دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

پاییز و زمستان ۱۴۰۴

استاد: علی شریفی زارچی
گردآورندگان: گرامی‌راد - پیمایی - بهزاداصل - کوچک‌نیا - رستمیان

تمرین سوم | شبکه های عصبی | مهلت ارسال: ۲۶ آذر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- درنظر داشته باشید ددلاین این تمرین شب قبل از آزمون میان‌ترم می‌باشد و امکان استفاده از تاخیر مجاز برای این تمرین وجود ندارد.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- 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.

سوالات نظری (۱۰۰ نمره)

۱. (۸ نمره)

بخش اول: گزاره‌های زیر صحیح است یا خیر؟ مختصر توضیح دهید.

آ. اگر به تعداد کافی پرسپترون و لایه داشته باشیم، می‌توانیم هر تابع پیوسته‌ای را با دقت دلخواه تقریب بزنیم.

ب. تابع هزینه در یک شبکه عصبی عمیق، به دلیل وجود توابع فعال‌سازی غیرخطی و ساختار چند لایه، همواره یک تابع غیرمحدب است.

پ. یک شبکه عصبی که تنها یک لایه دارد و از تابع فعال‌سازی Sigmoid برای خروجی، به همراه تابع هزینه Cross-entropy برای دسته‌بندی دودویی استفاده می‌کند، کاملاً با مدل Logistic regression یکسان است.

ت. پارامترهای گاما (γ) و بتا (β) در لایه نرمال‌سازی دسته‌ای، هایپرپارامتر هستند که باید قبل از شروع آموزش به صورت دستی تنظیم شوند و در حین Backpropagation به‌روزرسانی نمی‌شوند.

بخش دوم: به سوالات زیر پاسخ کوتاه دهید.

آ. چرا مقداردهی اولیه صفر برای تمام وزن‌های یک شبکه عصبی ایده بدی است و باعث چه مشکلی در فرآیند یادگیری می‌شود؟

ب. چرا توابع فعال‌سازی در لایه‌های مخفی یک شبکه عصبی باید غیرخطی باشند؟ اگر همه آن‌ها خطی باشند چه اتفاقی می‌افتد؟

پ. مکانیزم بهینه‌ساز «مومنتوم» (Momentum) را با روش «کاهش شیب تصادفی» (SGD) استاندارد مقایسه کنید. مومنتوم برای حل چه مشکلی در فرآیند بهینه‌سازی معرفی شده است؟

ت. روش «حذف تصادفی»[۱] چگونه کار می‌کند و نقش اصلی آن به عنوان یک روش تنظیم‌گری[۲] چیست؟ چرا این روش فقط در مرحله آموزش فعال است و در مرحله آزمون باید غیرفعال شود؟

[۱] Dropout
[۲] Regularization

۲. (۱۲ نمره) یک شبکه عصبی با ۲ ورودی، ۲ نورون در لایه مخفی[۳] با تابع فعال‌سازی ReLU و یک نورون خروجی با تابع فعال‌سازی Sigmoid داده شده است. پارامترها به صورت زیر هستند:

$$W^{(1)} = \begin{bmatrix} ۱ & -۱ \\ ۰/۵ & ۰/۵ \end{bmatrix}, \quad b^{(1)} = \begin{bmatrix} ۰ \\ ۰/۵ \end{bmatrix}, \quad W^{(2)} = \begin{bmatrix} ۱ & -۱/۵ \end{bmatrix}, \quad b^{(2)} = ۰/۲$$

برای ورودی $x = \begin{bmatrix} ۲ \\ -۱ \end{bmatrix}$ و برچسب $y = ۱$ ، موارد زیر را انجام دهید:

الف. یک مرحله پیش‌روی کامل[۴] را اجرا کرده و مقادیر $a^{(1)}$، h، $z^{(2)}$ و $\hat{y}$ را به دست آورید. (در اینجا h خروجی فعال‌شده لایه مخفی است).

ب. با استفاده از تابع هزینه[۵] Cross-Entropy که در زیر تعریف شده است، گرادیان‌های هزینه نسبت به تمام پارامترها را محاسبه کرده و مقادیر عددی آن‌ها را به دست آورید.

$$\mathcal{L}(\hat{y}, y) = -[y \ln \hat{y} + (۱ - y) \ln(۱ - \hat{y})]$$

گرادیان‌های مورد نیاز: $\frac{\partial \mathcal{L}}{\partial W^{(1)}}$، $\frac{\partial \mathcal{L}}{\partial b^{(1)}}$، $\frac{\partial \mathcal{L}}{\partial W^{(2)}}$، $\frac{\partial \mathcal{L}}{\partial b^{(2)}}$

پ. (۱) تعداد کل پارامترهای قابل آموزش برای یک شبکه عصبی تماماً متصل[۶] با n نورون در لایه ورودی، یک لایه مخفی با k نورون، و m نورون در لایه خروجی را به صورت یک فرمول کلی بنویسید.
(۲) با استفاده از فرمول به دست آمده، تعداد کل پارامترهای قابل آموزش برای شبکه تعریف شده در این سوال را محاسبه کنید.

