HW1

Seyed Hamid Azimidokht 9933213

سوال ۱

الف

درست است. به طور غیر مستقیم باعث این اتفاق میشود.
منتظم سازی باعث میشود که مدل در پی انتخاب مقادیر کوچک تر برای وزن ها برود و بنابراین مدل را ساده تر (به این معنی که مینیمم های محلی کمتر میشوند) میکند که این باعث میشود تابع هزینه اسموت(نرم) تر شود و احتمال گیر کردن در مینیمم محلی کمتر شود.

ب

درست است.

در حالتی که وای در فضای ویژگی ها قرار دارد سیستم معادلات متشکل جواب دارد و بحثی نیست

$$X\theta = t$$

ولی در حالتی که در فضای ویژکی ها قرار ندارد در واقع در پی حل یک مسئله بهینه سازی هستیم و میخواهیم تتایی را پیدا کنیم که مقدار خطای جمع مربعات را کمینه کند:

$$\arg\min_{\theta} ||t - X\theta||_2$$

و طبق قضیه که در کلاس مطرح و اثبات شد، نزدیک ترین بردار در یک فضا به یک بردار که در آن فضای نیست تصویر آن برداری که در آن فضا نیست در آن فضاست.
پس گذاره ذکر شده درست است.

ج

درست است.

چون مدل خیلی ساده است نمیتواند به خوبی بر روی داده فیت شود و الگو های پیچیده را یاد بگیرد و برای داده های جدید هم نمیتواند به خوبی عمومی سازی (جنرالایز) شود

البته اگر داده ها به نحوی باشد که آن مدل ساده بتواند بخوبی آن را مدل کند انگاه (زیربرازش رخ نمیدهد)

د

.درست است

توجه کنید که اگر چنین باشد ممکن است مدلی که با مجموعه ارزیابی انتخاب میشود روی این مجموعه عملکرد خوبی داشته باشد ولی چون توزیع متفاوتی با آزمایش دارد روی .مجموعه آزمایش خیلی بد عمل کند

ولی در مواردی ممکن است قصد ما این باشد که مدلی که آموزش می دهیم بر روی .داده های با توزیع جدید بررسی شود در چنین مواردی میتواند باشد

ه

نادرست است

اگر داده های ما باالانس نباشند انگاه روش نمونه برداری تصادفی برای مجموعه ترین باعث میشود که مدل روی داده های مربوط به کلاس با داده زیاد تر یادگیری را انجام دهد و کلاس با داده کم را بدرستی تشخیص ندهد و همچنین به دلیل تصادفی بودن همیشه این احتمال وجود دارد که توزیع داده ها برای مجموعه ها مختلف باشد که مشکلاتی بوجود می آورد

یا در حالت دیگری اگر داده های ما داده های سری های زمانی باشد روش تصادفی .کفایت نمیکند

.پس همواره کفایت نمیکند

و

.نادرست

تابع ام ای ای در نقاطی که مقدار تابع و پیشبینی برابر اند مشتق بپذیر نیست ولی .برای ام اس ای اینطور نیست

ز

نادرست است.

