



دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

یادگیری ژرف

نیم سال اول ۰۱-۰۲

استاد: دکتر حمید بیگی

گردآوردندگان: علی سطوتی - محمدعلی صدرایی جواهری - مهدی کافی - فاطمه مهدوی

بررسی و بازبینی: مهران حسینزاده

تمرین اول

مفاهیم اولیه

مهلت ارسال: ۱۷ آبان ۱۴۰۱

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.
- در طول ترم امکان ارسال با تاخیر تمرین ها بدون کسر نمره تا سقف ۱۰ روز (تا سقف ۳ روز برای هر تمرین) وجود دارد. محل بارگزاری جواب تمرین ها بعد از ۵ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ های ارسال شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- هم کاری و هم فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم فکری و یا استفاده از هر منابع خارج درسی، نام هم فکران 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.
- پاسخ تمامی سوالات (تئوری و عملی) را در یک فایل فشرده به صورت `DL_Hw1_[firstName]_[lastName]_[StudentId]` نام گذاری کرده و ارسال کنید.

سوالات نظری (۶۰ نمره)

۱. رگرسیون خطی (۱۲ نمره)

یکی از پر استفاده ترین الگوریتم های یادگیری ماشین، رگرسیون خطی است. در این روش، خروجی یک نمونه بر اساس مدلی خطی از متغیرهای مستقل، به صورت زیر، محاسبه می شود.

$$y = \sum_{i=0}^n x_i \cdot w_i$$

و خروجی چندین داده به صورت همزمان با ضرب ماتریسی زیر محاسبه خواهد شد.

$$y = Xw$$

(آ) (۴ نمره) با توجه به رابطه ضرب ماتریسی بالا، وزن های الگوریتم به راحتی به صورت $w = X^{-1}y$ قابل محاسبه خواهد بود. توضیح دهید چرا در اکثر مواقع قادر به استفاده از این رابطه نیستیم و در چنین شرایطی فرم بسته پاسخ رگرسیون خطی را محاسبه نمایید.

(ب) (۸ نمره) هنگام استفاده از الگوریتم رگرسیون خطی، فرضیاتی در نظر گرفته می شود. یکی از این فرضیات عدم وجود auto correlation است. این مفهوم را توضیح دهید. در عمل آزمایشی با عنوان Durbin Watson برای بررسی وجود auto correlation استفاده می شود. این روش را بر روی داده های زیر اعمال کنید و مشخص کنید که نتیجه آزمون، وجود یا عدم وجود auto correlation و یا عدم توانایی اظهار نظر درباره آن است. آستانه معناداری را ۵٪ در نظر بگیرید ($\alpha = 0.05$).

n	x	y
۱	۴۸/۹	۶۰/۸
۲	۵۰/۶	۶۲/۵
۳	۵۲/۹	۶۴/۶
۴	۵۵	۶۶/۱
۵	۵۶/۸	۶۷/۷
۶	۵۸/۸	۶۹/۱
۷	۶۱/۲	۷۱/۷
۸	۶۲/۵	۷۳/۵
۹	۶۴/۷	۷۶/۲
۱۰	۶۵	۷۷/۳

۲. توابع فعال سازی (۱۶ نمره)

به سوالات زیر پاسخ دهید.

(آ) (۸ نمره) توابع زیر را از نظر محو شدن گرادیان، مشتق‌پذیری، هزینه محاسبه و zero-centered بودن بررسی کنید. همچنین مشتق هر کدام را در صورت وجود، محاسبه کنید.