ت. از لحاظ مفهومی توضیح دهید که افزودن یک لایه مخفی جدید به یک شبکه عصبی، چه تأثیری بر ظرفیت مدل می‌گذارد و چرا این کار احتمال بیش‌برازش[۷] را افزایش می‌دهد؟

۳. (۲۵ نمره) تابع سافت‌مکس (Softmax) این خاصیت مطلوب را دارد که یک توزیع احتمال خروجی می‌دهد و اغلب به عنوان تابع فعال‌سازی در بسیاری از شبکه‌های دسته‌بندی چندکلاسه استفاده می‌شود. یک شبکه عصبی ۲ـ لایه برای دسته‌بندی Kـ کلاسه، با استفاده از فعال‌سازی سافت‌مکس و تابع هزینه Cross-Entropy مطابق تعریف زیر در نظر بگیرید:

$$\begin{aligned} z^{[۱]} &= W^{[۱]}x + b^{[۱]} \\ a^{[۱]} &= \text{LeakyReLU}(z^{[۱]}, \alpha = ۰/۰۱) \\ z^{[۲]} &= W^{[۲]}a^{[۱]} + b^{[۲]} \\ \bar{y} &= \text{softmax}(z^{[۲]}) \\ L &= -\sum_{i=۱}^{K} y_i \log(\bar{y}_i) \end{aligned}$$

که در آن مدل ورودی x به شکل $D_x \times ۱$ و برچسب One-hot ($y \in \{۰, ۱\}^K$) را می‌گیرد. فرض کنید لایه پنهان دارای D_a گره است، یعنی $z^{[۱]}$ یک بردار با اندازه $D_a \times ۱$ است. به یاد بیاورید که تابع سافت‌مکس به صورت زیر محاسبه می‌شود:

$$\hat{y} = \begin{bmatrix} \exp(z_۱^{[۲]})/Z \\ \vdots \\ \exp(z_K^{[۲]})/Z \end{bmatrix}, \quad Z = \sum_{j=۱}^{K} \exp(z_j^{[۲]})$$

---

[۳] Hidden Layer
[۴] Pass Forward
[۵] Loss Function
[۶] Fully Connected Neural Network
[۷] Overfit

آ. شکل (ابعاد) $W^{[۲]}$ و $b^{[۲]}$ چیست؟ اگر بر روی m مثال برداری‌سازی می‌کردیم، یعنی از یک دسته[۸] نمونه $X \in \mathbb{R}^{D_x \times m}$ به عنوان ورودی استفاده می‌کردیم، شکل خروجی لایه پنهان چه می‌بود؟

ب. مشتق $\frac{\partial \hat{y}_k}{\partial z_k^{[۲]}}$ چیست؟ پاسخ خود را بر حسب مؤلفه(های) $\hat{y}$ ساده کنید.

پ. به ازای $i \neq k$، مشتق $\frac{\partial \hat{y}_k}{\partial z_i^{[۲]}}$ چیست؟ پاسخ خود را بر حسب مؤلفه(های) $\hat{y}$ ساده کنید.

ت. فرض کنید برچسب y درایه kـام خود مقدار ۱ و در سایر درایه‌ها مقدار ۰ دارد. $\frac{\partial L}{\partial z_i^{[۲]}}$ چیست؟ پاسخ خود را بر حسب $\hat{y}_i$ ساده کنید. نکته: هر دو حالت $i = k$ و $i \neq k$ را در نظر بگیرید.

ث. مشتق $\frac{\partial z^{[۲]}}{\partial a^{[۱]}}$ چیست؟ این نتیجه را با نماد $\delta_۱$ نشان دهید.

ج. مشتق $\frac{\partial a^{[۱]}}{\partial z^{[۱]}}$ چیست؟ این نتیجه را با نماد $\delta_۲$ نشان دهید. می‌توانید از نماد آکولاد (تابع چندضابطه‌ای) استفاده کنید.

