400108547 تمرین ۱ تشریحی ML پرهام رضایی

سوال اول

۱. قرار می‌دهیم $E(X|Y)=g(Y)$ ، داریم که
متغیر تصادفی برحسب Y است.

طبق LOTUS *

$$E(E(X|Y)) = E_Y(g(Y)) \overset{*}{=} \sum_y g(y)P(Y=y) = \sum_y E(X|Y)P(Y=y) =$$

$$= \sum_y \Big(\sum_x x P(X=x|Y=y)\Big)P(Y=y) = \sum_y \sum_x x \underbrace{P(X=x|Y=y)P(Y=y)}_{\text{بیز}} = \sum_y \sum_x x P(X=x, Y=y)$$

$$= \sum_x \sum_y \underbrace{P(X=x, Y=y)}_{\text{مارجینال روی } Y} = \sum_x x P(X=x) = E(X) \qquad Q.E.D$$

۲. قرار می‌دهیم $g(Y)=E_X(X|Y)$ ، طبق الف (Adam's law) داریم که $E(g(Y))=E(X)$

$$E(Var(X|Y)) = E(E(Y^2|X) - E(Y|X)^2) = \underbrace{E(E(Y^2|X))}_{\text{طبق الف (adam)}} - E(g(Y)^2) = E(X^2) - E(g(Y)^2)$$

$$Var(E(X|Y)) = E(E(X|Y)^2) - E(E(X|Y))^2 = E(g(Y)^2) - \underbrace{E(g(Y))}_{E(X)}^2 = E(g(Y)^2) - E(X)^2$$

$$\Rightarrow E(Var(X|Y)) + Var(E(X|Y)) = E(X^2) - E(g(Y)^2) + E(g(Y)^2) - E(X)^2 = E(X^2) - E(X)^2 = Var(X)$$

Q.E.D

۳.

$$\sum_{i=1}^{n} P[X_i]\log P[X_i] + \sum_{i=1}^{n} P[Y_i]\log P[Y_i] \leq \sum_{i=1}\sum_{j=1} P[X=i,Y=j]\log P[X=i,Y=j]$$

⊛ معادلاً

$$0 \leq -\sum P(X_i)\log P(X_i) - \sum P(Y_i)\log P(Y_i) + \sum\sum P(X_i,Y_j)\log P(X_i,Y_j)$$

$$0 \leq H(X) + H(Y) - H(X,Y)$$

$$\sum_{i=1}^{n} P[X_i]\log P[X_i] = \sum_{i=1}^{n}\sum_{j=1}^{n} \underbrace{P(X_i,Y_j)}_{\text{مارجینال} \leftarrow \text{۲ متغیره}}\log P(X_i)$$

مشابهاً ←

$$\sum_{j=1}^{n} P(Y_j)\log P(Y_j) = \sum_{j=1}^{n}\sum_{i=1}^{n} P(X_i,Y_j)\log P(Y_j)$$

⊛ $\Rightarrow RHS = \sum_{i=1}^{n}\sum_{j=1}^{n} P(X_i,Y_j)\left(-\log P(X_i) - \log P(Y_j) + \log P[X_i,Y_j]\right)$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} P(X_i,Y_j)\log\frac{P(X_i,Y_j)}{P(X_i)P(Y_j)}$$

باید RHS ≥ 0 را نشان دهیم. درون لاگ را معکوس کرده تا مسئله معادل ≤ 0 شدن عبارت زیر شود.

روش اول: log تابعی کرب است پس (جمع ضرایب ۱ است ← میانگین وزن‌دار)

$$\sum_{i,j} P(X_i,Y_j)\log\frac{P(X_i)P(Y_j)}{P(X_i,Y_j)} \leq \log\sum_{i,j} P(X_i,Y_j)\frac{P(X_i)P(Y_j)}{P(X_i,Y_j)} = \log\sum_{i,j} P(X_i)P(Y_j)$$

حال پس کافی است $\log\sum_{i,j} P(X_i)P(Y_j) \leq 0$ شود.

معادلاً (با پایه بردن log به توان عبارت):

$$\sum_{i,j} P(X_i)P(Y_j) \overset{?}{\leq} 1$$

داریم که

$$(\square)\quad \sum_{i=1}^{n}\sum_{j=1}^{n} P(X_i)P(Y_j) = \sum_{i=1}^{n} P(X_i) \underbrace{\sum_{j=1}^{n} P(Y_j)}_{1} = \sum_{i=1}^{n} P(X_i) = 1$$

پس عبارت ثابت شد $\boxed{\checkmark}$

روش دوم : در عبارت $\sum_{i,j} P(X_i, Y_j) \log \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)} \geq 0$ که می‌خواهیم ثابت

کنیم، توزیع R را تعریف می‌کنیم $R(X_i, Y_j) = P(X_i)P(Y_j)$.

همواره بزرگتر مساوی صفر و طبق $(\square)$ مجموع احتمالش ۱ است. حال داریم که

$$\sum_{i,j} P(X_i, Y_j) \log \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)} = \sum_{i,j} P(X_i, Y_j) \log \frac{P(X_i, Y_j)}{R(X_i, Y_j)}$$

$= KL(P \| R)$ که می‌دانیم KL divergence که صفر است.

