به نام خدا

یادگیری ژرف

نیم‌سال اول ۴۰۳-۱۴۰۲

مدرس: دکتر بیگی

دانشکده‌ی مهندسی کامپیوتر

تمرین سری اول — مرور جبرخطی، بهینه‌سازی، منظم‌سازی، شبکه‌های عصبی و انتشار رو به عقب — نمره کل: ۵+۱۰۰ نمره

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- در هر کدام از سوالات، اگر از منابع خارجی استفاده کرده‌اید باید آن را ذکر کنید. در صورت همفکری با افراد دیگر هم باید نام ایشان را در سوال مورد نظر ذکر نمایید.
- پاسخ ارسالی واضح و خوانا باشد. در غیر این صورت ممکن است منجر به از دست دادن نمره شود.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد. به اسکرین‌شات از منابع یا پاسخ افراد دیگر نمره‌ای تعلق نمی‌گیرد.
- در صورتی که بخشی از سوال‌ها را جای دیگری آپلود کرده و لینک آن را قرار داده باشید، حتما باید تاریخ آپلود مشخص و قابل اعتنا باشد.
- تمام پاسخ‌های خود را در یک فایل با فرمت HW1_[StudentID]_[Fullname].zip روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۰ روز (تا سقف ۳ روز برای هر تمرین) وجود دارد.

سوالات نظری (۶۰ نمره)

سوال ۱: مرور جبر خطی (۱۲ نمره)

۱- نشان دهید ماتریس Hessian یک تبدیل مانند $y = \psi(u, v, z)$ را میتوان به صورت ماتریس ژاکوبی گرادیان این تبدیل نوشت. متغیر های u, v, z را تک بعدی و y را تابعی بر حسب آن‌ها در نظر بگیرید.

۲- موارد زیر را اثبات کنید $(y \in \mathbb{R}, x \in \mathbb{R}^n, a \in \mathbb{R}^n, X \in \mathbb{R}^{n \times n}, A \in \mathbb{R}^{n \times n})$:

الف_ $\frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a^T$

ب_ $\frac{\partial}{\partial y} tr(X) = tr(\frac{\partial}{\partial y} X)$

پ_ $\frac{\partial}{\partial x} tr(X^T A X) = X^T (A + A^T)$

ت_ $\frac{\partial}{\partial x} log(\det(X)) = X^{-T}$

راهنمایی: معکوس یک ماتریس مربعی به شکل زیر بدست می آید:

$$(X^{-1})_{ij} = \frac{1}{\det(A)} C_{ij}$$

که در آن C ماتریس cofactor است و داریم:

$$C_{ij} = (-1)^{i+j} \det(M_{ij})$$

که M_{ij} ماتریس X بدون سطر i و ستون j است.

۳- ماتریس تبدیل دوران در دو بعد به اندازه θ درجه به فرم زیر است:

$$R(\theta) = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix}$$

مقادیر و بردارهای ویژه‌ی این ماتریس را به دست آورید و صحت رابطه‌ی زیر را برای آن نشان دهید:

$$\det(X) = \prod_{i=1}^{n} |\lambda_i|$$

که در آن λ_i ها مقادیر ویژه‌ی ماتریس X هستند.

در نهایت با استفاده از تجزیه مقادیر ویژه نشان دهید:

$$R(n\theta) = R^n(\theta)$$

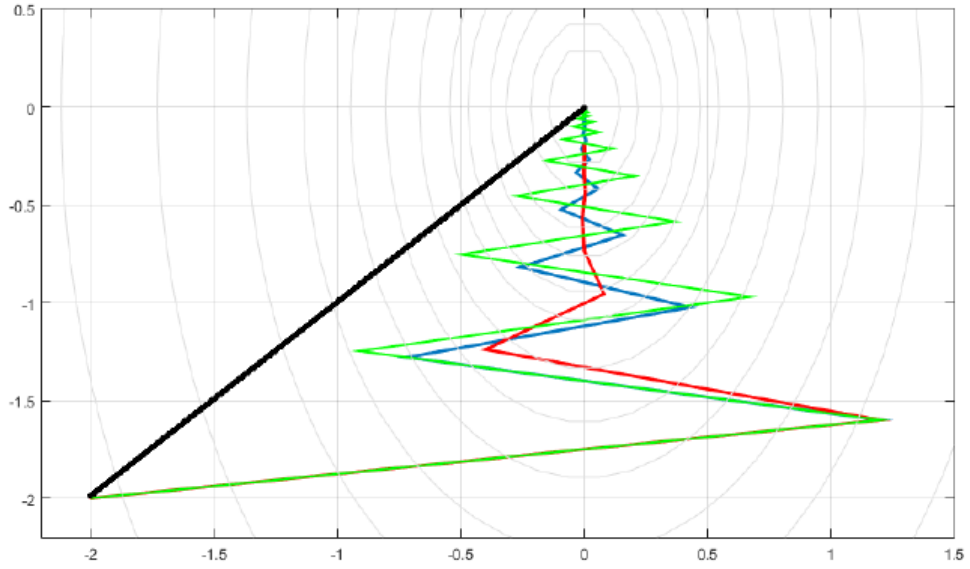
سوال ۲: بهینه سازی (۸ نمره)

۱- چرا در ابعاد بالا نقاط زینی بیشتر از نقاط کمینه محلی وجود دارند؟

۲- در شکل زیر مسیر بهینه سازی برای روش های Nestrov-Momentum و Momentum و GD و RMSprop برای یک تابع مرتبه ۲ از نقطه شروع $[-2, -2]$ رسم شده است. نمودار آبی رنگ مربوط به روش GD است.

الف) توضیح دهید سه نمودار سبز، قرمز و مشکی مربوط به کدام یک از سه روش نام برده دیگر است.

ب) مزایا و معایب سه روش را بیان کنید و بگویید چگونه این روش ها مشکلات GD را حل میکنند.



شکل ۱: مسیر همگرایی

پ) توضیح دهید روش ADAM چگونه مشکل روش Momentum را حل میکند و علت استفاده از Bias correction در این روش چیست؟

سوال ۳: منظم سازی (۱۰ نمره)

با توجه به مقاله Dropout به سوالات ۱ تا ۴ پاسخ دهید.

۱- توضیح دهید که چرا Dropout عمکردی مانند Ensemble-Learning دارد و علت برتری Dropout در شبکه های با تعداد پارامتر بالا نسبت به آن چیست؟

۲- توضیح دهید که چرا Dropout مانند منظم ساز عمل می‌کند؟

۳- تفاوت استفاده از Dropout در حین Train و Test چیست و علت این تفاوت را ذکر کنید؟

۴- (۵ نمره امتیازی) در مسئله رگرسیون خطی برای N داده با ماتریس داده $X \in \mathbb{R}^{N \times D}$ و خروجی هدف $y \in \mathbb{R}^N$ و تابع هزینه زیر اگر Dropout را بر روی ورودی اعمال کنیم(یعنی هر المان از ماتریس X با احتمال p در مسئله رگرسیون حضور داشته باشد.) اثبات کنید Dropout معادل استفاده از جمله منظم ساز در تابع هزینه است.(پاسخ کامل مورد نظر است).

$$J(w) = \|y - Xw\|_2^2$$

۵- علت اینکه در صورت استفاده از Batch-Normalization به راحتی می‌توان نرخ یادگیری وزن‌ها را تنظیم کرد چیست؟

۶- Batch-Normalization چگونه باعث منظم سازی میشود و با ذکر دلیل بیان کنید که بزرگ شدن سایز بچ چه تاثیری بر ویژگی منظم سازی آن خواهد داشت؟

سوال ۴: توابع فعال سازی (۸ نمره)

۱- برای هر مساله ی دسته بندی زیر مشخص کنید تابع فعالسازی مناسب برای لایه خروجی شبکه عصبی آن چیست؟علت آن را ذکر کنید.

(آ) دسته بندی را در نظر بگیرید که تصویر ورودی را دریافت می کند و میخواهد تعیین کند تصویر ورودی سگ است یا گربه.

(ب) دسته بندی را در نظر بگیرید که تصویر ورودی یک حیوان را دریافت می کند و میخواهد تعیین کند تصویر ورودی متعلق به کدام دسته از حیوانات از میان ۱۰۰ دسته مشخص است.