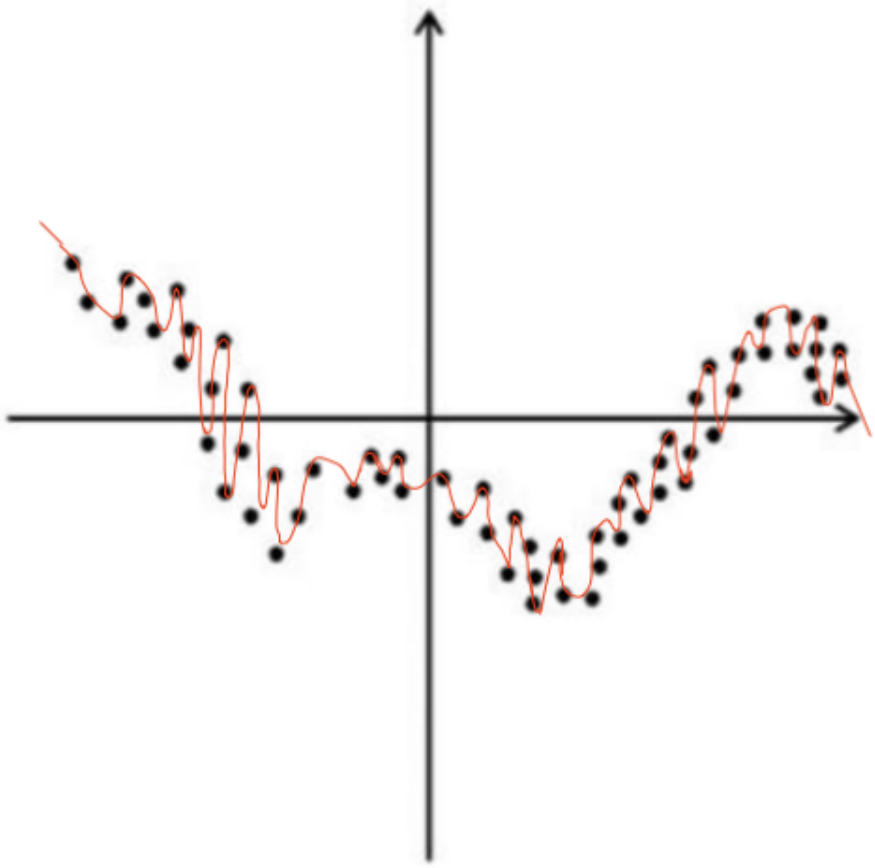
اگر منظور از گزارش این است که مدل را برای داده تست انتخاب کنیم:

بعد از انتخاب پیچیدگی مدل آن مدل را روی تمامی داده آموزش داده و سپس مدل بدست آمده را برای گزارش نهایی استفاده میکنیم.

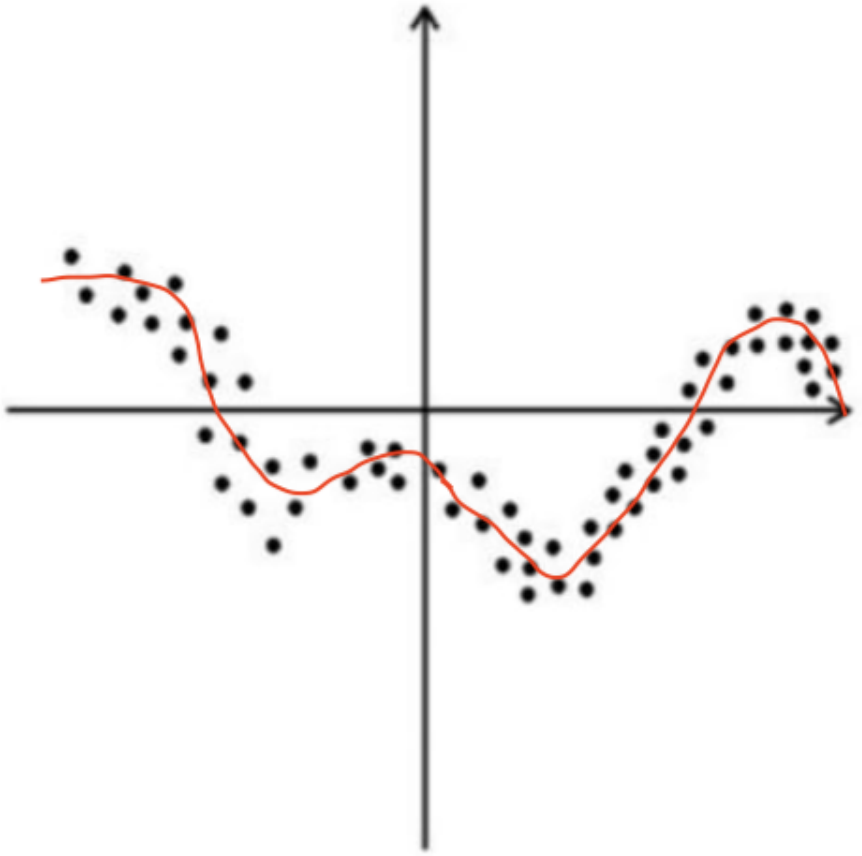
اما اگر منظور از گزارش این است که عملکرد ارزیابی کنیم:

باید میانگین نتایج حاصل از همه قسمت ها یا فولد ها رو گزارش کنیم.

