزین برعکس دو روش قبلی که فقط یک پارامتر بازگردانده می‌شد، کل عبارت $\mathbb{P}(w|\mathcal{X})$ محاسبه می‌شود و میتوانیم مقدار متوسط این عبارت را به عنوان مقداری مناسب برای w لحاظ کنیم. در این روش داریم:

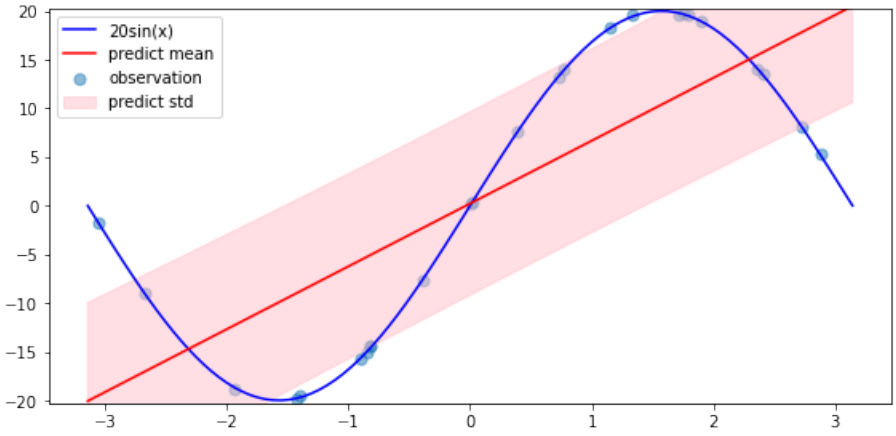
$$\mathbb{P}(w|\mathcal{X})=\frac{\mathbb{P}(w)\times\mathbb{P}(\mathcal{X}|w)}{\mathbb{P}(\mathcal{X})}$$

که برای محاسبه‌ی $\mathcal{X}$ داریم:

$$\mathbb{P}(x) = \int_w \mathbb{P}(\mathcal{X}|w)P(w)dw$$

حال برای اینکه بتوانیم این دو الگوریتم را با هم مقایسه کنیم، برای روش بیزین دو شبیه‌سازی برای تابع $sin(x)$ و $20sin(x)$ انجام شده است که نوت‌بوک ضمیمه شده است. برای تابع اول همانطور که مشخص است، واریانس روش بیزی بسیار واریانس مناسبی می‌باشد و باتوجه به اینکه تابع پیش‌بینی شده دارای شیب‌های تندی است، این عامل باعث نشده که واریانس مدل پایین بیاید. دلیل این امر آن است که برای تنظیم پارامترهای این مدل از مقدار متوسط کل پارامترها استفاده می‌شود که در نتیجه به دلیل همین خاصیت میانگین‌گیری، واریانس کاهش می‌یابد. برای تابع دوم نیز این امر مشهود است





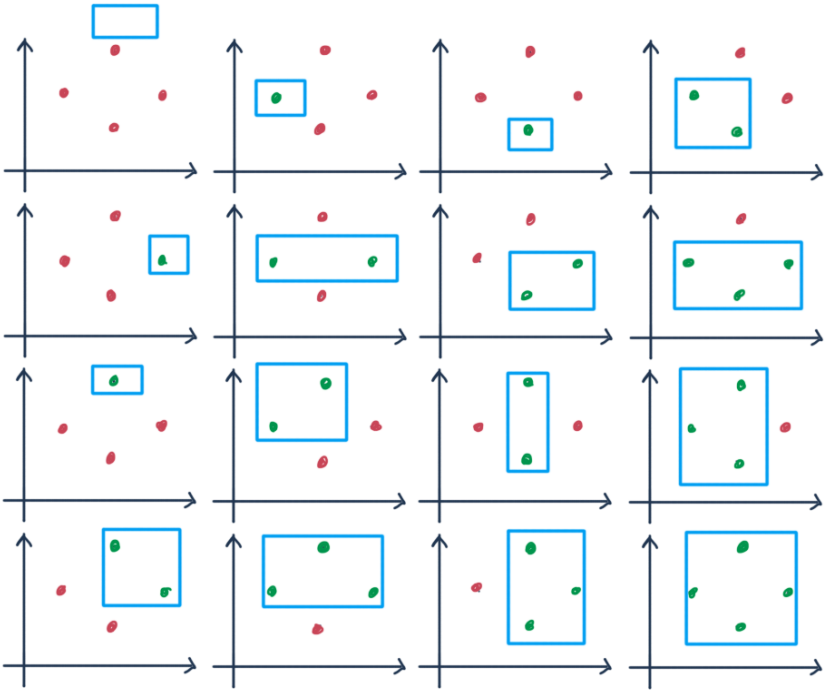
ولی در مورد بایاس، از آنجایی که بایاس به معنی فاصله‌ی بهترین فرضیه‌ی خطی ممکن از فرضیه‌ی ایده‌ال است در نتیجه این مقدار برای روش اول و دوم تفاوتی ندارد و همان مقدار است. پس از دیدگاه بایاس واریانس روش گوسی به دلیل انجام عمل میانگین‌گیری کاهش واریانس را در بر خواهد داشت.