سوال دوم

۱.

$$Y = \beta_1^* X + \beta_0^* + \epsilon$$

داریم که برای درجه یک

$$y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$J(\beta) = \|y - X\beta\|^2 \rightarrow \nabla_\beta J(\beta) = -2X^T(y - X\beta)$$

$$\nabla_\beta J(\beta) = 0 \Rightarrow X^TX\beta = X^Ty \Rightarrow \beta = (X^TX)^{-1}X^Ty \quad \Big\}\, *$$

$$y = X\beta^* + \epsilon \rightarrow J = \|X\beta^* + \epsilon - X(X^TX)^{-1}X^T(X\beta^* + \epsilon)\|^2 =$$

$$\|X\beta^* + \epsilon - X\underbrace{(X^TX)^{-1}X^TX}_{I}\beta^* - X(X^TX)^{-1}X^T\epsilon\|^2$$

$$= \|X\beta^* + \epsilon - X\beta^* - X(X^TX)^{-1}X^T\epsilon\|^2 = \|(I - X(X^TX)^{-1}X^T)\epsilon\|^2$$

هزینه روی دیتای ترین.

برای مدل درجه ۳ نسبت به * برای Φ داریم که:

$$\Phi = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ & & \vdots & \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$$

$$\beta = (\Phi^T\Phi)^{-1}\Phi^Ty$$

$$\Rightarrow J = \|X\beta^* + \epsilon - \Phi(\Phi^T\Phi)^{-1}\Phi^T(X\beta^* + \epsilon)\|^2 =$$

$$\|(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T)(X\beta^* + \epsilon)\|^2$$

استدلال:

hypothesis space

از آنجایی که در مدل درجه سه فضای فرضیه پیچیده تر است، approximation بهتری دارد

دیتای ترین را با قابلیت تعمیم پذیری کم.

از آنجایی که فضای چندجمله‌ای‌ها حداکثر درجه یک زیرمجموعه فضای چندجمله‌ای‌ها حداکثر درجه سه است، و mse که ارور ما است convex است، اگر مینیمم کردن نیم‌ها جواب درجه یکی می داد این جواب توسط درجه سه نیز قابلیت ایجاد داشت.

بنابراین ارور روی دیتای ترین در مدل درجه سه حداکثر اندازه ارور مدل درجه 1 است.

در اینجا در واقع اورفیت رخ می دهد و به دلیل فیت شدن با نویزها مسئله ارور مدل درجه سه کمتر خواهد بود در دیتای ترین ولی در کل توزیع (یا دیتای تست) ارور افزایش زیادی خواهد داشت.

* عملکرد روی دیتای تست درجه سه البته اگر دیتا خیلی زیاد باشد، بهبود می یابد.

۲.

تابع پیشنهادی:

$$f(x) = \frac{35}{68}x^4 - \frac{30}{68}x^2 + \frac{3}{68}$$

نحوه ایجاد تابع در ... در انتهای حل آمده و خواص + برایش اثبات شده است. (همین طوری ساخته شده است)

+ داریم که

$$\circledast \quad \int_{-1}^{1} f(x)\,dx = \int_{-1}^{1} x f(x)\,dx = \int_{-1}^{1} x^2 f(x)\,dx = \int_{-1}^{1} x^3 f(x)\,dx = 0$$

حال

$$y = f(x) + \epsilon(x)$$

برای مدل خطی داریم که

$$\hat{y} = \beta_1 x + \beta_0$$

n اندازه دیتا آموزشی:

$$\rightsquigarrow \quad mse = \frac{1}{n}\sum_{n} (y_i - \beta_1 x_i - \beta_0)^2$$

داریم که بررسی در $n \to \infty$ است.

در مورد بهینه

$$\frac{\partial mse}{\partial \beta_0} = 0 \Rightarrow \frac{-2}{n}\sum_{n}(y_i - \beta_1 x_i - \beta_0) = 0$$

$$\Rightarrow \lim_{n\to\infty} \frac{1}{n}\sum_{n} y_i - \frac{\beta_1}{n}\sum x_i - \beta_0 = 0$$

$$\Rightarrow \int_{-1}^{1} y(x)\,dx - \frac{\beta_1}{n}\underbrace{\int_{-1}^{1} x\,dx}_{\frac{x^2}{2}\Big|_{-1}^{1} = 0} - \beta_0 = 0 \Rightarrow \underbrace{\int_{-1}^{1} f(x)\,dx}_{=0 \ \circledast} + \underbrace{\int_{-1}^{1} \epsilon(x)\,dx}_{E_x(\epsilon(x)) = 0} - \beta_0 = 0 \Rightarrow \beta_0 = 0$$

$$\frac{\partial mse}{\partial \beta_1} = \frac{-2}{n}\sum_{n}(y_i - \beta_1 x_i - \beta_0)\,x_i = 0$$

$$\Rightarrow \lim_{n\to\infty} \frac{1}{n}\sum y_i x_i - \frac{\beta_1}{n}\sum x_i^2 - \underbrace{\beta_0}_{0} x_i = 0$$

$$\Rightarrow \frac{1}{n}\int_{-1}^{1} x y(x)\,dx - \frac{\beta_1}{n}\int_{-1}^{1} x^2\,dx = 0 \Rightarrow \underbrace{\int_{-1}^{1} x f(x)\,dx}_{=0 \ \circledast} + \int_{-1}^{1} x\epsilon(x)\,dx = \beta_1 \frac{x^3}{3}\Big|_{-1}^{1} = \frac{2\beta_1}{3}$$

$$\Rightarrow \int_{-1}^{1} x\,\epsilon(x)\,dx = \frac{2\beta_1}{3}$$

$$\square)\ \int_{-1}^{1} x\,\epsilon(x)\,dx = E(x\,\epsilon(x)) = E(x\,\epsilon(x)) - \underbrace{E(x)}_{0}\underbrace{E(\epsilon(x))}_{0} = \operatorname{cov}(x, \epsilon(x)) = 0$$

چون مستقل از $\epsilon(x_i)$ ها

$$\Rightarrow \frac{2\beta_1}{3} = 0 \Rightarrow \beta_1 = 0$$

$$E_x^{true}\left((y(x) - \underset{0}{\beta_1} x - \underset{0}{\beta_0})^2 \right) = \underbrace{E_x^{true}\left((f(x) + \epsilon(x))^2 \right)}_{0} =$$

$$\int_{-1}^{1} f(x)\,dx + 2\int_{-1}^{1} f(x)\,\epsilon(x) + \int_{-1}^{1} \epsilon(x)^2$$

