

یادگیری عمیق

پاییز ۱۴۰۲

استاد: دکتر فاطمیزاده

- گردآورندگان:



دانشگاه صنعتی شریف
دانشکده‌ی مهندسی برق

شبکه های پرسپترون، رگولاسیون، بهینه سازها

تمرين دوم

- مهلت ارسال پاسخ تا ساعت ۵:۵۹ ۲۳ روز مشخص شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همهی تمارین تا سقف ۵ روز و در مجموع ۲۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال شده پذیرفته نخواهد بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- هم‌کاری و همفکری شما در انجام تمرین مانع ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
(دقت کنید در صورت تشخیص مشابهت غیرعادی برخورد جدی صورت خواهد گرفت.)
- در صورت همفکری و یا استفاده از هر منابع خارج درسی، نام همفکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- نتایج و پاسخ‌های خود را در یک فایل با فرمت zip به نام Quera HW2-Name-StudentNumber در سایت قرار دهید. برای بخش عملی تمرین نیز لینک گیت‌هاب که تمرین و نتایج را در آن آپلود کرده‌اید قرار بدهید. دقต کنید هر سه فایل نوتبوک تکمیل شده بخش عملی را در گیت‌هاب قرار دهید.
- لطفا تمامی سوالات خود را از طریق کوئیرای درس مطرح بکنید (برای اینکه تمامی دانشجویان به پاسخ‌های مطرح شده به سوالات دسترسی داشته باشند و جلوی سوالات تکراری گرفته شود، به سوالات در بسترهای دیگر پاسخ داده نخواهد شد).
- دقت کنید کدهای شما باید قابلیت اجرای دوباره داشته باشند، در صورت دادن خطأ هنگام اجرای کدتان، حتی اگه خطأ بدلیل اشتباه تایپی باشد، نمره صفر به آن بخش تعلق خواهد گرفت.

سوالات نظری (۲۰۰ نمره)

۱. (۴۰ نمره)

(آ) نموداری رسم کنید که محور افقی آن تعداد داده در mini-batch و محور عمودی آن تعداد گام لازم برای همگرایی توسط الگوریتم SGD جهت رسیدن به مقدار از پیش تعیین شده ای از خطای آموزش باشد. بخش آغازین نمودار (بسته های کوچک داده) و بخش پایانی نمودار (بسته های بزرگ داده) را با ذکر دلیل توجیه نمایید.

(ب) آیا می‌توان گفت که در لایه Batch Normalization و در هنگام آموزش، مقداری نویز به توابع فعالیت لایه های مخفی تزریق می‌شود؟ چرا؟

(ج) شبکه‌ی عصبی Fully Connected ای را در نظر بگیرید که تمام توابع فعالیت آن سیگموید باشد و وزن های اولیه آن مقادیر مثبت بزرگ باشد. آیا این شبکه مناسبی برای طبقه بندی می باشد؟ چرا؟

(د) شبکه‌ی عصبی Fully Connected ای با ۵ لایه مخفی، که در هر کدام از لایه ها، ۱۰ نورون وجود دارد را در نظر بگیرید. ورودی این شبکه ۲۰ بعدی و خروجی آن اسکالر می باشد. تعداد کل پارامتر های قابل آموزش را در این شبکه محاسبه کنید.

(ه) یک مسئله طبقه بندی با نری می تواند با دو روش زیر حل شود:
روش اول: Logistic Regression ساده (یک نورون)

$$\hat{y} = \sigma(W_l x + b_l)$$

اگر $0.5 \leq \hat{y}$ کلاس صفر در غیر این صورت کلاس یک طبقه بندی می شود.

روش دوم: Softmax Regression ساده (دو نورون)

$$\hat{y} = softmax(W_s x + b_s) = [\hat{y}_1, \hat{y}_2]^T$$

اگر $\hat{y}_1 \geq \hat{y}_2$ کلاس صفر در غیر این صورت کلاس یک طبقه بندی می شود.

روش دوم دو برابر روش اول پارامتر دارد. آیا می توان گفت که روش دوم مدل های پیچیده تری

نسبت به روش اول یاد می گیرد؟

اگر بله، پارامترهای (W_s, b_s) تابعی که روش دوم می تواند آن را مدل کند مثال بزنید در غیر این صورت نشان دهید که (W_s, b_s) همیشه می تواند بر حسب (W_l, b_l) نوشته شود.