چ. مشتق $\frac{\partial L}{\partial z^{[۲]}}$ را با δ. نشان دهید. $\frac{\partial L}{\partial b^{[۱]}}$ و $\frac{\partial L}{\partial W^{[۱]}}$ چیست؟ می‌توانید از نمادهای بخش‌های قبلی استفاده مجدد کنید. نکته: به اشکال (ابعاد) دقت کنید.

ح. برای جلوگیری از مشکلات پایداری عددی، می‌توان از یک ترفند هنگام پیاده‌سازی تابع سافت‌مکس استفاده کرد. فرض کنید $m = \max_{j=۱}^{K} z_j^{[۲]}$ حداکثر مقادیر $z_i^{[۲]}$ باشد، سپس $\hat{y}_i$ به صورت زیر محاسبه می‌شود:

$$\hat{y}_i = \frac{\exp(z_i^{[۲]} - m)}{\sum_{j=۱}^{K} \exp(z_j^{[۲]} - m)}.$$

مشکل عددی در محاسبه اولیه سافت‌مکس چیست؟ چرا فرمول تغییر یافته به حل آن مشکل کمک می‌کند؟

خ. آیا اصلاح انتقال‌ـحداکثر (max-shift) در بخش آخر، مشتق نسبت به $z^{[۲]}$ را تغییر می‌دهد؟

۴. (۱۲ نمره) یک شبکه عصبی برای پیاده‌سازی تابع XOR برای چهار ورودی دودویی $x_۱, x_۲, x_۳, x_۴ \in \{۰, ۱\}$ طراحی شده است. این شبکه از توابع فعال‌سازی زیر استفاده می‌کند:

$$\text{ReLU}(z) = \max(۰, z) \quad , \quad \sigma(z) = \frac{۱}{۱ + e^{-z}}$$

خروجی به عنوان ۱ در نظر گرفته می‌شود اگر $y > ۰/۵$ و در غیر این صورت ۰ است. تابع هزینه، خطای مربع است:

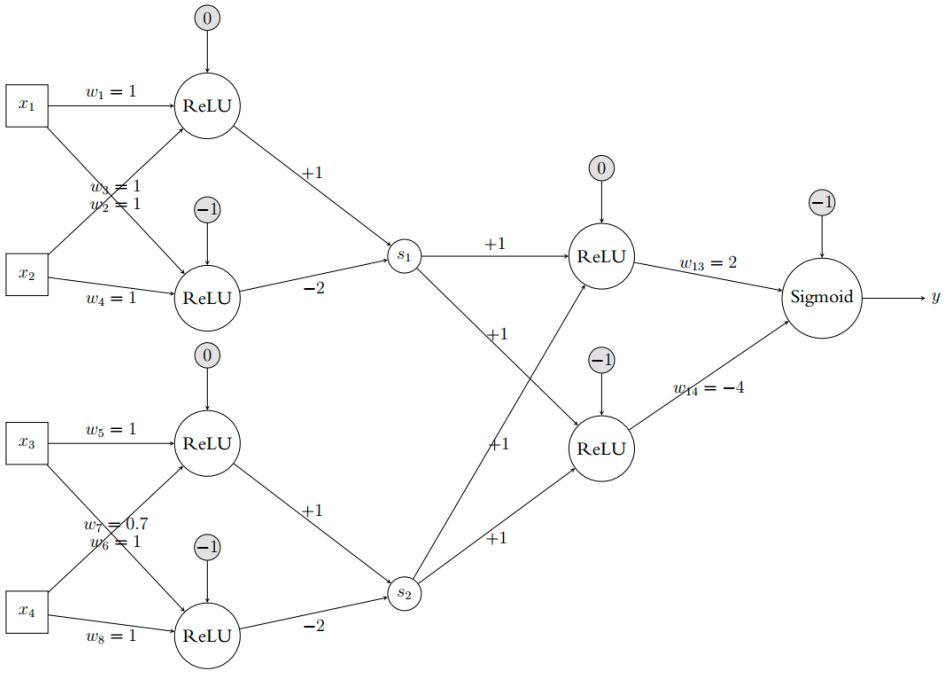
$$L(x) = (y_{\text{هدف}} - y)^۲.$$

ساختار شبکه

$$\begin{aligned}
h_۱ &= \text{ReLU}(x_۱ + x_۲), \\
h_۲ &= \text{ReLU}(x_۱ + x_۲ - ۱), \\
h_۳ &= \text{ReLU}(x_۳ + x_۴), \\
h_۴ &= \text{ReLU}(۰/۷x_۳ + x_۴ - ۱). \\
s_۱ &= h_۱ - ۲h_۲, \\
s_۲ &= h_۳ - ۲h_۴. \\
h_۵ &= \text{ReLU}(s_۱ + s_۲), \\
h_۶ &= \text{ReLU}(s_۱ + s_۲ - ۱). \\
y &= \sigma(۲h_۵ - ۴h_۶ - ۱).
\end{aligned}$$