برای مدل درجه ۳ داریم که

$$mse = \frac{1}{n} \sum_n (y_i - \beta_3 x_i^3 - \beta_2 x_i^2 - \beta_1 x_i^1 - \beta_0)^2$$

مشابه قبل دو حالت داریم که:

$$\frac{\partial mse}{\partial \beta_0} = 0 \Rightarrow \frac{1}{n}\sum y_i - \frac{\beta_3}{n}\sum x_i^3 - \frac{\beta_2}{n}\sum x_i^2 - \frac{\beta_1}{n}\sum x_i - \beta_0 = 0$$

$$= E_x(y(x)) - \beta_3 E_x(x^3) - \beta_2 E(x^2) - \beta_1 E(x) - \beta_0 = 0$$

$$= \int_{-1}^{1} y(x)\cdot\frac{1}{2}\,dx - \beta_3 \int_{-1}^{1} x^3 \frac{1}{2}\,dx - \beta_2 \int_{-1}^{1} x^2 \frac{1}{2}\,dx - \beta_1 \int_{-1}^{1} x\cdot\frac{1}{2}\,dx - \beta_0 = 0$$

$$= \frac{1}{2}\left(\underbrace{\int_{-1}^{1} f(x)\,dx}_{0} + \underbrace{\int_{-1}^{1} \epsilon(x)\,dx}_{=E(\epsilon(x))=0} - \beta_3 \frac{x^4}{4}\Big|_{-1}^{1} - \beta_2 \frac{x^3}{3}\Big|_{-1}^{1} - \beta_1 \frac{x^2}{2}\Big|_{-1}^{1} \right) - \beta_0 = 0$$

$$= \frac{1}{2}\left(-\frac{2\beta_2}{3}\right) - \beta_0 = 0 \Rightarrow \boxed{\beta_2 = -3\beta_0} \quad \mathrm{I}$$

$$\frac{\partial mse}{\partial \beta_1} = 0 \Rightarrow -\frac{1}{n}\sum_n \left(y_i - \beta_3 x_i^3 - \beta_2 x_i^2 - \beta_1 x_i - \beta_0\right) x_i = 0$$

$$\lim_{n\to\infty} \quad E(x\,y(x)) - \beta_3 E(x^4) - \beta_2 E(x^3) - \beta_1 E(x^2) - \beta_0 E(x) = 0$$

$$\times 2 \Rightarrow \underbrace{\int_{-1}^{1} x f(x)\,dx}_{0} + \int_{-1}^{1} x\,\epsilon(x)\,dx - \beta_3 \int_{-1}^{1} x^4\,dx - \beta_2 \underbrace{\int_{-1}^{1} x^3\,dx}_{0} - \beta_1 \int_{-1}^{1} x^2\,dx - \beta_0 \underbrace{\int_{-1}^{1} x\,dx}_{0} = 0$$

$$\Rightarrow \int_{-1}^{1} x\,\epsilon(x)\,dx - \beta_3 \left.\frac{x^5}{5}\right|_{-1}^{1} - \beta_1 \left.\frac{x^3}{3}\right|_{-1}^{1} =$$

$$\int_{-1}^{1} x\,\epsilon(x)\,dx - \frac{2\beta_3}{5} - \frac{2\beta_1}{3} = 0 \Rightarrow \frac{2\beta_3}{5} + \frac{2\beta_1}{3} = \int_{-1}^{1} x\,\epsilon(x)\,dx$$

$$\overset{\star}{=} 0 \Rightarrow \frac{\beta_3}{5} + \frac{\beta_1}{3} = 0 \Rightarrow \boxed{\beta_3 = -\frac{5}{3}\beta_1} \quad \text{II}$$

$$\star \quad \int_{-1}^{1} x\,\epsilon(x)\,\frac{1}{2}\,dx = E(x\epsilon(x)) = E(x\,\epsilon(x)) - \underbrace{E(x)}_{0}\underbrace{E[\epsilon(x)]}_{0} = Cov(x, \epsilon(x))$$

اما ϵ ما مستقل از x $\Leftarrow$ $=0$

$$\frac{\partial mse}{\partial \beta_2} \quad \frac{1}{n}\sum_n \left(y_i - \beta_3 x_i^3 - \beta_2 x_i^2 - \beta_1 x_i - \beta_0\right)^2 = 0$$

$$\Rightarrow \frac{-2}{n} \sum \left(y_i - \beta_3 x_i^3 - \beta_2 x_i^2 - \beta_1 x_i - \beta_0\right) x_i^2 = 0$$

$$\Rightarrow \frac{1}{n}\sum y_i x_i^2 - \frac{\beta_3}{n}\sum x_i^5 - \frac{\beta_2}{n}\sum x_i^4 - \frac{\beta_1}{n}\sum x_i^3 - \frac{\beta_0}{n}\sum x_i^2 = 0$$

$$\lim_{n\to\infty} \quad E(x^2 y(x)) - \beta_3 E(x^5) - \beta_2 E(x^4) - \beta_1 E(x^3) - \beta_0 E(x^2) = 0$$

$$= \int_{-1}^{1} x^2 y(x)\frac{1}{2}dx - \beta_3 \underbrace{\int_{-1}^{1} x^5 \frac{1}{2}dx}_{0} - \beta_2 \int_{-1}^{1} x^4 \frac{1}{2}dx - \beta_1 \underbrace{\int_{-1}^{1} x^3 \frac{1}{2}dx}_{0} - \beta_0 \int_{-1}^{1} x^2 \frac{1}{2}dx = 0$$