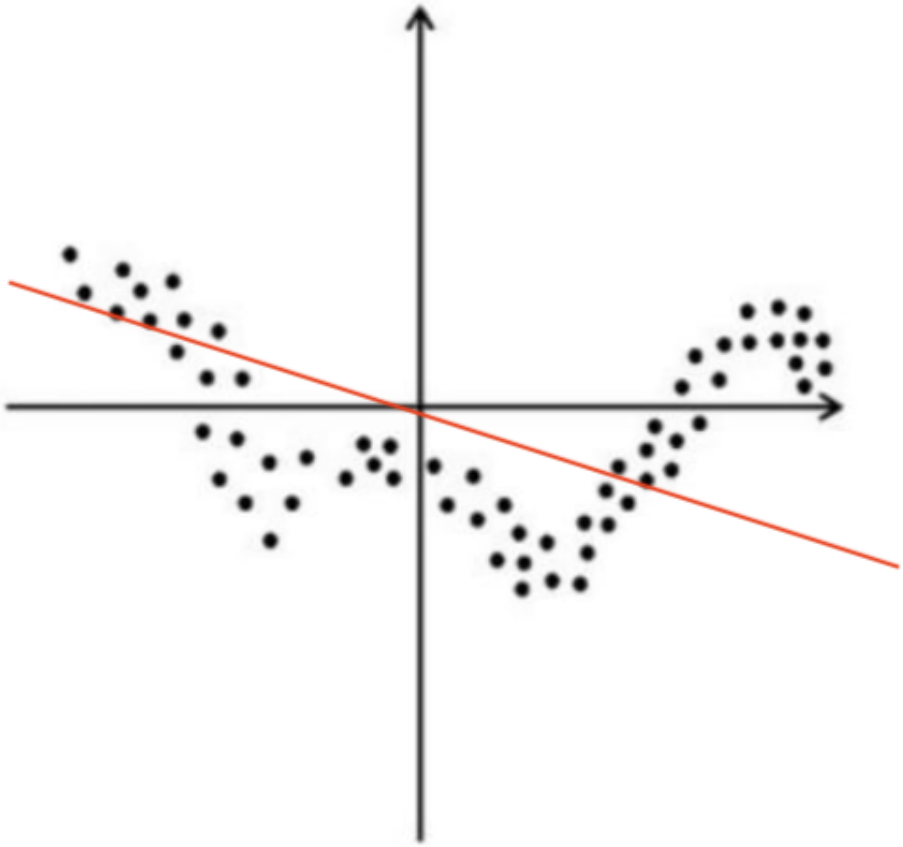
سوال ۲



Ovefit



Fit



Undefit

سوال ۳

ستون ای مربوط به مقدار منفی بینهایت برای لگاریتم طبیعی لاندا است. مقدار آن خیلی کوچک می شود و عملا انگار ترم منتظم سازی در تابع هزینه نیست. پس مدل راحت تر میتواند مقادیر زیاد برای وزن ها در نظر بگیرد بدون جریمه ای که در صورت وجود ترم منتظم سازی وجود داشت.

ستون بی مربوط به مقدار ۱۸- برای لگاریتم طبیعی لاندا است. زیرا مقدار آن کوچک میشود ولی نسبت به حالت قبل که خیلی نزدیک به صفر بود بیشتر است و تاثیر ترم منتظم سازی بیشتر خواهد بود پس مدل برای اتخاذ مقادیر بزرگ برای وزن ها نسبت به

حالتی که ترم منتظم سازی وجود نداشت محدود تر است و جریمه می شود. پس مقادیرش کوچک تر خواهند بود .

