مبانی یادگیری ماشین

پاییز ۱۴۰۳

استاد: دکتر حامد ملک

دانشکده مهندسی و علوم کامپیوتر

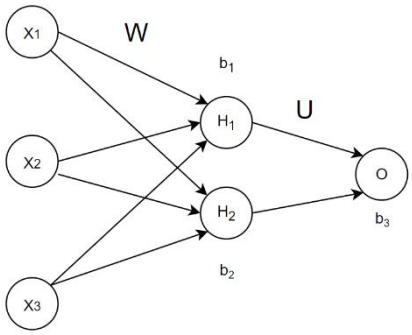
تمرین شماره ۳

- قبل از شروع تمرین، فایل مربوط به قوانین حل و تحویل تمرین‌ها را مطالعه کنید.
- سؤالات و مشکلات خود را درباره این تمرین می‌توانید در گروه تلگرامی درس یا با طراحان این تمرین مطرح کنید.
- نوشتن گزارش کامل و تفسیر نتایج اجباری است. جزئیاتی مانند روش‌های مورد استفاده، تاثیر هر روش در نتیجه نهایی و بهبود حاصل‌شده به همراه ارائه معیارهای ارزیابی خواسته‌شده در گزارش ضروری است. با هر تغییر و هر بهبود، تغییر مقادیر معیارهای ارزیابی نیز ذکر شود. ضمنا برای گزارش سؤالات عملی، نیاز به فایل دیگری نیست و در همان محیط کدزنی، در قالب بلاک‌های مارک‌داون ، توضیحات مورد نیاز را قرار دهید.
- پاسخ سؤالات تئوری را در یک فایل پی‌دی‌اف[1] با عنوان گزارش قرار دهید.
- برای سوالات عملی، نوتبوک‌های داده شده را تکمیل نمایید.
- طراحان این تمرین: امیرحسین ایزدی – امیررضا افتخاری

سؤالات تئوری — 5 + 35 نمره

سؤال ۱ (۱۴ نمره)

یکی از اساسی‌ترین اجزا در یک شبکه ژرف، انتشار به عقب[2] است. شبکه عصبی پیش‌رونده[3] با معماری زیر را در نظر بگیرید:



- لایه اول: لایه تماماً متصل[4] با تابع فعالساز سیلو[5]
- لایه دوم: لایه تماماً متصل با تابع فعالساز سیگموئید[6]

[1] PDF
[2] Backpropagation
[3] Feedforward
[4] Fully Connected
[5] SiLU
[6] Sigmoid

اگر وزن‌های ابتدایی و ورودی شبکه مقادیر زیر باشند و همچنین تابع خسارت[7]، تابع Cross-Entropy و بهینه‌ساز ما در این مسئله گرادیان کاهشی[8] به همراه مومنتوم[9] و پارامتر های η و γ به ترتیب ۰.۸ و ۰.۹ باشند، به موارد زیر پاسخ دهید.

نمونه های ورودی:

$$x^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}, y^{(1)} = 1$$

$$x^{(2)} = \begin{bmatrix} 0.1 \\ 0.5 \\ 1 \end{bmatrix}, y^{(2)} = 0$$

$$x^{(3)} = \begin{bmatrix} 0 \\ 1 \\ 0.7 \end{bmatrix}, y^{(3)} = 0$$

وزن های ابتدایی شبکه:

$$w_{11} = 0.4 \quad w_{12} = -0.3 \quad w_{21} = 0.2 \quad w_{22} = 0.2 \quad w_{31} = -0.5 \quad w_{32} = 0.1$$

$$b_1 = 0 \quad b_2 = -0.4 \quad b_3 = 0.1$$

$$u_1 = 0.3 \quad u_2 = -0.1$$

الف) گراف محاسباتی[10] را رسم کنید.

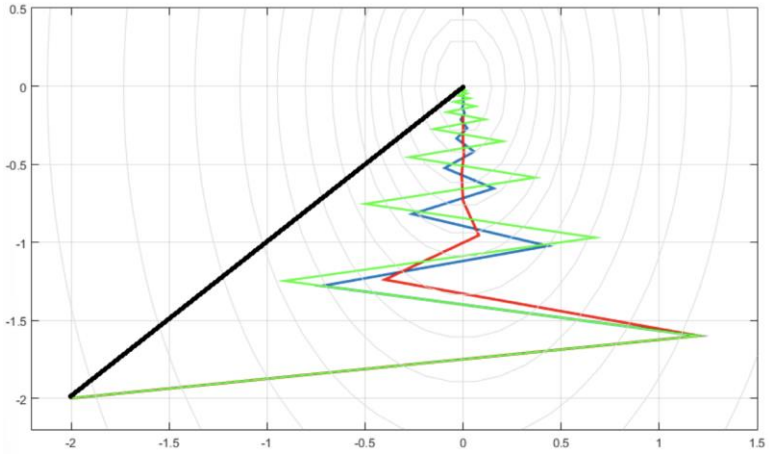
ب) عملیات انتشار رو به جلو[11] و انتشار به عقب را برای دو بار بروزرسانی وزن‌ها انجام دهید و وزن‌های جدید به همراه میزان تابع خطا را پس از بروزرسانی وزن‌ها گزارش کنید. توجه داشته باشید تمامی محاسبات شما باید بصورت ماتریسی انجام شود در غیر این صورت نمره‌ای از این سوال کسب نخواهید کرد.

سؤال ۲ (۱۲ نمره)

به سوالات زیر پاسخ دهید.

الف) در شکل زیر مسیرهای بهینه‌سازی برای روش‌های Nestrov-Momentum و Momentum و GD[12] و RMSProp برای یک تابع مرتبه ۲ از نقطه شروع [-2,2] رسم شده است. نمودار آبی مربوط به روش GD است.