$$\underset{\times 2}{\Longrightarrow} \underbrace{\int_{-1}^{1} x^2 f(x)\,dx}_{0} + \int_{-1}^{1} x^2 \epsilon(x) - \beta_2 \frac{x^5}{5}\Big|_{-1}^{1} - \beta_0 \frac{x^3}{3}\Big|_{-1}^{1} =$$

$$\int_{-1}^{1} x^2 \epsilon(x) - \frac{2\beta_2}{5} - \frac{2\beta_0}{3} = 0 \overset{**}{\Longrightarrow} \frac{\beta_2}{5} + \frac{\beta_0}{3} = 0 \Rightarrow \boxed{\beta_2 = -\frac{5}{3}\beta_0} \quad \text{III}$$

$$\xrightarrow{\text{I}, \text{III}} \beta_2 = \underbrace{-\frac{5}{3}\beta_0 = -3\beta_0} \Rightarrow \boxed{\beta_0 = 0 \Rightarrow \beta_2 = 0}$$

$$** \int_{-1}^{1} x^2 \epsilon(x) \frac{1}{2}dx = E(x^2\epsilon(x)) = E(x^2\epsilon(x)) - E(x^2)\underbrace{E(\epsilon(x))}_{0} = cov(x^2, \epsilon(x))$$

$0 \leftarrow$ مستقل اند

$$\frac{\partial mse}{\partial \beta_3} \quad \frac{1}{n}\sum_{n} (y_i - \beta_3 x_i^3 - \beta_2 x_i^2 - \beta_1 x_i - \beta_0)^2 = 0$$

$$\Longrightarrow \frac{-2}{n}\sum_{n} (y_i - \beta_3 x_i^3 - \beta_2 x_i^2 - \beta_1 x_i - \beta_0)\, x_i^3 = 0$$

$$\Rightarrow \frac{1}{n}\sum_{n} y_i x_i^3 - \frac{\beta_3}{n}\sum x_i^6 - \frac{\beta_2}{n}\sum x_i^5 - \frac{\beta_1}{n}\sum x_i^4 - \frac{\beta_0}{n}\sum x_i^3$$

$$\lim_{n\to\infty} \quad E(x^3 y(x)) - \beta_3 E(x^6) - \beta_2 E(x^5) - \beta_1 E(x^4) - \beta_0 E(x^3)$$

$$= \int_{-1}^{1} x^3 y(x)\frac{1}{2}dx - \beta_3 \int_{-1}^{1} x^6 \frac{1}{2}dx - \beta_2 \underbrace{\int_{-1}^{1} x^5 \frac{1}{2}dx}_{0} - \beta_1 \int_{-1}^{1} x^4 \frac{1}{2}dx - \beta_0 \underbrace{\int_{-1}^{1} x^3 \frac{1}{2}dx}_{0}$$

$$\underset{\times 2}{\Longrightarrow} \underbrace{\int_{-1}^{1} x^3 f(x)\,dx}_{0} + \int_{-1}^{1} x^3 \epsilon(x) - \beta_3 \frac{x^7}{7}\Big|_{-1}^{1} - \beta_1 \frac{x^5}{5}\Big|_{-1}^{1} =$$

$$\int_{-1}^{1} x^3 \epsilon(x)\,dx - \frac{2\beta_3}{7} - \frac{2\beta_1}{5} = 0$$

$\star\star \Rightarrow \quad \frac{\beta_3}{7} + \frac{\beta_1}{5} = 0 \Rightarrow \boxed{\beta_3 = -\frac{7\beta_1}{5}}$ IV

$\star\star$: $\int_{-1}^{1} x^3 \epsilon(x) \frac{1}{2} dx = E(x^3 \epsilon(x)) = E(x^3 \epsilon(x)) - E(x^3) \underbrace{E(\epsilon(x))}_{0} = \mathrm{cov}(x^3, \epsilon(x))$

۰ چون مستقل از ϵ ها.

$\text{II, IV} \Rightarrow \quad \beta_3 = \frac{-5}{3}\beta_1 = \frac{-7}{5}\beta_1 \Rightarrow \boxed{\beta_1 = 0 \Rightarrow \beta_3 = 0}$

$\Rightarrow$

$$E_x^{true}\left((y - \beta_3 x^3 - \beta_2 x^2 - \beta_1 x - \beta_0)^2 \right) =$$

$$E_x^{true}\left((f(x) + \epsilon(x))^2 \right) \quad \square$$

model 1 model 2

$\bigcirc, \square \Rightarrow \quad E_x^{true}(err(m_1^f)) = E_x^{true}(err(m_2^2))$

$$|D| \to \infty : E_{test}\left(|err(m_1) - err(m_2)| \right) = 0$$

نکات قابل ذکر:

اساساً نیازی به امید ریاضی نیست چرا که با $\lim_{n\to\infty}$ ضرایب به درآمدیمی

برای تست دیتای m عنصری $\frac{1}{m}\sum (y_i - f(x_i))^2$

$= \frac{1}{m}\sum y_i^2$

یعنی برای هر دو مدل‌ها ارور در تست دیتای یکسان برابر می‌شود.

اما در تست‌های مختلف داریم که اگر ارور را $\hat{R}_n$ بگذاریم داریم که همان R امید ریاضی است که در صفحه قبل ثابت کردیم برای هر دو برابر است.
روی توزیع

طبق نامساوی چبیشف: $P(|\hat{R}_n(f) - R(f)| \geq \epsilon) \leq \frac{var(\hat{R}_n(f))}{\epsilon^2} = \frac{\sigma_L^2}{n\epsilon^2}$

که $\sigma_L^2 = Var(L_i)$ که $L_i = l(f(x_i), y_i)$ (برای $\lim_{|D|\to\infty}$ ← اندازه دیتای ترین)

* که واریانس خروجی هایمان مقدار مشخص ثابتی است مانند حداکثر c ثابت.