۲. (۳۰ نمره) می دانیم حتی یک شبکه عصبی یک لایه نیز می تواند طبقه بندی ارقام را با دقت خوبی انجام دهد. راه های متعددی برای ارتقای دقت مدل وجود دارد. این **مقاله** روشنی ساده برای ارتقای عملکرد مدل، بدون تغییر در ساختار آن پیشنهاد آموزش چند مدل مشابه است. مقاله را بخوانید و به سوالات زیر پاسخ دهید:

(آ) کمیته ۱ چیست و چطور به بهبود عملکرد مدل کمک می کند؟

(ب) پیش پردازش انجام شده در مقاله را شرح دهید. چگونه این پیش پردازش از وابستگی زیاد خطای مدل ها جلوگیری می کند؟

۳. (۳۰ نمره)

(آ) یک شبکه عصبی با ورودی x را در نظر بگیرید. برای بدست آوردن خروجی محاسبات زیر بر روی x انجام می شود.

$$z = wx + b$$

$$y = \sigma(z)$$

$$L = \frac{1}{2}(y - t)^2$$

$$R = \frac{1}{2}w^2$$

$$L_{reg} = L + \lambda R$$

گراف محاسباتی این مسئله را رسم کنید و مشتقات L_{reg} را نسبت به همه متغیرها بدست آورید.

(ب) پارامتر های یک شبکه عصبی در ابتدا به صورت تصادفی و با مقادیر کوچک مقداردهی می شوند. توضیح دهید در صورت عدم رعایت این دو ویژگی در مقداردهی چه مشکلاتی بروز پیدا می کند.

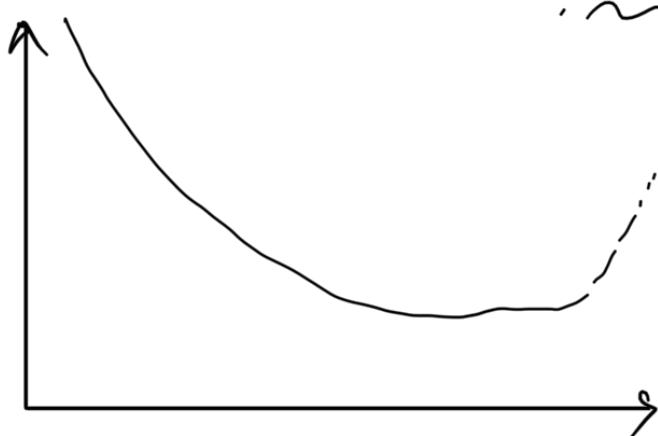
(ج) وزن های شبکه عصبی بدست آمده در قسمت اول را با مقادیر تصادفی دلخواه مقداردهی کنید و برای یک ورودی دلخواه، با توجه به مشتقاتی که در قسمت اول بدست آورید، با اعمال بهینه سازی گرادیان کاهشی برای یک ایپاک با نرخ یادگیری 0.1 ، وزن های شبکه را آپدیت کنید.

owell

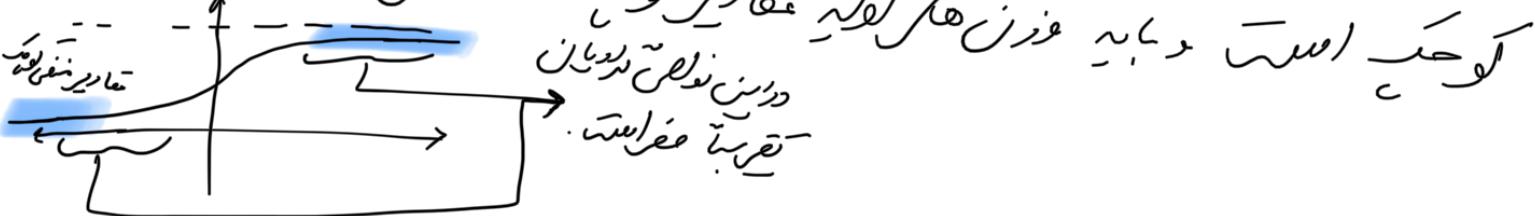
آ: بازگشتن اندیشه به درتعاریف کنیم، مقدار دارایی و تضییع را برایان چنین
نماییم که باید از بالغ فرض نظریه باشد، محدودیت سیم آرخیوی داشته باشد.
کوچک داراییان را بودجه روزگار میگذرانند تفسیس خواهد شدند.

در تعاریف بزرگتر به عملکرد ساده و فقر این عده اندیشه
Saddle points نزدیک میگیرد است بجزءی از تفاوت زیستی

کوچکی - به نیازهای اندیشه.