۳ سوال سوم

۱.۳ آ

۱.۱.۳ قسمت ۱

برای یک و دو و سه نقطه مشخص است که به بیشتری تعداد دایکاتمی می‌رسیم.برای چهار نقطه به شکل زیر می‌توان به همه‌ی ۱۶ دایکاتمی متفاوت رسید.

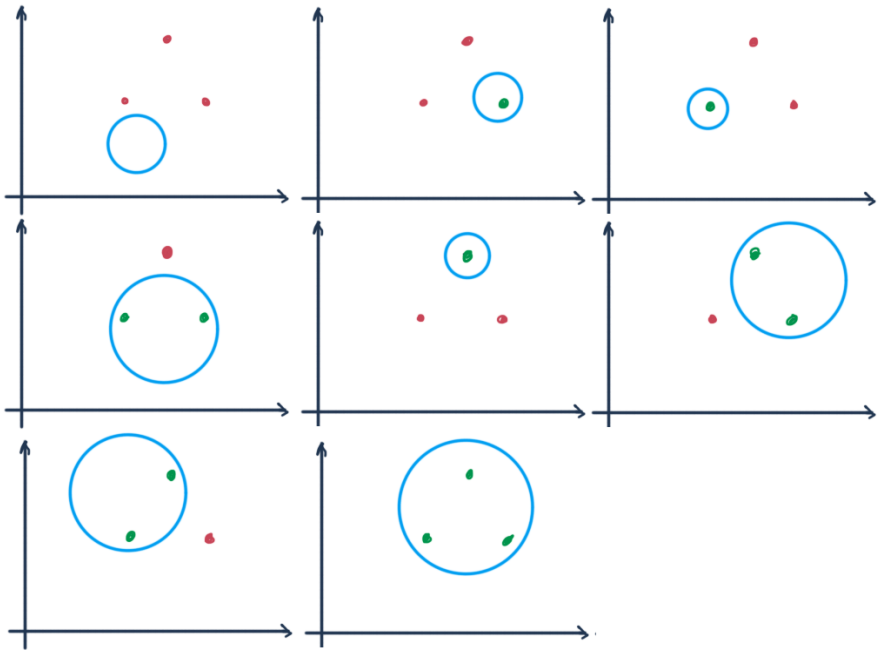


برای پنج نقطه، برای اینکه بتوانیم به همه‌ی دایکاتمی های ممکن برسیم آن ها را حول یک پنج ضلعی منتظم قرار می‌دهیم. ولی حتی در این حالت نیز به هیچ وجه نمی‌توانیم چنین برچسب‌دهی را بسازیم:

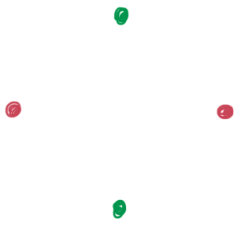
در نتیجه بعد VC این فضا برابر ۴ است.