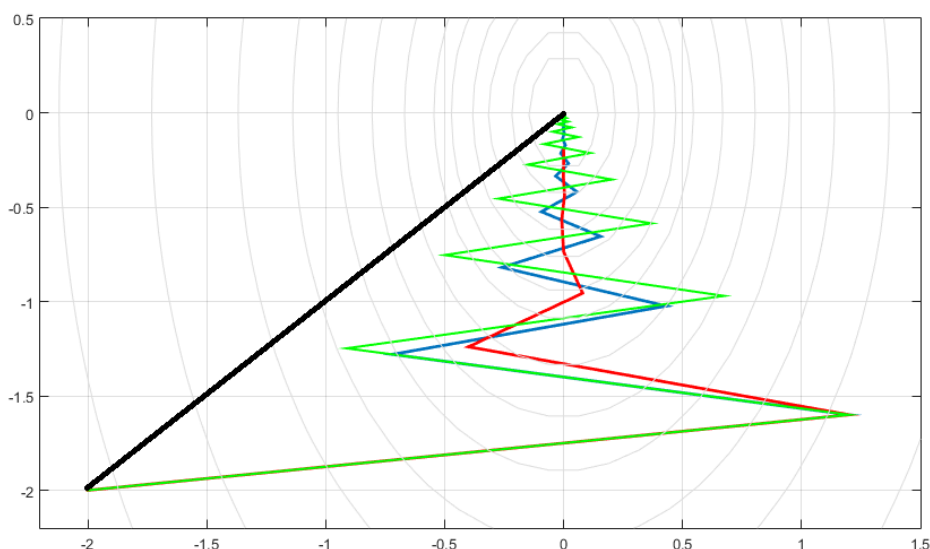
- Binary step
- Sigmoid
- Tanh
- ReLU
- Leaky ReLU

(ب) (۸ نمره) تابع فعال سازی‌ای در سال ۲۰۱۷ توسط گوگل با عنوان swish معرفی شده است. فرمول این تابع و مشتق آن را بنویسید و نمودار آنها را رسم کنید. با مطالعه [مقاله](#) مرتبط با این تابع فعال سازی، نتایج دقت مدل‌ها را در دو حالت استفاده از این تابع یا استفاده از تابع ReLU با یکدیگر مقایسه کنید و بررسی کنید که کدام تابع به صورت کلی عملکرد بهتری داشته است. همچنین این تابع را از نظر محو شدن گرادیان و zero-centered بودن بررسی کنید. تابع دیگری با عنوان hard swish نیز معرفی شده است. رابطه این تابع را بنویسید و مزیت آن را نسبت به swish ذکر کنید.

۳. بهینه‌سازی و منظم‌سازی (۱۹ نمره)

به سوالات زیر پاسخ دهید.

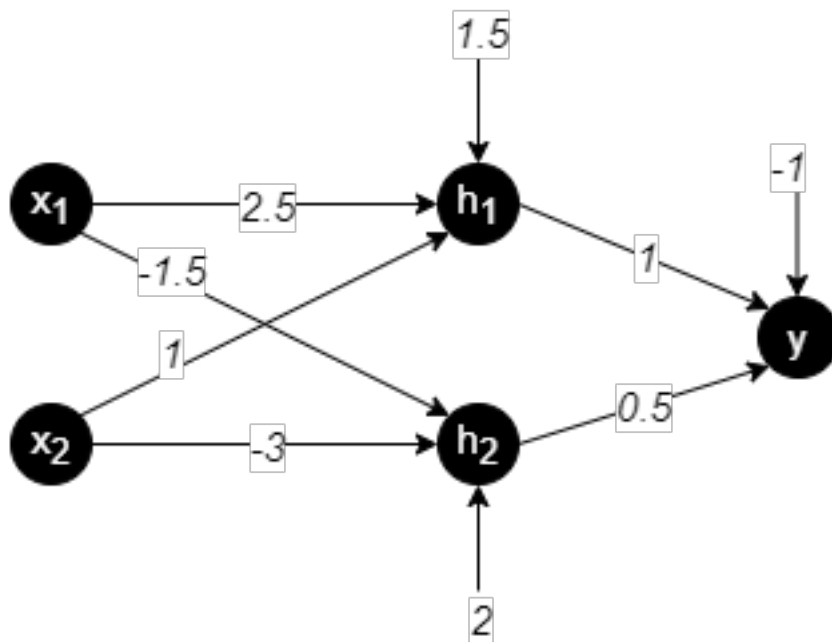
(آ) (۹ نمره) در شکل زیر مسیر بهینه‌سازی برای روش‌های Nestrov-Momentum و Momentum و GD و RMSprop برای یک تابع مرتبه ۲ از نقطه شروع $[-2, -2]$ رسم شده است. نمودار آبی رنگ مربوط به روش GD است.