ب: بازی، حمله و درگیری زیستی از میان این دو اندیشه ملایم است
که کوچک اسماهه صریح و از آنجایی که به همراه محدودیت اندیشه محدود است
مقدار میان این دو اندیشه میگذرد که معمولاً تفاوت زیستی بودجه و دارایی روزگار میگذرد
ج: بازی، حمله و پیغامهای سیمی بین دو اندیشه بزرگ دیدار میگیرند
کوچکی کوچکی - حمله و پیغام دو اندیشه را متفق درین زمینی بیارکنند بودجه بین این
دو اندیشه برقراری وزن های خوبی آوریتی فرم کنند که در معادله انتقامیان بدل
در فرم زنده برقراری وزن های خوبی آوریتی فرم کنند که در معادله انتقامیان بدل



٩ ك

$$20 \times 10 + 10 \times 1 = 200 + 400 + 10 = 610 \quad (7)$$

bias es: $5 \times 10 + 1 = 51 \Rightarrow$ total trainable parameters: $610 + 51 = 661$

weights

مقدار مجموع المدخلات هو $\hat{y}_1 + \hat{y}_2 = 1$ ، مما يدل على أن المدخلات مترافق

$\hat{y}_1 = \text{Softmax} \left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right)_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2}}$

$\hat{y}_2 = \text{Softmax} \left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right)_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2}}$

$\hat{y}_1 + \hat{y}_2 = 1 \rightarrow \hat{y}_2 = 1 - \hat{y}_1$

$w_{3x} + b_3 = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ ، $w_{1x} + b_1 = z_1$

$\hat{y}_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2}} = \frac{1}{1 + e^{z_2 - z_1}} = \sigma(z_2 - z_1)$

$\hat{y}_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2}} = \frac{1}{1 + e^{z_1 - z_2}} = \sigma(z_1 - z_2)$

$\hat{y}_1 + \hat{y}_2 = 1 \rightarrow \sigma(z_2 - z_1) + \sigma(z_1 - z_2) = 1$

$$z_1 = w_{11}^s x_1 + w_{12}^s x_2 + b_1^s$$

$$z_2 = w_{21}^s x_1 + w_{22}^s x_2 + b_2^s$$

$$z_1 - z_2 = \pi_1(w_{11}^s - w_{21}^s)$$

in first model with sigmoid:

$$\hat{y} = \sigma(w_1^* x_1 + b_1^*) \quad ①$$

$$w_1^* = [w_1^s \quad w_2^s] \rightarrow z = w_1^s x_1 + w_2^s x_2 + b^*$$

$$\begin{array}{l} \xrightarrow{\text{①②}} w_1^s \equiv w_1^* \quad w_2^s \equiv w_2^* \quad b^s \equiv b^* \\ \hline \end{array} \quad ②$$

Suppose \underline{x} has two features:

$$\rightarrow z_1 - z_2 = \pi_1(w_{11}^s - w_{21}^s) + \pi_2(w_{12}^s - w_{22}^s) + b_1^s - b_2^s$$

$$\hat{y}_1 = \sigma(z_1 - z_2)$$

$$\hat{y}_2 = 1 - \hat{y}_1 = \sigma(z_1 - z_2)$$

دوال

ا: لئه دروامع ده تهی از مصلح مخلف (دیپ) (MLPs) میباشد باع
از مصلح رئیس فرهنگ و ارتباطات معاون از سریس بزرگ میباشد

choose the class probs from \leftarrow majority voting - 1 } final answer
 then \Rightarrow classifiers
 average the class probs from the $n=9$ classifiers and choose the class with highest prob. Average - 2 }
 take the median of the class probs from the $n=9$ classifiers and choose the class with highest median Median - 3 }