$\rightarrow P(|\hat{R}_n(f) - R(f)| \geq \epsilon) \leq \frac{c^2}{n\epsilon^2}$

مشابهاً: مدل دوم $P(|\hat{R}_n(g) - R(g)| \geq \epsilon) \leq \frac{c^2}{n\epsilon^2}$

$R(g) = R(f)$ *

نامساوی مثلث

$|\hat{R}_n(g) - R(g)| + |R(f) - \hat{R}_n(f)| \geq |\hat{R}_n(g) - \hat{R}_n(f)|$

$\Rightarrow P(|\hat{R}_n(g) - \hat{R}_n(f)| \geq \epsilon) \leq P(|\hat{R}_n(g) - R(g)| + |\hat{R}_n(f) - R(f)| \geq \epsilon) \leq$

احتمال حداقل یکی $< \frac{\epsilon}{2}$

$$P\left(|\hat{R}_n(f) - R(f)| \geq \frac{\epsilon}{2}\right) + P\left(|\hat{R}_n(g) - R(g)| \geq \frac{\epsilon}{2}\right) \leq$$

$$\frac{4c^2}{n\epsilon^2} + \frac{4c'^2}{n\epsilon^2} = \frac{4(c^2 + c'^2)}{n\epsilon^2}$$

پس اگر $n \to \infty$ داریم که RHS به صفر میرود یعنی

$$\lim_{n\to\infty} \left[P\left(|\hat{R}_n(f) - \hat{R}_n(g)| \geq \epsilon\right) \right] = 0$$

$(\epsilon > 0)$

$$\Rightarrow \lim_{n\to\infty} P\left(|\hat{R}_n(f) - \hat{R}_n(g)| \leq \epsilon\right) = 1$$

★ چون مستقل اند داریم که $\underbrace{cov_{f,\epsilon}}_{\text{مستقل}} = E(f\epsilon) - E(f)\underbrace{E(\epsilon)}_{0}$ ← مقدار ضریب تقریب بدست

آمده در اینجا ۲ بجز $E(\epsilon^2)$ ها ثابت اند.

$$\sigma_\epsilon = Var(\epsilon) = E(\epsilon^2) - \underbrace{E(\epsilon)}_{0}^2 \sim \text{مقدار ثابت}$$

محاسبه تابع :

$f(x) = ax^4 + bx^3 + cx^2 + dx + e$

$\int_{-1}^{1} f(x)\,dx = 0 \Rightarrow \left[\frac{ax^5}{5} + \frac{bx^4}{4} + \frac{cx^3}{3} + \frac{dx^2}{2} + ex \right]_{-1}^{1} = \frac{2a}{5} + \frac{2c}{3} + 2e = 0$

$\Rightarrow \frac{a}{5} + \frac{c}{3} + e = 0 \quad i$

$\int x f(x)\,dx = 0 \Rightarrow \left[\frac{ax^6}{6} + \frac{bx^5}{5} + \frac{cx^4}{4} + \frac{dx^3}{3} + \frac{ex^2}{2} \right]_{-1}^{1} = \frac{2b}{5} + \frac{2d}{3} = 0$

$\Rightarrow \frac{b}{5} + \frac{d}{3} = 0 \quad ii$

$\int x^2 f(x)\,dx = 0 \Rightarrow \left[\frac{ax^7}{7} + \frac{bx^6}{6} + \frac{cx^5}{5} + \frac{dx^4}{4} + \frac{ex^3}{3} \right]_{-1}^{1} = \frac{2a}{7} + \frac{2c}{5} + \frac{2e}{3} = 0$

$\Rightarrow \frac{a}{7} + \frac{c}{5} + \frac{e}{3} = 0 \quad iii$

$\int x^3 f(x)\,dx = 0 \Rightarrow \left[\frac{ax^8}{8} + \frac{bx^7}{7} + \frac{cx^6}{6} + \frac{dx^5}{5} + \frac{ex^4}{4} \right]_{-1}^{1} = \frac{2b}{7} + \frac{2d}{5} = 0$

$\Rightarrow \frac{b}{7} + \frac{d}{5} = 0 \quad iv$

$ii, iv \rightarrow \begin{cases} b = -\frac{5}{3}d \\ b = -\frac{7}{5}d \end{cases} \Rightarrow d = b = 0$

$i, iii \rightarrow e = -\frac{a}{5} - \frac{c}{3} \Rightarrow \frac{a}{7} + \frac{c}{5} - \frac{a}{15} - \frac{c}{9} = 0$

ادامه

$\Rightarrow \frac{8a}{105} + \frac{4c}{45} = 0 \Rightarrow a = \frac{-\cancel{4} \times \cancel{105}^{\,7} c}{\cancel{8}_{2} \times \cancel{45}_{3}} = \frac{-7}{6} c \Rightarrow \begin{matrix} a = 7k \\ c = -6k \end{matrix} \xrightarrow{k = 5r} \begin{matrix} a = 35r \\ c = -30r \end{matrix}$