[۸] Batch

نمودار شبکه



شکل ۱ : نمودار شبکه عصبی مورد استفاده در سوال

آ. تمام بردارهای ورودی $x = (x_۱, x_۲, x_۳, x_۴)$ که در آن مدل پیش‌بینی نادرست انجام می‌دهد (یعنی $y_{\text{مدل}} \neq y_{\text{هدف}}$) را شناسایی کنید.

ب. یک گام SGD روی وزن $w_۷ = ۰/۷$ با استفاده از بردار ورودی $(x_۱, x_۲, x_۳, x_۴) = (۰, ۰, ۱, ۱)$ انجام دهید. نرخ یادگیری $\eta = ۰/۱$ را فرض کنید. مقدار جدید $w_۷$ را تعیین کنید.

۵. (۲۰ نمره) ماتریس طراحی $\mathbf{X}$ (که سطر iـام آن نقطه نمونه $\mathbf{x}_i^T$ است) و یک بردار nـتایی از برچسب‌ها $\mathbf{y} = [y_1, \ldots, y_n]^T$ به شما داده شده است. برای سادگی، فرض کنید $\mathbf{X}$ سفید (whitened) شده است، به طوری که $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$. جمله بایاس یا بعد اضافی اضافه نکنید؛ برای ورودی صفر، خروجی همیشه صفر است. فرض کنید $\mathbf{x}_{*i}$ ستون iـام ماتریس $\mathbf{X}$ را نشان می‌دهد.

آ. نشان دهید که تابع هزینه برای کمترین مربعات با تنظیم‌گر ℓ_1، (Lasso) یعنی:

$$J_1(\mathbf{w}) \triangleq \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1 \quad (\text{که در آن } \lambda > 0)$$

می‌تواند به صورت زیر بازنویسی شود:

$$J_1(\mathbf{w}) = \|\mathbf{y}\|^2 + \sum_{i=1}^{d} f(\mathbf{x}_{*i}, w_i)$$

که در آن $f(\cdot, \cdot)$ یک تابع مناسب است که آرگومان اول آن یک بردار و آرگومان دوم آن یک اسکالر است.

ب. با استفاده از پاسخ خود به بخش (آ)، شرایط لازم و کافی برای مولفه iـام بهینه‌ساز $\mathbf{w}^*$ تابع $J_1(\cdot)$ را استخراج کنید به طوری که هر یک از سه خاصیت زیر را برآورده کند: $w_i^* > 0$، $w_i^* = 0$ و $w_i^* < 0$.

پ. برای بهینه‌ساز $\mathbf{w}^\#$ تابع هزینه کمترین مربعات با تنظیم‌گر ℓ_2، (Ridge) یعنی:

$$J_2(\mathbf{w}) \triangleq \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_2^2 \quad (\text{که در آن } \lambda > 0)$$

یک شرط لازم و کافی برای $w_i^\# = 0$ استخراج کنید، که در آن $w_i^\#$ مولفه iـام بردار $\mathbf{w}^\#$ است.

ت. یک بردار را تُنُک[۹] می‌نامیم اگر بیشتر مولفه‌های آن صفر باشند. بر اساس پاسخ‌های خود به بخش‌های (ب) و (پ)، کدام یک از بردارهای $\mathbf{w}^*$ و $\mathbf{w}^\#$ به احتمال زیاد تُنُک‌تر است؟ چرا؟

۶. (۱۰ نمره) تابع هزینه یک مدل $J(\theta)$ دارای حداقل‌های محلی است و می‌خواهیم از الگوریتم کاهش شیب تصادفی (SGD) برای بهینه‌سازی پارامترهای آن (θ) استفاده کنیم. فرض کنید از یک نرخ یادگیری با کاهش نمایی[۱۰] به شکل زیر استفاده می‌کنیم:

$$\alpha_t = \alpha_0 e^{-kt}$$

که در آن t شماره گام (epoch) است. نرخ یادگیری اولیه $\alpha_0 = 0.1$ و ضریب کاهش $k = 0.05$ داده شده است.

آ. قاعده به‌روزرسانی پارامترها (θ) را با استفاده از SGD بنویسید.

ب. به صورت مفهومی توضیح دهید که چرا استفاده از نرخ یادگیری متغیر با کاهش[۱۱] می‌تواند هم به سرعت همگرایی و هم به پایداری فرآیند آموزش کمک کند.