(ج) دسته بندی را در نظر بگیرید که تصویر ورودی را دریافت می کند و میخواهد تعیین کند تصویر ورودی شامل کدام حیوانات است. (تصویر ممکن است شامل چندین حیوان باشد)

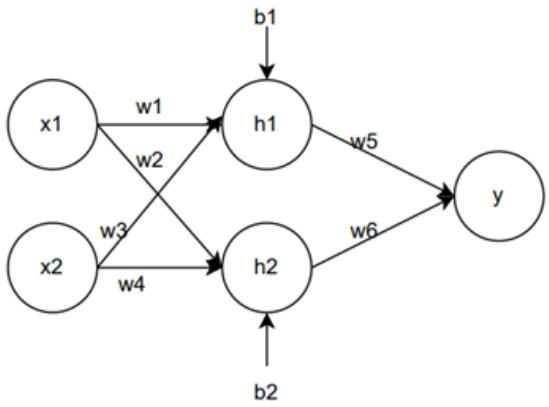
۲- Relu به عنوان یکی از توابع پرکاربرد برای حل مساله محوشدگی گرادیان کاربرد دارد با این حال این تابع در مقادیر منفی مقدار صفر دارد و باعث می شود در فرآیند آموزش برخی از واحد ها بروز نشوند. برای حل این مساله نسخه های مختلفی از این تابع فعالساز ارایه شده است.یکی از جدید ترین نمونه های آن تابع فعالساز DPReLU است. با مراجعه به مقاله این تابع فعالساز، ابتدا تابع مورد استفاده دراین تابع را با تابع Relu مقایسه کنید و مشخص کنید هر پارامتر در این تابع فعالساز چه کاربردی دارد. سپس بررسی کنید که نسخه های قبلی این تابع فعالساز چه مشکلاتی داشتند و این تابع فعالساز چگونه توانسته است آن ها را بهبود دهد.

سوال ۵: شبکه های عصبی و انتشار رو به عقب (۲۲ نمره)

۱- شبکه عصبی زیر را با وزن ها و بایاس های داده شده در نظر بگیرید. تابع فعالسازلایه میانی Lrelu و تابع فعالساز لایه خروجی تابع سیگموید است

محاسبات مربوط به الگوریتم انتشار به عقب را برای یک گام انجام دهید. تابع هزینه را mean square error در نظر بگیرید.(نرخ یادگیری ۰.۱)

*برای محاسبات تا سه رقم اعشار را در نظر بگیرید.



شکل ۲: شبکه عصبی دولایه

$$[x_1, x_2] = [0, 1] \quad (۱)$$

$$[y] = [1] \quad (۲)$$

$$[w_1, w_2, w_3, w_4] = [0.3, 0.2, 0.2, -0.6] \quad (۳)$$

$$[w_5, w_6] = [0.5, -1] \quad (۴)$$

$$[b_1, b_2] = [0.2, -1.4] \quad (۵)$$

$$Lrelu(x) = 1(x \geq 0)(x) + 1(x < 0)(0.2x) \quad (۶)$$

$$L = \frac{1}{2}(\hat{y} - y)^2 \quad (۷)$$

۲- تکنیک batch normalization که در لایه های مختلف شبکه می تواند استفاده شود، با نرمالسازی دادگان یک batch در لایه به بهبود عملکرد مدل بر روی دادگان جدید کمک می کند. فرض کنید $[x_1, x_2]$به عنوان ورودی لایه batch normalization داده می شود و متغیر های میانی $[\hat{x_1}, \hat{x_2}]$تولید می شود و در نهایت با رابطه ی $\hat{y_k} = \gamma\hat{x_k} + \beta$ به خروجی لایه تبدیل می شود. با رسم گراف محاسباتی لایه batch normalization روابط بین میانگین و واریانس را با ورودی و خروجی لایه بنویسید(پارامتر های قابل یادگیری لایه را γ, β در نظر بگیرید). سپس روابط مشتقات تابع هزینه شبکه را با فرض داشتن $\frac{\partial L}{\partial \hat{y_2}}$ و $\frac{\partial L}{\partial \hat{y_1}}$ نسبت به β و γ و x_1 و x_2 محاسبه کنید.

سوالات عملی (۴۰ نمره)

سوال ۶: سوال عملی

- به نوت بوک مراجعه شود