۲.۱.۳ قسمت ۲

برای یک و دو نقطه مشخص است که می‌توانیم به بیشترین تعداد دایکاتمی های ممکن برسیم. برای سه نقطه به شکل زیر می‌توان به همه‌ی ۸ دایکاتمی ممکن رسید



برای چهار نقطه، برای اینکه بتوانیم به همه‌ی دایکاتمی های ممکن برسیم آن ها را حول یک چهار صلعی منتظم قرار می‌دهیم. ولی حتی در این حالت نیز به هیچ وجه نمی‌توانیم چنین برچسب‌دهی را بسازیم:



پس برای این فضای فرضیه بعد VC برابر ۳ است.

۳.۱.۳ قسمت ۳

این مساله همان مساله‌ی پرسپترون است که در آن $w_0 = 0$ زیرا ابرصفحه‌ی آن حتما باید از مرکز مختصات عبور کند. پس برای هر $x = (x_1, x_2, ..., x_d)$ داده شده، برچسب را به شکل زیر به آن داده نسبت می‌دهیم.

$$h(x) = sign(\sum_{i=1}^{d} w_i x_i) = sign(w^T x)$$

مشابه کاری که در کلاس انجام شد برای اینکه ثابت کنیم $d_{vc} = d$ دو عبارت را ثابت می‌کنیم:

$$d_{vc} = d \Leftrightarrow d_{vc} \leq d, d_{vc} \geq d$$

برای اثبات اینکه $d_{vc} \geq d$ کافی است ثابت کنیم d نقطه وجود دارند که توسط این الگوریتم Shatter می‌شوند. بدین منظور داده‌ها را به شکل زیر می‌سازیم:

$$X = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ . & . & . & ... & 0 \\ 0 & 0 & 0 & ... & 1 \end{bmatrix}$$

حال برچسب‌دهی های زیر برای این داده‌ها ممکن است :

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_d \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ . \\ . \\ . \\ \pm 1 \end{bmatrix}$$

می‌خواهیم بردار وزنی به نام w بیابیم که

$$Xw = y \Rightarrow sign(Xw) = y$$

چون X ماتریس همانی بوده معکوس‌پذیر و معکوس آن برابر خودش است، داریم:

$$Xw = y \Rightarrow w = X^{-1}y = Xy = y$$

حال به سراغ بخش دوم اثبات می‌رویم و نشان می‌دهیم که $d_{vc} \leq d$ یعنی هیچ $d+1$ نقطه‌ای نداریم که توسط این فضای فرضیه $Shatter$ شوند. با فرض خلف $d+1$ نقطه در نظر می‌گیریم که توسط این فضای فرضیه $Shatter$ می‌شوند:

$$x_1, ..., x_d, x_{d+1}$$

هیچکدام از این نقاط بردار ۰ نیستند. زیرا برای بردار صفر داریم :

$$h(\vec{0}) = sign(\sum_{i=1}^{d} 0) = sign(0)$$

یعنی بردار صفر یک مقدار ثابت $sign(0)$ دارد. اگر علامت ۰ را مثبت تعیین کنیم این بردار همواره برچسب مثبت میخورد وگرنه برچسب منفی. پس به هر حال نمیتوان همه‌ی دایکاتمی ها را با این تقطه به دست آورد. حال از جبر خطی می‌دانیم اگر تعداد بردارها از تعداد ابعاد صفحه بیشتر شود، یکی از بردارها، ترکیب خطی بردارهای دیگر است. پس :

$$\exists j : x_j = \sum_{i \neq j} a_i x_i$$

چون هیچ‌کدام از x_i ها بردار صفر نیستند پس همه‌ی a_iها مقدار صفر ندارند. پس حداقل یک a_i غیر صفر داریم. روی این نقاط دایکاتمی زیر را تعریف می‌کنیم:

$$y_j = -1$$
$$\forall i \neq j, a_i \neq 0, y_i = sign(a_i)$$
$$\forall i \neq j, a_i = 0, y_i = +1$$

طبق رابطه‌ی بالا داریم:

$$x_j = \sum_{i \neq j} a_i x_i \Rightarrow w^T x_j = \sum_{i \neq j} a_i w^T x_i$$

اما از آنجایی که برای a_iهای غیر صفر:

$$sign(a_i) = y_i = sign(w^T x_i)$$

پس حاصل $a_i w^T x_i$ مثبت است و در نتیجه حاصل سیگما مثبت است. پس $y_j = +1$ که با $y_j = -1$ در تناقض است.