ستون سی مربوط به مقدار ۰ برای لگاریت طبیعی لاندا است زیرا مقدار لاندا برابر با ۱ خواهد بود و لذا ترم منتظم سازی نسبت به حالت های قبل بیشتر تاثیر میگذارد و باعث میشود وزن ها کوچک تر شوند.

سوال ۴

$$L(\theta) = \prod_{n=1}^{N} p(x^n|\theta) = \prod_{n=1}^{N} \frac{1}{\theta} = (\frac{1}{\theta})^n$$

$$\log L(\theta) = n \log(\frac{1}{\theta}) = -n \log \theta$$

توجه داریم که چون داده ها از این توزیع انتخاب شده اند لذا باید برای همه مقادیر آی داشته باشیم:

$$\theta > x_i$$

از طرفی تابع بدست آمده برای لوگ لایکی هود یه تابع نزولی است. پس برای اینکه این مقدار را ماکسیمم کنیم باید مقدار تتا را تا حد امکان کوچک در نظر بگیریم.
پس با توجه به دو شرطی که داریم باید تتا برابر با بزرگ ترین داده ما باشد یعنی:

$$\theta = \max(x_i)$$

for i=1,…,n

سوال ۵

$$\arg\min_{w} -\log p(w|X,y) \tag{1}$$
$$= \arg\min_{w} -\log \frac{p(X,y|w)p(w)}{p(X,y)} \tag{2}$$
$$= \arg\min_{w} -\log p(y|X,w)p(w) \tag{3}$$
$$= \arg\min_{w} -\log \left[p(w) \prod_{n=1}^{N} p(y_n|X_n,w) \right] \tag{4}$$
$$= \arg\min_{w} -\log \left[\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \prod_{n=1}^{N} p(y_n|X_n,w) \right] \tag{5}$$
$$= \arg\min_{w} -\log \Big[\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \Big] - \log \Big[\prod_{n=1}^{N} p(y_n|X_n,w) \Big] \tag{6}$$
$$= \arg\min_{w} -\log \Big[p^{\alpha-1}(1-p)^{\beta-1} \Big] + \log B(\alpha,\beta) - \log \Big[\prod_{n=1}^{N} p^{x}(1-p)^{1-x} \Big] \tag{7}$$

$$\frac{d}{dp} \left[-\log \Big[p^{\alpha-1}(1-p)^{\beta-1} \Big] + \log B(\alpha,\beta) - \log \Big[\prod_{n=1}^{N} p^{x_n}(1-p)^{1-x_n} \Big] \right] \tag{8}$$
$$= \frac{d}{dp} \left[-\log \Big[p^{\alpha-1}(1-p)^{\beta-1} \Big] - \log \Big[\prod_{n=1}^{N} p^{x_n}(1-p)^{1-x_n} \Big] \right] \tag{9}$$
$$= \frac{d}{dp} \left[(1-\alpha)\log \Big[p \Big] + (1-\beta)\log \Big[(1-p) \Big] - \sum_{n=1}^{N} x^n \log p + (1-x^n) log(1-p) \right] \tag{10}$$
$$= \frac{1-\alpha}{p} + \frac{\beta-1}{1-p} - \frac{1}{p}\sum_{n=1}^{N} x^n - \frac{1}{1-p}\sum_{n=1}^{N} x^n - 1 = 0 \tag{11}$$

$$\implies p = \frac{1-\alpha-\sum_{n=1}^{N} x^n}{2-(\alpha+\beta)+N}$$

سوال ۶

$$LL(\theta) = \sum_{n=1}^{N} \log P(x^i|\theta) = 5\log(\theta) + 5\log(1-\theta) - 16\log(2) + 3\log(3)$$

$$\frac{dLL(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$$

$$\implies \theta = 0.5$$

سوال ۷

ستون بی مربوط به تابع هزینه لاسو است و ستون ای مربوط به تابع هزینه ریج

دلیل: در تابع هزینه ریج به دلیل اینکه وزن ها به توان دو رسیده اند، وقتی که مقدار وزن ها کوچک میشود با به توان دو رسیدن و ضرب شدن در خودش عددی خیلی کوچک تولید میشود که تاثیر خیلی زیادی در مقدار تابع هزینه نخواهد داشت (نسبت به لاسو) برای همین مقادیر وزن ها (نسبت به لاسو) مقدار صفر نمیشود و عددی کوچک و نزدیک به صفر خواهد بود.
همانطور که مشاهده می شود در ستون بی مقدار صفر بیشتری برای وزن ها موجود است که نشان دهنده لاسو بودن تابع هزینه برای آن است.