[7] Loss Function
[8] Gradient Descent
[9] Momentum
[10] Computational Graph
[11] Forward Propagation
[12] Gradient Descent



- با ذکر دلیل کافی توضیح دهید سه نمودار قرمز، مشکی و سبز هر کدام مربوط به کدام یک از سه روش دیگر است؟ می‌توانید برای بررسی دقیق‌تر نحوه‌ی عملکرد این الگوریتم‌ها، به این لینک نیز مراجعه کنید.
- مزایا و معایب سه روش دیگر را بیان کنید و بگویید چگونه این روش‌ها مشکلات GD را حل می کنند؟
- الگوریتم AdaGrad را در نظر بگیرید. ضمن مقایسه آن با RMSProp بطور جامع، توصیف کنید اگر مسیر این الگوریتم را در شکل بالا قصد داشتیم حدس بزنیم، چگونه بود؟

ب) با توجه به گرادیان محاسبه شده در یک نقطه، بهینه ساز Adam سه مرحله مجزا دارد: اول به روزرسانی میانگین متحرک؛ دوم اعمال تصحیح بایاس و سوم به‌روزرسانی پارامترها. میانگین متحرک مربع گرادیان‌ها را بصورت بازگشتی که در ادامه داده شده است در نظر بگیرید:

$$s_t = \beta_۲ s_{t-۱} + (۱ - \beta_۲) g_t^۲$$

- عبارت S_t را فقط بر حسب گرادیان های $g_۰, g_۱, \ldots, g_t$ بنویسید.
- با توجه به عبارت داده شده در قسمت قبل، $E[S_t]$ را برحسب $E[g_t^۲]$ و $\beta_۲$ بنویسید. می‌توانید فرض کنید که g_i ها از هم مستقل و هم‌توزیع هستند. فرمول‌های زیر ممکن است مفید باشند:

$$\sum_{i=۰}^{n-۱} (a + id) = \frac{n}{۲}(۲a + (n-۱)d)$$

$$\sum_{i=۰}^{n-۱} ar^i = a\frac{r^n - ۱}{r - ۱}$$

- با استفاده از نتیجه گیری خود در قسمت قبل، توضیح دهید که اگر مرحله تصحیح بایاس را انجام نمی‌دادید، چه اتفاقی می‌افتاد؟
- الگوریتم Adam را با RMSProp مقایسه کنید. هر یک چه مزیت و معایبی نسبت به همدیگر دارند؟

سؤال ۳ (۸ نمره)

به سوالات زیر در رابطه با نرمالسازی دسته‌ای[۱۳] پاسخ دهید.

الف) فرمول زیر را در نظر بگیرید و توضیح دهید اگر به جای y، $\hat{x}$ را درنظر بگیریم چه مشکلی بوجود می‌آید؟

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i$$
$$\sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2$$
$$\widehat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$
$$y_i = \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$$

ب) توضیح دهید زمانی که سایز دسته‌ها کوچک باشد، استفاده از نرمالسازی دسته‌ای چه تأثیری در فرآیند آموزش دارد؟ استفاده از نرمالسازی دسته‌ای با اندازه دسته ۱ و مقدار اولیه شیفت نرمالسازی دسته‌ای برابر با صفر، چه مشکلی دارد؟

ج) شما در حال اعمال نرمالسازی دسته‌ای یک لایه کاملا متصل با اندازه ورودی ۱۰ و اندازه خروجی۲۰ هستید. با احتساب پارامترهای نرمالسازی دسته‌ای، این لایه چند پارامتر آموزشی دارد؟

د) نرمالسازی لایه‌ای[۱۴] تکنیک نرمالسازی دیگری است که برای غلبه بر معایب نرمالسازی دسته‌ای طراحی شده است. فرمول نرمالسازی لایه‌ای به صورت زیر است:

$$\hat{x} = \frac{x - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}} \qquad y = \gamma\hat{x} + \beta$$

$$\mu_L = \frac{1}{d}\sum_{j=1}^{d} x_j \qquad \sigma_L = \sqrt{\frac{1}{d}\sum_{j=1}^{d}(x_j - \mu_L)^2}$$

با توجه به تعریف نرمالسازی لایه‌ای، بزرگ‌ترین تفاوت بین نرمالسازی دسته‌ای و لایه‌ای چیست؟ در چه شرایطی نرمالسازی دسته‌ای نسبت به نرمالسازی لایه‌ای ترجیح داده می‌شود؟ و برعکس آن چطور؟ چرا؟

[13] Batch Normalization
[14] Layer Normalization

سؤال ٤ (٦ نمره)

در مورد دراپ‌اوت[15] به موارد زیر پاسخ دهید.

الف) توضیح دهید چرا دراپ اوت مانند منظم‌ساز عمل می‌کند؟
ب) آیا می‌توان گفت دراپ‌اوت عملکردی شبیه یادگیری جمعی[16] دارد؟
ج) یک شبکه عصبی با N نود را در نظر بگیرید که هر کدام از نودها می‌توانند در طول آموزش بصورت مستقل با احتمال $0 < p < 1$ حذف شوند. تعداد کل مدل‌های منحصر به فردی که می‌توان با اعمال دراپ‌اوت تحقق بخشید، چقدر است؟
د) استفاده از دراپ‌اوت هنگام آموزش و آزمایش چه تفاوتی با هم دارد؟ چرا؟

سؤالات عملی — 15 + 65 نمره

سوالات عملی در نوتبوک‌های پیوست شده به تمرین قابل مشاهده هست.

سالم و موفق باشید.

[15] Dropout
[16] Ensemble Learning