- (۱) (۳ نمره) توضیح دهید سه نمودار قرمز، مشکی و سبز مربوط به کدام یک از سه روش نام برده دیگر است.
- (۲) (۳ نمره) مزایا و معایب سه روش دیگر را بیان کنید و بگویید چگونه این روش‌ها مشکلات GD را حل می‌کنند.
- (۳) (۳ نمره) توضیح دهید روش ADAM چگونه مشکل روش Momentum را حل می‌کند و علت استفاده از Bias Correction در این روش چیست؟
- (ب) (۴ نمره) روش‌های multi-task learning و parameter sharing را به طور مختصر توضیح دهید و بگویید که چگونه باعث افزایش generalization می‌شوند.
- (ج) (۶ نمره) به سوالات زیر در مورد drop-out پاسخ دهید:
- (۱) (۲ نمره) توضیح دهید که چرا drop-out مانند منظم‌ساز عمل می‌کند؟
- (۲) (۲ نمره) توضیح دهید که چرا drop-out عملکردی شبیه ensemble-learning دارد؟
- (۳) (۲ نمره) استفاده از drop-out در حین train و test چه تفاوتی دارد و این تفاوت به چه علت است.

۴. Backpropagation (۱۳ نمره)

- (آ) (۸ نمره) فرض کنید شبکه عصبی با دو ورودی $x = (x_1, x_2)$ ، یک لایه پنهان با دو نورون و یک نورون خروجی داریم. هر نورون یک ورودی بایاس دارد و تابع سیگموئید به عنوان تابع فعال‌ساز در نورون‌های لایه پنهان و خروجی استفاده می‌شود. در شکل زیر مقادیر وزن‌ها و بایاس‌ها داده شده است.



فرض کنید x_1 و x_2 و مقدار واقعی y به ترتیب صفر، ۱ و ۱ باشد. محاسبات مربوط به الگوریتم انتشار به عقب را تنها برای یک گام انجام دهید. همچنین تابع هزینه logistic regression را در نظر بگیرید. (نرخ یادگیری را برابر ۰/۱ در نظر بگیرید)

(ب) (۵ نمره) شبکه زیر را در نظر بگیرید:

$$\begin{aligned} z_1 &= w_1 x + b_1 \\ h_1 &= \text{Relu}(z_1) \\ z_2 &= w_2 h_1 + b_2 \\ h_2 &= \text{Relu}(z_2) \\ \hat{y} &= \text{Softmax}(h_2) \\ J &= - \sum_i y_i \log(\hat{y}_i) \end{aligned}$$

ابعاد پارامترها به این صورت می باشند:

$$x \in \mathbb{R}^{D_x \times 1}, w_1 \in \mathbb{R}^{H \times D_x}, b_1 \in \mathbb{R}^H, w_2 \in \mathbb{R}^{D_y \times H}, b_2 \in \mathbb{R}^{D_y}, \hat{y} \in \mathbb{R}^{D_y \times 1}$$

مقادیر $\frac{\partial J}{\partial w_1}$, $\frac{\partial J}{\partial w_2}$, $\frac{\partial J}{\partial b_1}$ و $\frac{\partial J}{\partial b_2}$ را برای یک نمونه ورودی با جزئیات مراحل محاسبه کنید.

سوالات عملی (۴۰ نمره)

۱. دسته‌بندی تصاویر با استفاده از پایتورچ (۴۰ نمره)

برای حل این سوال به نوتبوک ضمیمه شده مراجعه کنید و همه‌ی پاسخ‌های خواسته شده را داخل همان نوتبوک ارائه دهید.