$e = -\left(\frac{a}{5} + \frac{c}{3}\right) = -(-3r) = 3r$

$\Rightarrow f_r(x) = 35rx^4 - 30rx^2 + 3r$

برای اینکه در بازه $[-1,1]$ باشد ، چون

$\forall x \in [-1,1]: \quad |f_r(x)| = |35rx^4 - 30rx^2 + 3r| \leq$

$|35r||x^4| + |30r||x^2| + |3r| \underset{|x|\leq 1}{\leq} |35r| + |30r| + |3r|$

$\overset{r>0}{=} (35+30+3)r = 68r$

$r = \frac{1}{68}$ قرار می دهیم

$\Rightarrow f(x) = \frac{35}{68}x^4 - \frac{30}{68}x^2 + \frac{3}{68}$

سوال سوم

.۱

ابتدا حالت ماتریسی می کنم چون تمیز تره :)

$$J(\theta) = \|y - X\theta\|_2^2$$

گام نیوتن : $\theta_{n+1} = \theta_n - H_J^{-1}(\theta_n) \nabla J(\theta_n)$

ماتریس هسیان

چون که $x \leftarrow x - \frac{f'(x)}{f''(x)}$

$$J(\theta) = (y - X\theta)^T (y - X\theta) = (y^T - \theta^T X^T)(y - X\theta)$$

$$= y^T y - y^T X\theta - \theta^T X^T y + \theta^T X^T X\theta$$

$$\nabla J(\theta) = 0 - X^T y - X^T y + 2X^T X\theta = 2(X^T X\theta - X^T y)$$

$$H_J(\theta) = 2X^T X - 0 = 2X^T X$$

اجرا الگوریتم :

$$\theta_0 = \text{initialized } \theta$$

$$\Rightarrow \theta_1 = \theta_0 - (2X^T X)^{-1} 2(X^T X\theta_0 - X^T y) =$$

$$\theta_0 - \underbrace{(X^T X)^{-1}(X^T X)}_{I}\theta_0 + (X^T X)^{-1} X^T y = \theta_0 - \theta_0 + (X^T X)^{-1} X^T y$$

$$= (X^T X)^{-1} X^T y$$

که این همان جواب بهینه است که با گرادیان هم بدست می آمد.

یعنی با 1 گام نیوتون به جواب بهینه می رسیم.

.۲

wlog داریم که $x^i \in \mathbb{R}^{n+1}$ یک 1 اضافه کردم , $\theta = [\theta \; \theta_0]$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} y_i \log h_\theta(x^i) + (1 - y_i)\log(1 - h_\theta(x^i))$$

$$h_\theta(x^i) = \sigma(\theta^T x^i)$$

$$* \quad \sigma'(x) = \frac{d}{dx}\frac{1}{1+e^{-x}} = \frac{1}{(1+e^{-x})^2}e^{-x} = \left(\frac{1}{1+e^{-x}}\right)\left(1-\frac{1}{1+e^{-x}}\right) = \sigma(x)(1-\sigma(x))$$

$$\nabla J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{y_i}{\sigma(\theta^T x^i)} + \frac{(1-y_i)(-1)}{1-\sigma(\theta^T x^i)}\right)\frac{\partial}{\partial\theta}\sigma(\theta^T x^i)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{y_i}{\sigma(\theta^T x^i)} - \frac{(1-y_i)}{(1-\sigma(\theta^T x^i))}\right)\sigma(\theta^T x^i)\,(1-\sigma(\theta^T x^i))\,x^i$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(y_i(1-h_\theta(x^i)) - (1-y_i)\,h_\theta(x^i)\right)x^i$$

$$= \frac{1}{m}\sum_{i=1}^{m}(y_i - h_\theta(x^i))\,x^i$$

مشابه

$$H_J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}(\nabla h_\theta(x^i))\,x^i = \frac{-1}{m}\sum_{i=1}^{m}h_\theta(x^i)(1-h_\theta(x^i))(x^i)(x^i)^T$$

$$= \frac{-1}{m}\sum_{i=1}^{m}(h_\theta(x_i) - h_\theta(x_i)^2)\,x_i\,x_i^T$$

قرار می‌دهیم y بردار y_iها. h_θ بردار $h_\theta(x^i)$ها، X ماتریس x_iها

$$\nabla J(\theta) = X^T(y - h_\theta)$$

$$\sum\begin{pmatrix} x^1_1 & x^1_2 & \cdots & x^1_p \\ x^2_1 & \cdots & & x^2_p \end{pmatrix}\begin{bmatrix} y_1 - h_\theta(x^1) \\ \vdots \end{bmatrix} \Rightarrow X^T(y - h_\theta)$$

$$H_J(\theta) = X^T\Omega X \qquad \Omega = \begin{bmatrix} h_\theta(x^1)(1-h_\theta(x^1)) & 0 & 0 & 0 & 0 \\ 0 & h_\theta(x^2)(1-h_\theta(x^2)) & 0 & 0 & 0 \\ 0 & & \ddots & & \\ 0 & & & & \cdot \end{bmatrix}$$

قطری $h(1-h)$ها.

گام نیوتن:

$$\theta \leftarrow \theta - (H_J(\theta))^{-1} \nabla J(\theta)$$

تا زمان کانورج کردن

$$\Rightarrow \theta \leftarrow \theta - (X^T \Omega X)^{-1} X^T (y - h_\theta)$$

سوال چهارم

۱. بله یا خیر ؟ ← برای هر دانشجو یک توزیع برنولی داریم. پارامتر توزیع را θ می‌نامیم. می‌دانیم iid اند پس همگی پارامتر θ اند و مستقل

$$l = \prod_i p(x_i) = \prod_{i=1}^{n} \hat{\theta}^{x_i} (1-\hat{\theta})^{1-x_i} =$$

x_i بله یا خیر شدن (۰ یا ۱) است.