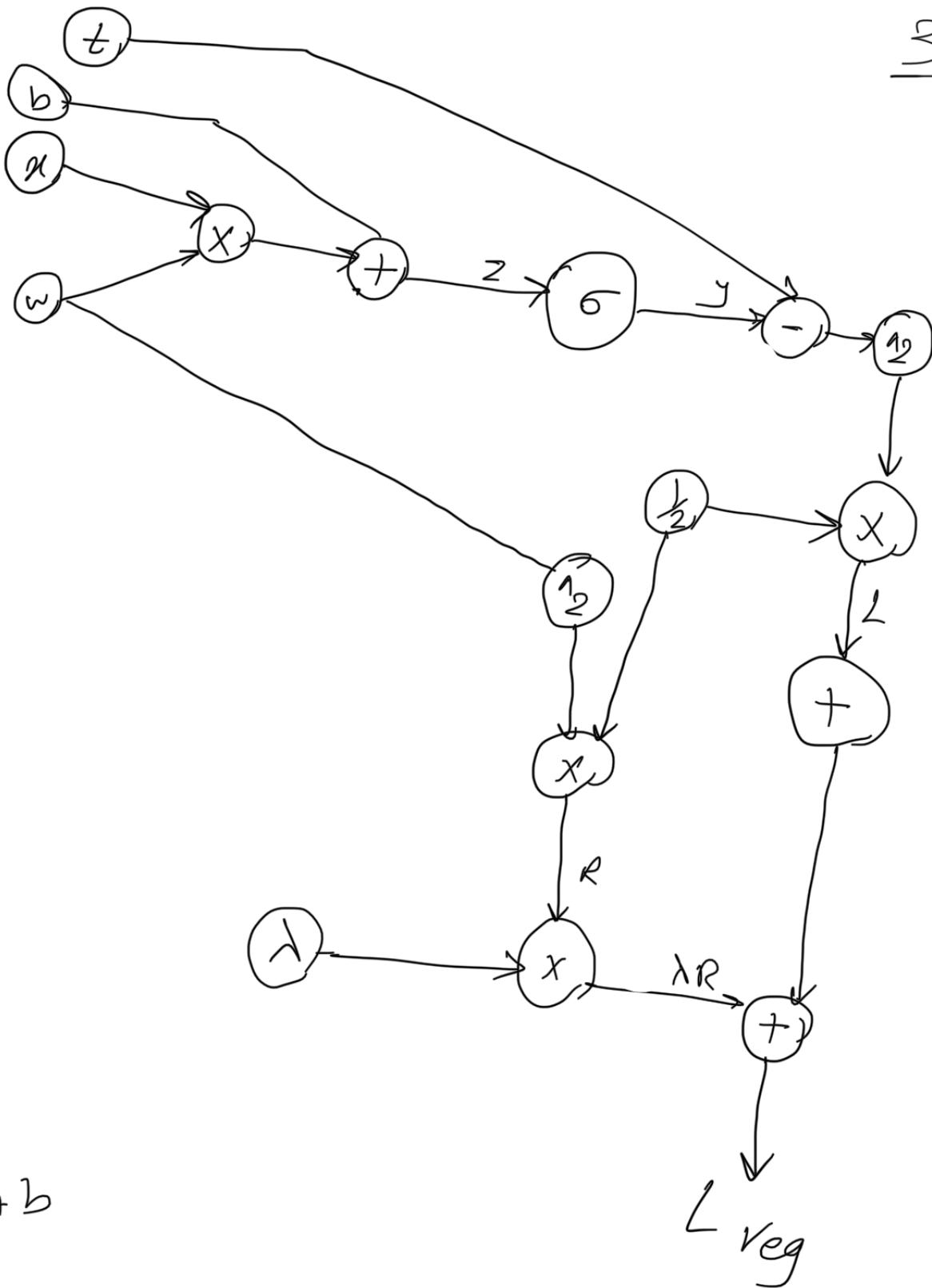
2 Nov

بـ : در اینجا می‌خواهیم یک میله 20x20 باز و چهار زوایای مختلف داشت
 ۸ دیگر را باز کنیم که از پیشنهاد شده در اینجا می‌خواهیم
 ۷ دیگر را باز کنیم که از پیشنهاد شده در اینجا می‌خواهیم
 بـ ۱ دیگر را باز کنیم که از پیشنهاد شده در اینجا می‌خواهیم
 (desanted) میله را باز کنیم که از پیشنهاد شده در اینجا می‌خواهیم
 سـ از اینجا آغاز کنید و از پیشنهاد شده در اینجا می‌خواهیم
 لـ از اینجا آغاز کنید و از پیشنهاد شده در اینجا می‌خواهیم
 PC
 (horizontally) بـ $\tan(\alpha) \times d$ از اینجا آغاز کنید و از پیشنهاد شده در اینجا می‌خواهیم
 (vertical) از اینجا آغاز کنید و از پیشنهاد شده در اینجا می‌خواهیم
 (top) از اینجا آغاز کنید و از پیشنهاد شده در اینجا می‌خواهیم
 (bottom) از اینجا آغاز کنید و از پیشنهاد شده در اینجا می‌خواهیم

و مکانیزم این را (پسندیده باشید) مخفی و 800 نفری (مخفی)
در این ریاست مکان آتش داده اند از این امر باعث شد که
در این ریاست مکان آتش داده اند از این امر باعث شد که