برتری:

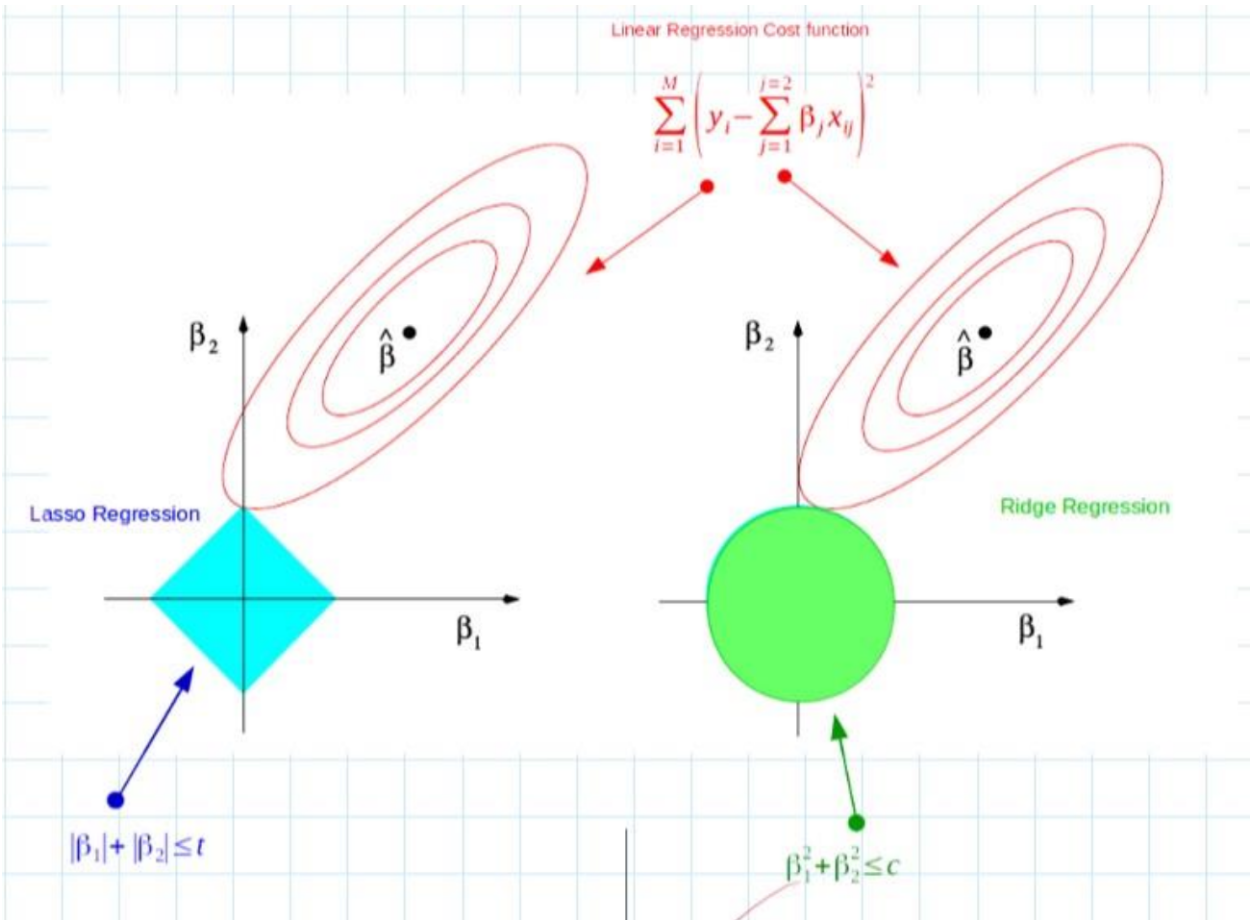
در لاسو به دلیل اینکه توانایی کوچک کردن وزن ها تا صفر شدن آنها را دارد برای فیچر سلکشن بهتر است.