۲.۳ ب

طبق باند هافدینگ داریم:

$$\mathbb{P}(|E_{in} - E_{out}| > \epsilon) \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

در اینجا می‌خواهیم دقت 0.95 باشد. پس داریم:

$$\mathbb{P}(|E_{in} - E_{out}| > 0.05) \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}(0.05)^2 N}$$

همچنین برای فضای H_c داریم:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{vc}} \binom{N}{i} = \sum_{i=0}^{3} \binom{N}{i} = \frac{N^3 + 5N + 6}{6}$$

پس:

$$m_{\mathcal{H}}(2N) \leq \frac{8N^3 + 10N + 6}{6}$$

در نتیجه:

$$\mathbb{P}(|E_{in} - E_{out}| > 0.05) \leq 4(\frac{8N^3 + 10N + 6}{6})e^{-\frac{1}{8}(0.05)^2 N}$$

حال چون می‌خواهیم احتمال یادگیری بالای ۹۰ درصد باشه، قرار می‌دهیم:

$$4(\frac{8N^3 + 10N + 6}{6})e^{-\frac{1}{8}(0.05)^2 N} \leq 0.1$$

که با حل این نامعادله به وسیله کامپیوتر به پاسخ زیر می‌رسیم:

$$N \geq 125424$$

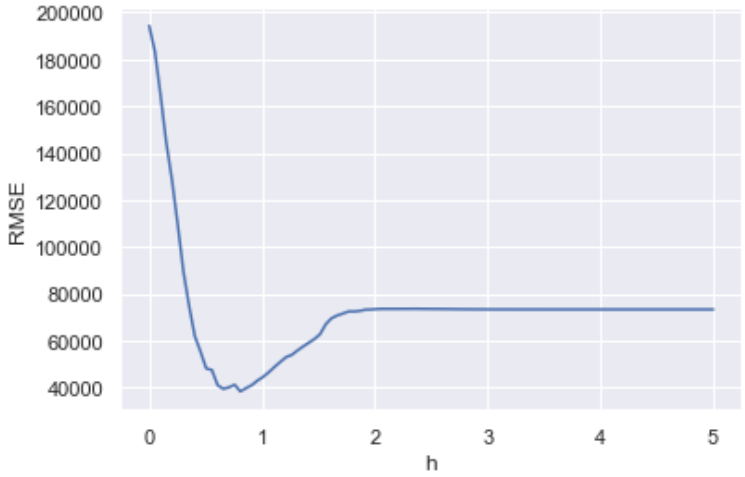
۳.۳ پ

از همان ایده‌ی گوی ها استفاده می‌کنیم. فرض می‌کنیم برای هریک از عناصر فضای حالت یک ظرف داریم که هر کدام به تعدادی گوی داریم که هر گوی یکی از عناصر توزیعی است که داده از آن توزیع آمده است. اگر فرضیه‌ی مربوطه نمونه‌ی درون ظرف را درست برچسب زده باشد،گوی سبز وگرنه قرمز است.از هرکدام از آن ها ۲ نمونه‌ی Nتایی اسخراج کرده و تعداد گوی‌های قرمز درونشان را E_{in} و E'_{in} می‌نامیم. اگر فضای فرضیه به اندازه‌ی کافی بزرگ باشد در ۹۰ درصد ظرف‌ها اختلاف تعداد قرمزها در E_{in} و E'_{in} حداکثر ۵ درصد است.