پ. اگر الگوریتم در یک ناحیه با شیب تند از تابع هزینه قرار گیرد، رفتار آن را با نرخ یادگیری ثابت در مقایسه با نرخ یادگیری دارای کاهش تحلیل و مقایسه کنید. این دو رویکرد چه تفاوتی در مسیر رسیدن به نقطه بهینه دارند؟

---

[۹] sparse
[۱۰] exponential decay
[۱۱] decaying learning rate

ت. یکی از شرایط کلاسیک برای تضمین همگرایی SGD (تحت فروض مشخص)، شرایط رابینز-مونرو[۱۲] برای نرخ یادگیری α_t است:

$$\sum_{t=۱}^{\infty} \alpha_t = \infty \quad \text{و} \quad \sum_{t=۱}^{\infty} \alpha_t^۲ < \infty$$

بررسی کنید که آیا نرخ یادگیری با کاهش نمایی ($\alpha_t = \alpha_۰ e^{-kt}$) در این شرایط صدق می‌کند یا خیر. برای تحلیل می‌توانید از تقریب انتگرال برای سری‌ها استفاده کنید. نتیجه‌ای که به دست می‌آورید چه مفهومی در مورد تضمین همگرایی این نوع نرخ یادگیری دارد؟

۷. (۱۳ نمره) فرض کنید یک لایه نرمال‌سازی دسته‌ای[۱۳] یک‌بعدی روی یک Mini-batch با اندازه m اعمال می‌شود. ورودی‌های اسکالر برای یک نورون خاص را با $x_۱, \ldots, x_m$ نمایش می‌دهیم. لایه BN آماره‌ها و خروجی‌ها را به صورت زیر محاسبه می‌کند:

$$\begin{aligned}
\mu_B &= \frac{۱}{m} \sum_{i=۱}^{m} x_i \\
\sigma_B^۲ &= \frac{۱}{m} \sum_{i=۱}^{m} (x_i - \mu_B)^۲ \\
\hat{x}_i &= \frac{x_i - \mu_B}{\sqrt{\sigma_B^۲ + \epsilon}} \\
y_i &= \gamma \hat{x}_i + \beta
\end{aligned}$$

که در آن γ و β پارامترهای قابل یادگیری هستند و ϵ یک عدد کوچک ثابت برای پایداری عددی است. تابع هزینه L تابعی از خروجی‌های $y_۱, \ldots, y_m$ است و گرادیان نسبت به خروجی‌ها $g_i := \frac{\partial L}{\partial y_i}$ می‌باشد.

آ. با استفاده از قاعده زنجیره‌ای، فرمول بسته‌ای برای گرادیان‌های $\frac{\partial L}{\partial \gamma}$ و $\frac{\partial L}{\partial \beta}$ بر حسب g_i و $\hat{x}_i$ به دست آورید.

ب. گرادیان $\frac{\partial L}{\partial x_i}$ را برای هر i محاسبه کنید. توجه داشته باشید که μ_B و $\sigma_B^۲$ توابعی از تمام x_jها هستند. پاسخ نهایی باید تنها بر حسب x_i، μ_B، $\sigma_B^۲$، ϵ، γ و g_j باشد. هدف رسیدن به رابطه زیر است:

$$\frac{\partial L}{\partial x_i} = \frac{\gamma}{\sqrt{\sigma_B^۲ + \epsilon}} \left(g_i - \frac{۱}{m} \sum_{j=۱}^{m} g_j - \frac{\hat{x}_i}{m} \sum_{j=۱}^{m} g_j \hat{x}_j \right)$$

راهنمایی: ابتدا $\frac{\partial L}{\partial \hat{x}_i}$ را بر حسب g_i و γ بنویسید، سپس گرادیان‌ها را از طریق $\hat{x}_i$، μ_B و $\sigma_B^۲$ به عقب منتقل کنید.

پ. با استفاده از رابطه‌ای که در قسمت (ب) به دست آوردید، به صورت تحلیلی ثابت کنید که:

$$\sum_{i=۱}^{m} \frac{\partial L}{\partial x_i} = ۰$$

در یک یا دو جمله توضیح دهید که این ویژگی چه مفهوم شهودی‌ای در مورد رفتار لایه BN نسبت به شیفت دادن تمام ورودی‌های دسته با یک مقدار ثابت (مثلاً $x_i \leftarrow x_i + c$) دارد.

---

[۱۲] Robbins-Monro conditions
[۱۳] Batch Normalization (BN)