$\hat{\theta}$ تخمین θ.

$$= \hat{\theta}^{\sum_{i=1}^{n} x_i} (1-\hat{\theta})^{\sum_{i=1}^{n}(1-x_i)}$$

$$\underset{\text{log likelihood}}{ll} = \log\left(\hat{\theta}^{\sum_{i=1}^{n} x_i} (1-\hat{\theta})^{\sum_{i=1}^{n}(1-x_i)}\right) = \sum_{i=1}^{n} x_i \log\hat{\theta} + \sum_{i=1}^{n}(1-x_i)\log(1-\hat{\theta})$$

$$\nabla ll = 0 \underset{\log' x = \frac{1}{x}}{\Rightarrow} \frac{1}{\hat{\theta}} \sum_{i=1}^{n} x_i + \frac{-1}{1-\hat{\theta}} \sum_{i=1}^{n} (1-x_i) = 0$$

$$\Rightarrow \frac{\sum x_i}{\hat{\theta}} = \frac{\sum(1-x_i)}{1-\hat{\theta}} = \frac{n - \sum x_i}{1-\hat{\theta}} \xrightarrow{\text{طرفین وسطین}}$$

$$n\hat{\theta} - \underbrace{\hat{\theta}\sum x_i} = \sum x_i - \underbrace{\hat{\theta}\sum x_i} \Rightarrow n\hat{\theta} = \sum x_i \Rightarrow \hat{\theta} = \frac{\sum x_i}{n}$$

μ برابر درصد بله‌ها است. $\hat{\theta}$ تخمین نسبت کل بله ها است. $x \sim \text{Bernoulli}(\theta)$

$$E(X) = 1 \times \theta + 0 \times (1-\theta) = \theta$$

$$\Rightarrow \underset{\text{تخمین درصد} \leftarrow}{\hat{\mu}} = 100 \times \hat{\theta} = \frac{100 \times \sum_{i=1}^{n} x_i}{n} = \frac{100}{n} \times \#(X_i = 1)$$

پس از پرسش از n نفر، تعداد بله‌ها به کل ضرب در ۱۰۰ برابر تخمین μ که درصد علاقه‌مندان است می‌شود.

۲. تخمین اول

برای $\hat{\theta}$ داریم : $0 \le \theta \le 1$

طبق Hoeffding's Inequality :

$$P(|\hat{\theta}-\theta| > \epsilon) \le 2e^{\frac{-2n^2\epsilon^2}{\underbrace{\sum (b_i-a_i)^2}_{\text{سر و ته بازه}}}} \rightarrow n\times 1$$

$$\Rightarrow P(|\hat{\theta}-\theta| > \epsilon) \le 2e^{\frac{-2n\epsilon^2}{1}} = 2e^{-2n\epsilon^2}$$

۱- می‌خواهیم $P(|\hat{\theta}-\theta| > 0.1) < 0.05$ پس کافی است

$$2e^{-2n(0.1)^2} < 0.05 \Rightarrow \frac{5}{200} > e^{\frac{-2n}{100}} \Rightarrow \ln\left(\frac{1}{40}\right) > \frac{-2n}{100}$$

$$\Rightarrow n > -50\ln\left(\frac{1}{40}\right) \approx 184.4 \Rightarrow$$ حداقل از 185 نفر بپرسیم

(این رو نوشتم سر محاسبه‌ی غلط سوال ولی حل همین بالاییه (و معادل صفحه‌ی بعدیش ۱ ای)

۲- اگر منظور از μ همان درصد باشد : $0 \le \hat{\mu}, \mu \le 100$

$$P(|\hat{\mu}-\mu| > 0.1) \le 2e^{\frac{-2n(0.1)^2}{(100)^2}}$$

$$\Rightarrow 0.05 > 2e^{-2n\times 10^{-6}} \Rightarrow \frac{1}{40} > e^{-2\times 10^{-6}\times n}$$

$$\Rightarrow \ln\left(\frac{1}{40}\right) > -2\times 10^{-6}\times n \Rightarrow n > -\frac{10^6}{2}\ln\left(\frac{1}{40}\right) \approx 1844439.72$$

$\Rightarrow$ n حداقل باید 1844440 باشد.

تعیین دوم: داریم $P(|\hat{\mu}-\mu| \leq 0.1) \geq 0.95$

پس به احتمال 95% نامساوی رخ دهد یا به عبارت دیگر بازه اطمینان با 95% تعیین را می‌خواهیم.

برای هر شخص $E(X_i) = \theta$, $Var(X_i) = \theta(1-\theta)$ چون برنولی اند -

انحراف معیار

حال σ_{max} را می‌یابیم. $\max_{\theta} \theta(1-\theta) \Rightarrow \frac{d}{d\theta}\theta(1-\theta)=0$

$\Rightarrow 1-2\theta=0 \Rightarrow \theta=\frac{1}{2}$

$\Rightarrow \sigma_{max} = \sqrt{\theta(1-\theta)} = \sqrt{\frac{1}{2}\cdot\frac{1}{2}} = \frac{1}{2}$

طبق تعیین بازه داریم که $P(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$

هنگامی که σ_{max} داریم در احتمال بالای قرار با σ جایگزین کرد.

در نتیجه بازه اطمینان $1-\alpha$ در مثال ما برابر است با $\left[\hat{\theta} - Z_{\frac{\alpha}{2}}\frac{\overset{\sigma_{max}}{\frac{1}{2}}}{\sqrt{n}}, \hat{\theta} + Z_{\frac{\alpha}{2}}\frac{\frac{1}{2}}{\sqrt{n}}\right]$

(برای n مقدار بزرگ)

پس چون $1-\alpha=0.95 \Rightarrow \alpha=0.05 \Rightarrow \frac{\alpha}{2}=0.025 \Rightarrow Z_{\frac{\alpha}{2}}=1.96$

$$\Rightarrow P\left(|\hat{\theta}-\theta| \leq \frac{1.96}{2\sqrt{n}}\right) \geq 0.95$$

حال پس لازم است $\frac{1.96}{2\sqrt{n}} \leq 0.1$ باشد که این نامساوی برایش محقق شود.