۴ سوال چهارم

۱.۴ آ

برای این قسمت و قسمت بعدی توضیحات کامل درون نوت‌بوک نوشته شده و در قالب کامنت آمده است. برای این قسمت با توجه به انجام مراحل نرمال‌سازی قبل از پیاده‌سازی مدل، نمودار زیر به دست آمد:

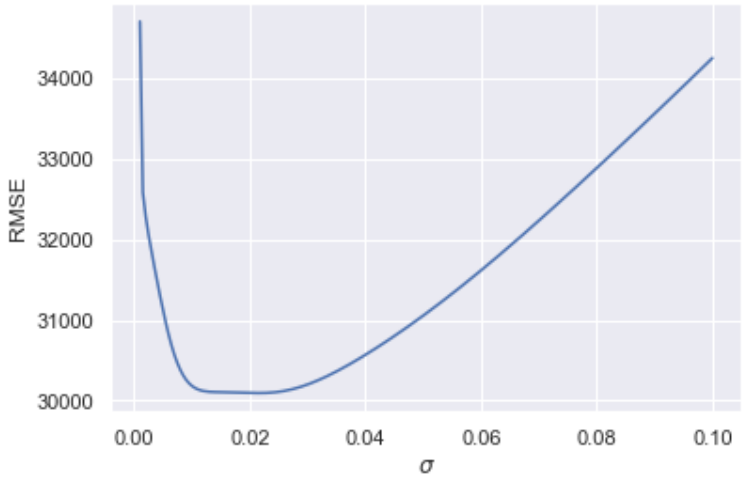


که کمترین خطادر نقطه‌ی $h = 0.8$ به دست آمد که برابر 38291 می‌باشد.

۲.۴ ب

۱.۲.۴ قسمت ۱

برای این قسمت با توجه به انجام مراحل نرمال‌سازی قبل از پیاده‌سازی مدل، نمودار زیر به دست آمد:



که کمترین خطادر نقطه‌ی $\sigma = 0.02$ به دست آمد که برابر 30091 می‌باشد.

۲.۲.۴ قسمت ۲

هرچه تعداد ویژگی‌های موثر کمتر شود نقاط به شکل صحیحی از هم فاصله نمی‌گیرند و در یک یا چند بعد به هم نزدیک‌تر می‌شوند. همین عامل باعث کوتاه شدن اندازه‌ی بردار $x_i - x_j$ می‌شود برای آن نقاط می‌شود. برای اینکه بر این کم شدن فاصله فائق بیاییم باید مخرج $x_i - x_j$ در فرمول کرنل گوسی یعنی σ را هم کاهش دهیم تا وزن کلی کرنل روی آن نقاط دچار تغییرات ناگهانی نشود و عملکرد مدل کمتر مختل شود. پس با کاهش تعداد ویژگی‌های موثر، باید σ هم کاهش یابد.

۳.۲.۴ قسمت ۳

۳.۴ پ

کرنل اول برای محاسبه‌ی فاصله‌ی بین دو نقطه یک معیار سفت و سخت دارد و فقط نقاطی را شبیه یک نقطه در نظر می‌گیرد که در فاصله‌ی مشخصی از آن باشند و در غیر این‌صورت شباهت آنها را صفر تعیین می‌کند. در حالی که هسته گوسی حتی دور ترین نقاط به نقطه‌ی فعلی را با یک احتمال مشخصی در انتخاب برچسب برای آن نقطه دخیل می‌کند. با توجه به این موضوع در اکثر موارد هسته گوسی عملکرد بهتری از خود نشان خواهد داد. اگر داده‌های موجود چگال باشند به شکلی که حول هر نقطه به تعداد زیادی نقطه موجود باشد در آن صورت کرنل اول نیز می‌تواند عملکرد بهتری از خود نشان دهد.

۴.۴ ت

وقتی داده‌ها تنک و یا Categorical باشند در آن صورت باتوجه به اینکه مقادیر زیادی اختیار نمی‌کنند، وجود یک نویز در داده‌ها کافی است که عملکرد رگرسیون خطی مختل شود. زیرا یک نویز می‌تواند تاثیر زیادی روی معیار خطای رگرسیون خطی بگذارد. اما اگر از کرنل گوسی استفاده شود، با توجه به اینکه این کرنل همه‌ی نقاط را با درصد مشخصی (بر حسب معکوس قاصله) برای انتخاب برچسب نقطه‌ی فعلی دخیل می‌کند نسبت به داده‌های تنک و یا Categorical حساسیت کمتری دارد و می‌تواند آن‌ها را به شکل بهتری پیش بینی کند.