3 جول

; 1



$$z = w\mathbf{n} + b$$

$$\frac{\partial z}{\partial b} = 1 \quad \frac{\partial z}{\partial w} = \mathbf{n}$$

$$\sigma(z) \rightarrow \frac{\sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z))$$

$$\frac{\partial R}{\partial w} = \mathbf{w}, \quad \frac{\partial L}{\partial w} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w}$$

$$\frac{\partial \mathcal{L}}{\partial y} = y - t$$

$$\frac{\partial y}{\partial z} = \sigma(z)(1-\sigma(z))$$

$$\frac{\partial z}{\partial w} = \lambda$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial w} = (\sigma(z) - t) \underbrace{(\sigma(z)(1-\sigma(z)))}_{\lambda}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial y} \frac{\partial y}{\partial z} \underbrace{\frac{\partial z}{\partial b}}_1 = (\sigma(z) - t) \underbrace{(\sigma(z)(1-\sigma(z)))}_{\lambda}$$

$$\frac{\partial \lambda R}{\partial b} = 0$$

$$\frac{\partial \lambda R}{\partial w} = \lambda w$$

$$\frac{\partial \mathcal{L}_{reg}}{\partial w} = \frac{\partial \mathcal{L}}{\partial w} + \lambda w$$

$$\frac{\partial \mathcal{L}_{reg}}{\partial b} = \frac{\partial \mathcal{L}}{\partial b}$$

حول ۳

→ آنچه از این مکاناتی است که با تعداد روحانیت های فعال شونده همانند
در عمل ۱ نهاد بزرگترین و بارزینه باشند (با اینکه باعث میگردند
برادران تعداد بزرگ (abs: در ماهیت انسانی) از این مکانات
در درون زمینه متفاوت باشند (مکانیسم) (جهود مدنی)
→ هنوز دلیل در مکانیسم رفاقت رسانی با افراد مکالمه فعال های به تبعیت (آیینه
نموده اند که نموده اند فصل های این مکانات را برای پیشگیری در اینجا
تصالحی بین نزدیکی فصل های انتظامی انتقام از نزدیکی
بگذشت مکانیزم ایجاد نموده باشد

$$\theta_{\text{new}} = \theta_{\text{old}} - \epsilon \nabla_{\theta} L_{\text{reg}}$$

۴. (۲۵ نمره)

الگوریتم آدام^۲ برای آموزش وزن های یک شبکه عصبی به صورت تکراری گام های زیر را اجرا می کند:

$$\begin{aligned} \textcircled{1} \quad g_t &\leftarrow \nabla_{\theta} f_t(\theta_{t-1}) \\ \textcircled{2} \quad m_t &\leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ \textcircled{3} \quad v_t &\leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\ \textcircled{4} \quad \hat{m}_t &\leftarrow \frac{m_t}{1 - \beta_1^t} \\ \textcircled{5} \quad \hat{v}_t &\leftarrow \frac{v_t}{1 - \beta_2^t} \\ \textcircled{6} \quad \theta_t &\leftarrow \theta_{t-1} - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \end{aligned}$$

(آ) الگوریتم بالا را خط به خط توضیح دهید.

(ب) نشان دهید چرا مقادیر m_t به سمت صفر بایاس دارند؟ چرا مقدار \hat{m}_t که به شکل $\frac{m_t}{1 - \beta_1^t}$ محاسبه می شود

(در t های با مقادیر کمتر)، با این مشکل رویرو نمی شود؟

(توجه کنید که مقدار اولیه $m_0 = 0$ است)

۵. (۵۰ نمره) تصور کنید که تابع هدف یک مدل یادگیری ماشین به صورت $w^T H w$ باشد که اگر از تجزیه مقادیر ویژه استفاده کنیم خواهیم داشت:

$$H = Q \lambda Q^T$$

(آ) اگر از روش گرادیان کاهشی با طول گام ϵ استفاده کنیم، فرمول یادگیری ضرایب به چه صورت است؟

(ب) با شروع از حالت اولیه w ضرایب در گام t به چه صورت خواهد بود؟

(ج) تحت چه شرایطی این الگوریتم همگرا می شود؟

(د) حال بررسی کنید اگر از روش نیوتن استفاده کنیم، یادگیری به چه صورت خواهد بود؟ چند گام طول می کشد تا همگرا شویم؟

(ه) چرا با وجود اینکه روش مرتبه ۲ نیوتن از روش مرتبه ۱ گرادیان کاهشی بسیار سریع تر همگرا می شود، در آموزش شبکه های عمیق از آن استفاده نمی شود؟

۶. (۲۵ نمره) تابع خطا در یک شبکه با اعمال Dropout گوسی-جمعی به شکل زیر است:

$$J_1 = \frac{1}{2} (y_d - \sum_{k=1}^n (w_k + \delta_k) x_k)^2$$

که در آن $\delta_k \sim N(0, \alpha w_k^2)$ می باشد.