ولی ریج برای دیتاست هایی که فیچر ها خیلی وابسته هستند مناسب تر است. زیرا مقدار وزن هارو صفر نمیکند تا تاثیر فیچر از بین برود.

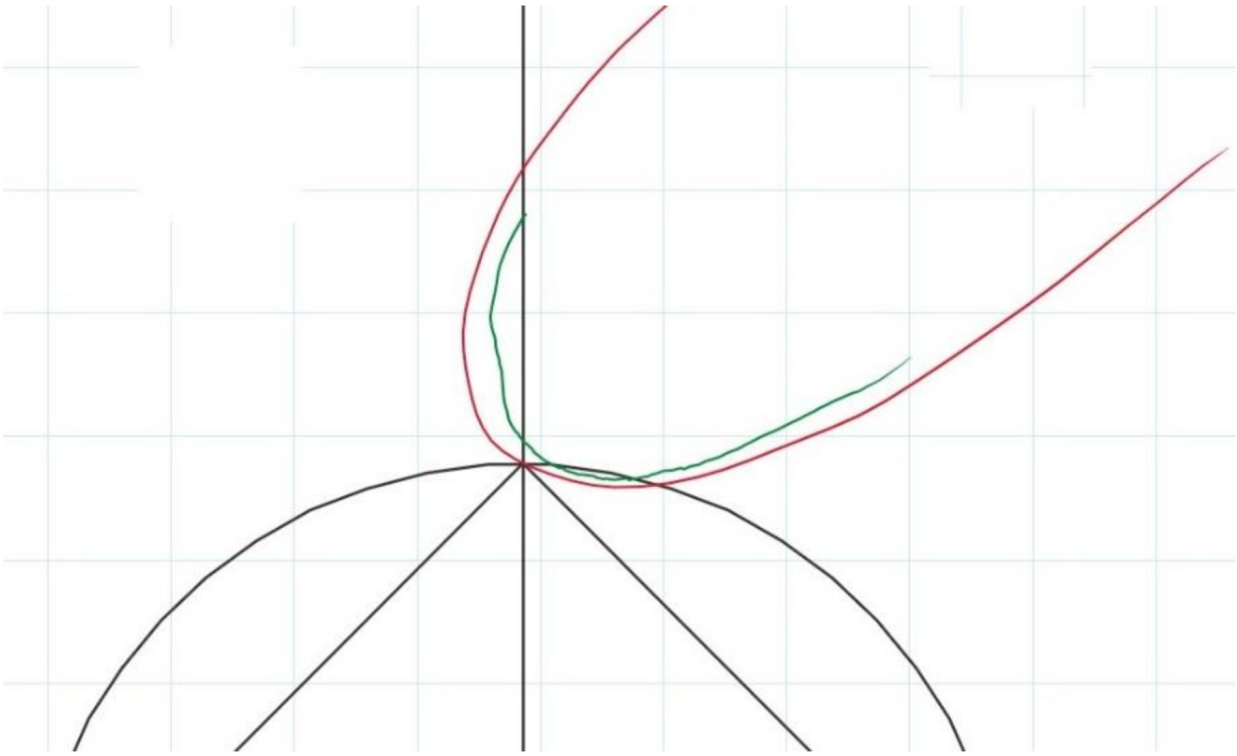
رسم نمودار و توضیح

اگر لاسو و ریج را در نظر بگیریم و برای دو بعد فرض کنیم که:

$$\beta_2 = w_5, \beta_1 = w_2$$



:با بررسی دقیق تر خواهیم داشت



با فرض اینکه کانتور سبز رسم شده بر دایره مماس است این توضیحات را خواهیم داشت:

توجه داریم که با توجه به نحوه قرار گیری نقطه بهینه کانتور، نقطه بالای لوزی که متناظر با شرط لاسو است به آن نزدیک تر است که در آن یکی از ضرایب صفر شده است.