$\Rightarrow \sqrt{n} \geq \frac{19.6}{2} \Rightarrow n \geq 96.04 \Rightarrow$ n حداقل 97 است.

☆ اگر مقصود درصد بود داریم که $\hat{\mu}$ و μ برابر $100\hat{\theta}$ و 100θ اند.

$\Rightarrow$ به جای 0.1 داشتیم $\frac{1}{1000}$ $\Rightarrow$ $10^{-3} \geq \frac{1.96}{2\sqrt{n}} \Rightarrow n \geq \left(\frac{1960}{2}\right)^2$

$\Rightarrow n \geq 960400 \Rightarrow$ n حداقل 960400 می‌شود.

سوال پنجم

(پوزش: در جاهای $\preceq$ و $\succeq$ شبیه $\leq$ و $\geq$ شده)

نکته: داریم که $A(I-A) \preceq B(I-B)$ بنابراین

$$B(I-B)-A(I-A) = B-B^2-A+A^2$$

مثبت نیمه معین است. ← این ویژگی را در حل psd می‌نامم.

ادعا: $(B-A)(\frac{I}{2}-B+\frac{I}{2}-A)$ psd است.

اثبات: کافی است برای هر بردار v، $v^T (\;) v \geq 0$ باشد. پس فرض خلف کنید که v وجود دارد که منفی شود.

$$v^T(B-A)\left(\frac{I}{2}-B+\frac{I}{2}-A\right)v<0$$

$$\parallel$$

$$\Rightarrow v^T\left(\frac{1}{2}B-B^2+\frac{1}{2}B-BA-\frac{1}{2}A+AB-\frac{1}{2}A+A^2\right)v<0$$

$$\parallel$$

$$v^T(B-B^2-A+A^2-BA+AB)v<0$$

صحت نکته ۱: پس $v^T(B-B^2-A+A^2)v \geq 0$

$$\underbrace{v^T(B-B^2-A+A^2)v}_{\geq 0}+v^T(AB-BA)v<0$$

پس $v^T(AB-BA)v<0$

اما این نیز ممکن است چرا که میدانیم A و B متقارن اند پس $A^T=A,\ B^T=B$

$$\Rightarrow\quad v^TABv = \langle v, ABv\rangle = \langle ABv, v\rangle = (ABv)^Tv = v^TB^TA^Tv = v^TBAv$$

پس $v^T(AB-BA)v=0$ است و فرض کردیم منفی باشد. $\times$. پس ادعا ثابت شد.

حال ادعای دوم: مقادیر ویژه های $B-A$ نامنفی اند. مقدار ویژه λ از آن را در نظر بگیرید با بردار ویژه v.

داریم که X psd باشد X^T هم psd است چرا که

$$\underbrace{v^TXv}_{\text{عدد}}\ge 0 \Rightarrow (v^TXv)^T = v^TX^Tv$$

پس $(B-A)(\frac{I}{2}-B+\frac{I}{2}-A) \xrightarrow{T} (\frac{I}{2}-B^T+\frac{I}{2}-A^T)(B^T-A^T)$ psd

A,B متقارن $\longrightarrow = (\frac{I}{2}-B+\frac{I}{2}-A)(B-A)$ psd است.

$$\to v^T(\frac{I}{2}-B+\frac{I}{2}-A)\underbrace{(B-A)v}_{=\lambda v}\ge 0 \quad\Rightarrow\quad v^T(\frac{I}{2}-B+\frac{I}{2}-A)\lambda v\ge 0$$

فرض خلف که λ منفی باشد پس باید:

$$v^T(\frac{I}{2}-B+\frac{I}{2}-A)v\le 0$$

$$\| $$

$$\underbrace{v^T(\frac{I}{2}-B)v}_{\ge 0}+\underbrace{v^T(\frac{I}{2}-A)v}_{\ge 0}$$

اما $\frac{I}{2}-A$ و $\frac{I}{2}-B$ psd اند پس

در نتیجه فقط حالت تساوی می‌تواند رخ دهد:

$$\Rightarrow\quad v^T(\frac{I}{2}-B)v=0=v^T(\frac{I}{2}-A)v$$

$0 = v^T(\frac{I}{2}-A)v - v^T(\frac{I}{2}-B)v = v^T\underbrace{(B-A)v}_{=\lambda v} = v^T\lambda v$ از هم کم می کنیم پس

فرض کردیم

$0 = v^T v = \|v\|^2$ $\Leftarrow v^T v = 0$ $\Leftarrow$ پس ناصفره است $\lambda \neq 0$

$\Rightarrow v = 0$ $\rightsquigarrow$ اما بردار ویژه ها ناصفرند. ✗

پس تناقض فرض کردیم یعنی $\lambda \geq 0$ است، پس مقدار ویژه ها نامنفی اند.

حال ادعا می کنم ماتریس X حقیقی متقارن با مقدار ویژه ها نامنفی، psd است.

اثبات: چون X متقارن است ← اگه حقیقی نبود جای T باید $*$ می بود.

طبق spectral decomposition می توان X را بصورت UDU^T نمایش داد D قطری نامنفی و

U unitary.

ماتریس قطری

حال چون قطر D نامنفیه برای $D^{\frac{1}{2}}$ تعریف کنیم که جذر قطرهایش است.

$Y = D^{\frac{1}{2}}U^T \Rightarrow Y^T Y = (D^{\frac{1}{2}}U^T)^T(D^{\frac{1}{2}}U^T) = U(D^{\frac{1}{2}})^T D^{\frac{1}{2}}U^T = UDU^T = X$

پس X را Y^*Y نوشتیم.

حال داریم که

هر بردار $v^T X v = v^T Y^T Y v = (Yv)^T(Yv) = \|Yv\|^2 \geq 0$

پس psd است.

پس $B-A$ psd است یعنی $A \preceq B$