(آ) مقدار امید ریاضی گرادیان تابع هدف نسبت به متغیر w_k را محاسبه و تا حد امکان ساده کنید.

$$E\left[\frac{\partial J_1}{\partial w_i}\right]$$

(ب) آیا می توانید تعبیری از رگولاسیون با استفاده از این نوع Dropout ارائه دهید؟

Adam^۳

- خطا ناهم در محاسبه مارکوفی است ۴ Nov

: T - خط اول: در اینجا بطریق ریاضی θ_t را بدستور θ_{t-1} و β_1, β_2 می‌توان محاسبه کرد. (بجزی اول مسند ملک باعث)

- خط دوم: در اینجا همان روش را که در First momentum لفته معرف کردیم برای دویندینه بازنگشتن - خطا دویندینه θ_t را بدستور θ_{t-1} و β_1, β_2 (برای این دویندینه) می‌توان محاسبه کرد. و پس با مرور تدریجی θ_t و θ_{t-1} را محاسبه کرد.

است، جمع معرفت.

- خط سوم: در اینجا مقادیر θ_t را بدستور θ_{t-1} و β_1, β_2 می‌توان محاسبه کرد. (برای این دویندینه) می‌توان صفت هسته ای را بدستور N_{t-1} و β_1, β_2 محاسبه کرد. و در این دویندینه دو مرور تدریجی برای محاسبه θ_t انجام داده شود.

است، جمع معرفت.

- خط چهارم و پنجم: در اینجا باید این است منظور θ_t را با محاسبه θ_{t-1} و β_1, β_2 بدستور θ_t باشیم. (ازن کل بیت صدید تا در نزدیکی θ_t و θ_{t-1} به θ_t بازگشتی باشیم) θ_t را با محاسبه θ_{t-1} و β_1, β_2 بدستور θ_t باشیم. (است در این دویندینه Corrected Bias به نظر می‌رسد) و نهان زدنی نیست این است که θ_t را با محاسبه θ_{t-1} و β_1, β_2 بدستور θ_t باشیم.

- خط پنجم: در این خط پنجم با روش ۴: مجموع فندها و استنباط θ_t را بدستور θ_{t-1} و β_1, β_2 می‌توان محاسبه کرد. آنرا از تقدیر بضریب α بخوبی نهایت نهاده کن. و در این دویندینه دو مرور تدریجی برای محاسبه θ_t انجام داده شود.

- در این دویندینه دو مرور تدریجی برای محاسبه θ_t انجام داده شود.

$$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$\xrightarrow{\begin{array}{l} \beta_1 = 0.9 \approx 1 \\ 1 - \beta_1 \approx 0 \end{array}}$

$$m_t \leftarrow m_{t-1}$$

Als U.S. ist m_t nicht mit $N(m_t | C_t, m_0 = 0)$ zu verwechseln.

العملية:

: ea

$$w^T = w^{T-1} - \varepsilon \frac{\partial L}{\partial w} = w^{T-1} - \varepsilon(H^T + H)(w)$$

$$= w^{T-1} - \varepsilon (Q A Q^T + Q A^T Q^T) w^{T-1}$$

$$w^T = w^{T-1} - 2\varepsilon Q A Q^T w^{T-1}$$

$$w^T = w^{T-1} - 2\varepsilon Q A Q^T w^{T-1} = w^{T-1} - 2\varepsilon H w^{T-1}$$

$$w^T = w^o - 2\varepsilon Q A Q^T w^o, \quad w^{T-1} = w^{T-2} - 2\varepsilon Q A Q^T w^{T-2}, \dots$$

$$\Rightarrow w^t = (I - 2\varepsilon Q A Q^T) w^{T-1} = (I - 2\varepsilon Q A Q^T)^2 w^{T-2} = \dots$$

$$= \underbrace{(I - 2\varepsilon Q A Q^T)^t}_{A} w^o$$

رسالة 1-2: دلائل على صحة الـ gradient descent

$$A = Q A Q^T \rightarrow w^t = Q A^T Q^T w^o \rightarrow \underbrace{I - 2\varepsilon A^T (I - 2\varepsilon A w^o)}$$

$$f(w) = f(w^*) + (w - w^*)^T \nabla f(w^*)$$

$$+ \frac{1}{2} (w - w^*)^T H (w - w^*)$$

$$\nabla f = 0 \rightarrow \nabla_w f(w) = \nabla f(w^*) + \nabla^2 f(w^*)(w - w^*) = 0 \rightarrow w - w^* = -H^{-1} \nabla f(w^*)$$

$$\rightarrow w^{t+1} = w^t - w^t = 0$$

رسالة 2-1: دلائل على صحة الـ gradient descent

حول:

: \tilde{T}

$$E[J] = \frac{1}{2} E \left[\left(y_d - \sum_k (w_k + \delta_k) x_k \right)^2 \right]$$

$$= \frac{1}{2} E \left[\left(y_d^2 - 2y_d \sum_k (w_k + \delta_k) x_k + \left(\sum_k (w_k + \delta_k) x_k \right) \left(\sum_{l \neq k} (w_l + \delta_l) x_l \right) \right) \right]$$

$$= \frac{1}{2} E \left[\left(-2(w_i + \delta_i) x_i y_d + (w_i + \delta_i)^2 x_i^2 + 2(w_i + \delta_i) x_i \sum_{k \neq i} (w_k + \delta_k) x_k \right) \right]$$

$$= \left(-w_i x_i y_d + 0.5 x_i^2 w_i^2 + 0.5 \alpha x_i^2 x_i^2 + (w_i x_i) \sum_{k \neq i} w_k x_k \right)$$

$$E \left(\frac{\partial J}{\partial w_i} \right) = \frac{\partial}{\partial w_i} E[J] = \left[-x_i y_d + w_i x_i^2 + \alpha w_i x_i^2 + x_i \sum_{k \neq i} w_k x_k \right]$$

$$= \left[-x_i y_d + \sum_{k \neq i} w_k x_k x_i + \alpha w_i x_i^2 \right]$$

$$J_c = \frac{1}{2} \left(y_d - \sum_{k=1}^n (w_k + \delta_k) x_k \right)^2 + \frac{1}{2} \alpha \sum_{k=1}^n (w_k x_k)^2$$

Regularization term \rightarrow

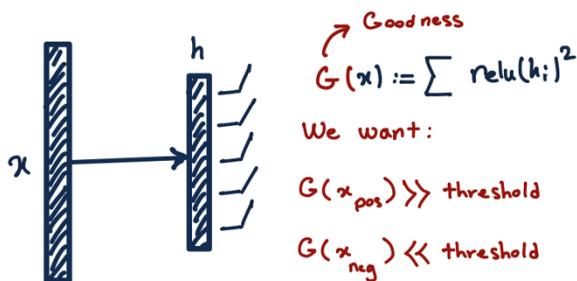
سوالات عملی (۳۰۰ نمره)

۱. (۱۵۰ نمره) در این تمرین شبکه عصبی Fully-Connected را پیاده سازی خواهید کرد. در این پیاده سازی با طراحی شبکه، انتشار رو به عقب (backpropagation)، انواع بهینه سازها، Dropout برای رگولاژن شبکه آشنا خواهید شد. فایل zip داده شده شامل یک نوتبوک و یک فایل fully-connected-networks.py می باشد. برای هر قسمت شما باید فایل پایتون py. را کامل کنید و با تست پیاده سازی خود در فایل نوتبوک از صحبت پیاده سازی درست خود اطمینان حاصل کنید.

توجه!

- (آ) هیچ کدی را خارج از بلاک های مشخص شده در نوتبوک و فایل پایتون ننویسید یا تغییر ندهید.
- (ب) سلول های نوتبوک را اضافه یا کم نکنید.
- (ج) در نهایت هر دو فایل py. و نوتبوک را به عنوان پاسخ نهایی ارسال نمایید.

۲. (۱۵۰ نمره) هدف از این مسئله پیاده سازی الگوریتم forward – forward و آموزش یک شبکه عصبی به وسیله آن بر روی دیتاست MNIST است. این الگوریتم در مقاله ای از ارائه جفری هینتون در NeurIPS ۲۰۲۲ شرح داده شده است که شما از طریق این [لينک](#) می توانید آن را مطالعه کنید. به طور خلاصه، الگوریتم forward-forward جایگزینی برای EBP است که از دو مسیر forward استفاده میکند. ایراد EBP عدم تطابق آن با مدل مغز، عدم پیاده سازی بهینه آن برای کاربردهای real time و نیاز آن به وجود یک مدل کامل است. از این جهات الگوریتم Forward-Forward می تواند جایگزین EBP باشد. در این الگوریتم دیتا به همراه لیبل درست تبدیل به یک وکتور و همچنین دیتا به همراه لیبل غلط به یک وکتور دیگر تبدیل می شوند سپس به کمک یکتابع هزینه محلی وزن ها تعیین می شوند. اینتابع هزینه محلی برای هر لایه تعریف می شود و در نتیجه دیگر نیازی به back propagation نمی باشد. هم چنین مفهومی به نام goodness تعریف می شود که لازم است مقدار آن برای دیتاهای صحیح بیشینه و برای دیتا های غلط کمینه باشد. دقت کنید خروجی هر لایه نیاز به نورمالایز شدن دارد تا از تحریک لایه های بعد جلوگیری شود.