پس لاسو این نقطه را انتخاب میکند (دو شرط نزدیکی به بهینه و شرط منتظم سازی لاسو)

حال اگر نمودار مربوط به شرط ریج را در نظر بگیریم میبینیم که نقطه ای که دو شرط نزدیکی به بهینه و شرط منتظم سازی ریج را بر آورده کند نقطه ای نیست که مقدار وزنی را صفر کند.

این توضیحات بالا، توضیحات استاد سر کلاس درس بود.
توضیحی که در ادامه می آورم به نحوه بهتری این مطلب را نشان میدهد که ریج وزن هارا صفر نمیکند:

با درنظر گرفتن یکی از وزن ها مثل:

$$w = w_i$$

داریم:

$$L_2 = (y - xw)^2 + \lambda \sum_{i=1}^{N} w_i^2$$

این تابع کاست مربوط به ریج است. قصد ما کمینه کردن این تابع است بنابر این آن را نسبت به وزنی که در نظر داریم مشتق میگیریم:

$$-2xy + 2x^2w + 2w\lambda = 0$$
$$\implies w = \frac{xy}{x^2 + \lambda}$$

مشاهده میشود که اگر وزن بخواهد صفر شود باید مقدار لاندا بینهایت باشد ولی مقدار لاندا توسط ما ست میشود و هیچ گاه نمیتواند بینهایت باشد.
بنابر این وزن صفر نمیشود.
در حالی که اگر همین کار را برای لاسو انجام دهیم خواهیم داشت:

$$w = \frac{2xy - \lambda}{2x^2}$$

که صورت کسر میتواند صفر شود و وزن مقدار صفر بگیرد.

سوال ۸

من برای این سوال کدش رو زدم و همراه این پی دی اف ارسال کردم.
لطفا اون رو بررسی کنید.

الف

$$MAE = \frac{1}{m} \sum_{i=1}^{N} |pred - y| = 18.945000000000007$$

$$MSE = \frac{1}{2m} \sum_{i=1}^{N} (pred - y)^2 = 182.8028750000001$$

ب

$$L^i(W) = (\hat{Y}^i - Y^i)^2$$

$$\nabla L^i(W) = 2(\hat{Y}^i - Y^i)X^i$$

$$W_{i+1} = W_i - lr \times \nabla L^i(W)$$

$$\nabla L^0(W) = [-41.6, -1708.88, -5751.84]$$

$$W_1 = [-55.332, 170.738, 575.784]$$

$$\nabla L^1(W) = [190326.552, 7993715.18400001, 29119962.45600002]$$

$$W_2 = [-19087.9872, -799200.7804, -2911420.4616]$$

$$\nabla L^2(W) = [-9.38428102e+08, -3.47218398e+10, -1.41702643e+11]$$

$$W_3 = [9.38237222e+07, 3.47138478e+09, 1.41673529e+10]$$

$$\nabla L^3(W) = [4.08807092e+12, 1.88051262e+14, 5.43713433e+14]$$

$$W_4 = [-4.08713269e+11, -1.88016549e+13, -5.43571759e+13]$$

$$MSE = 3.702183927607888e+31$$

ج

$$\nabla L(W) = \frac{1}{m} X^T(Y - \hat{Y}) = [-18.945, -781.7125, -2737.17]$$

$$W_{i+1} = W_i - lr \times \nabla L(W)$$

$$W_1 = [-57.6055, 78.02125, 274.317$$

$$MSE = 908587593.425293$$

د

.مقدار خطا در روش تصادفی نسبت به روش نزول گرادیان خیلی بیشتر است
و وزن های حاصل از روش از نظر اندازه خیلی بزرگ تر از روش نزول گرادیان معمولی
اند.

دلیل این امر این است که در تصادفی چهار بار اپدیت انجام شده ولی در روش معمولی
تنهای یکبار

.توجه داریم که در هر دوحالت به دلیل بزرگ بودن مقدار لرنینگ ریت الگورتم واگرا میشود