در ابتدا می خواهیم این الگوریتم را به روش Supervised پیاده کنیم. برای پیاده سازی این الگوریتم به ترتیب گام های زیر را انجام دهید:

- (آ) **لود کردن دیتاست:** دیتاست مورد استفاده در این بخش دیتاست MNIST است. با تعریف یک dataloader مناسب داده ها را به شیوه مناسب بخوانید.
- (ب) **تولید داده:** برای این کار ابتدا لیبل ها را به یک بردار one-hot تبدیل کرده و آن را در تصویر هر دیتا جایگذاری نمایید. از لیبل های اشتباه برای تولید دیتای منفی استفاده کنید.
- (ج) **پیاده سازی شبکه:** می دانیم هر لایه به صورت محلی محاسبات مربوط به آپدیت وزن های و خطای را انجام می دهد. با تعریف مناسب تابع خطای نحوی که شرط لازم goodness برای داده های مثبت و منفی برقرار باشد، هر لایه را به صورت مناسب تعریف کنید. loss برابر است با:

$$loss = \text{mean}(\log(1 + e^{[(\text{threshold}-\text{positivedata}), (\text{negatedata}-\text{threshold})]]}))$$

توضیح دهید چرا تعريف loss به شکل بالا میتواند شروط لازم goodness را برآورده کند در نهایت با استفاده از لایه های پیاده سازی شده، شبکه یادگیری را تعريف کنید. توجه داشته باشید که شبکه باید قابلیت آموزش و پیش بینی داشته باشد. دقت کنید در هر ایپاک باید از batch دیتای درست، دیتای غلط را تولید و به همراه دیتای درست forward کنید.

(د) **گزارش نتایج:** در انتها مقدار خطا و دقت مدل را برای داده های آموزش و تست بدست آورده و گزارش دهید. هم چنین توضیحات لازم راجع به الگوریتم و نحوه پیاده سازی آن را در گزارش خود ذکر کنید.

حال این الگوریتم را به روش Unsupervised پیاده سازی می کنیم. دقت کنید در این روش مجاز به استفاده از لیل ها در فرآیند آموزش شبکه FF نمی باشید:

(آ) **تعريف mask مناسب:** با استفاده از روش توضیح داده شده در مقاله mask مناسب را برای تولید داده های منفی تعريف کنید.

(ب) **تولید داده:** به کمک mask ای که در قسمت قبل تعريف کردید، تابعی برای تولید داده های هیبرید که به عنوان داده های منفی استفاده می شوند بنویسید.

(ج) **پیاده سازی شبکه:** مجدداً با تعريف مناسبتابع loss به نحوی که شروط لازم goodness برای داده های مثبت و منفی برقرار باشد، لایه را تعريف کنید loss برابر است با:

$$loss = \text{mean}(\log(1 + e^{[(\text{threshold}-\text{positivedata}), (\text{negatedata}-\text{threshold})]]}))$$

در نهایت شبکه یادگیری را به نحوه مناسب پیاده سازی کنید. بدیهتا این شبکه باید قابلیت آموزش و پیش بینی را داشته باشد. سپس با استفاده از تابعی که برای داده های منفی تعريف کرده بودید، شبکه FF را آموزش دهید. دقت کنید در هر ایپاک باید از batch دیتای درست، دیتای غلط را تولید و به همراه دیتای درست forward کنید. توضیح دهید خروجی شبکه شما چیست و چه ارتباطی با لیل ها دارد؟

(د) **پیاده سازی طبقه بند خطی:** یک linear classifier تعريف کنید. دقت کنید فقط مجاز به استفاده از شبکه FF در حالت inference هستید. چرا باید طبقه بند خطی تعريف و آموزش داده شود؟

(ه) **گزارش نتایج:** در انتها مقدار خطا و دقت مدل را برای داده های آموزش و تست بدست آورده و گزارش دهید. هم چنین توضیحات لازم راجع به الگوریتم و نحوه پیاده سازی آن را در گزارش خود ذکر کنید.