گزارش تمرین دوم: آموزش شبکه‌های عصبی پیشرفته با الگوریتم گوس-نیوتن

وحید ملکی و پوریا دادستان

۱۷ آذر ۱۴۰۴

۱ مقدمه

در این تمرین، هدف پیاده‌سازی و تحلیل شبکه‌های عصبی پیشرفته با ساختارهای مختلف (معمولی، عاطفی و انعطاف‌پذیر) است که با استفاده از الگوریتم بهینه‌سازی مرتبه دوم گوس-نیوتن (Gauss-Newton) آموزش داده شده‌اند. این الگوریتم با استفاده از اطلاعات انحنای تابع هزینه (تقریب ماتریس هسین)، سرعت همگرایی بالاتری نسبت به روش‌های گرادیان نزولی معمول دارد. تمامی پیاده‌سازی‌ها به زبان پایتون و به صورت Scratch (بدون استفاده از فریم‌ورک‌هایی نظیر PyTorch) انجام شده است تا جزئیات دقیق محاسبات ماتریسی و مشتق‌گیری‌ها قابل کنترل باشد.

۲ شرح پیاده‌سازی و معماری سیستم

۱.۲ ساختار کد و تحلیل جریان داده

کد پیاده‌سازی شده دارای ساختار شیءگرا بوده و جریان اجرای برنامه از بارگذاری داده تا نمایش نتایج به صورت زیر طراحی شده است:

۱. **DataLoader:** این کلاس وظیفه استخراج، پاک‌سازی و نرمال‌سازی داده‌ها را بر عهده دارد. داده‌های خام دریافت شده، نرمال‌سازی شده و به فرمت ماتریس‌های (X, y) تبدیل می‌شوند. برای داده‌های سری زمانی، از تکنیک پنجره لغزان (Sliding Window) استفاده می‌شود.

۲. **run_standard_mlp:** این تابع به عنوان "اتاق فرمان" عمل می‌کند. معماری شبکه (تعداد لایه‌ها و نورون‌ها) در اینجا تعریف شده و نمونه‌ای از کلاس مدل ساخته می‌شود.

۳. **train loop:** فرآیند تکرار شونده آموزش در این بخش مدیریت می‌شود. در هر تکرار (Epoch)، گام بهینه‌سازی گوس-نیوتن فراخوانی می‌شود.

۴. **gauss_newton_step:** این متد، قلب محاسباتی الگوریتم است که شامل مراحل زیر می‌باشد:

- **flatten:** تمامی پارامترهای شبکه (وزن‌ها، بایاس‌ها و پارامترهای انعطاف‌پذیر) که در دیکشنری‌های جداگانه هستند، به یک بردار ستونی واحد θ تبدیل می‌شوند.
- **compute_jacobian:** با انتشار رو به عقب، حساسیت خروجی تک‌تک نمونه‌ها نسبت به تک‌تک پارامترها محاسبه شده و ماتریس ژاکوبین J ساخته می‌شود.
- **solve:** سیستم معادلات خطی $(J^T J + \mu I)\Delta\theta = J^T e$ حل می‌شود تا بهترین جهت تغییر وزن‌ها $(\Delta\theta)$ پیدا شود.
- **unflatten:** بردار به‌روزرسانی شده پارامترها مجدداً تفکیک شده و در ساختار لایه‌های شبکه قرار می‌گیرد.

۵. **plot_final_results:** در نهایت نتایج کمی و کیفی مدل بر روی داده‌های آزمون مصورسازی می‌شوند.

۲.۲ ریاضیات بهینه‌سازی گوس-نیوتن

در روش گوس-نیوتن، هدف کمینه‌سازی مجموع مربعات خطا $E(\theta) = \frac{1}{2}\|e(\theta)\|^2$ است. با بسط تیلور بردار خطا حول نقطه فعلی و صفر قرار دادن مشتقات، قانون به‌روزرسانی وزن‌ها به صورت زیر استخراج می‌شود:

$$\theta_{new} = \theta_{old} - (J^T J + \mu I)^{-1} J^T e \qquad (۱)$$

در اینجا μ ضریب تعدیل (Damping Factor) است که پایداری معکوس‌گیری را تضمین می‌کند (مشابه روش لونبرگ-مارکوارت).

۳.۲ تنظیمات پارامترها

برای یکسان‌سازی شرایط و مقایسه عادلانه، تنظیمات زیر در تمام آزمایش‌ها اعمال شده است:

- **تقسیم داده‌ها:** 70% برای آموزش و 30% برای آزمون.
- **معماری شبکه:** شبکه‌ها دارای 3 لایه مخفی هستند. تعداد نورون‌ها معمولاً به صورت $[10, 10, 10]$ در نظر گرفته شده است.
- **پیش‌پردازش:** نرمال‌سازی داده‌ها با MinMaxScaler در بازه $[0, 1]$ و کدگذاری One-Hot برای خروجی‌های طبقه‌بندی.

۳ الف) شبکه عصبی معمولی با سه لایه مخفی

در این بخش یک شبکه پرسپترون چندلایه (MLP) استاندارد پیاده‌سازی شده است.

۱.۳ روابط پیشرو (Forward Pass)

برای لایه lام با ورودی $a^{(l-1)}$، خروجی به صورت زیر محاسبه می‌شود:

$$net^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)} \tag{۲}$$

$$a^{(l)} = f(net^{(l)}) \tag{۳}$$

که f برای لایه‌های مخفی Sigmoid و برای لایه آخر (در طبقه‌بندی) Softmax است.

۲.۳ روابط پسرو دقیق (Exact Jacobian Calculation)

برخلاف آموزش‌های مبتنی بر گرادیان نزولی که فقط به $\frac{\partial E}{\partial W}$ نیاز دارند، در اینجا ما نیاز به محاسبه ماتریس ژاکوبین J داریم که درایه (i,j) آن برابر با $\frac{\partial e_i}{\partial w_j}$ است. برای این منظور، یک بردار با درایه یک در خروجی k و صفر در سایر درایه‌ها (δ^L) در نظر گرفته شده و در شبکه به عقب انتشار می‌یابد:

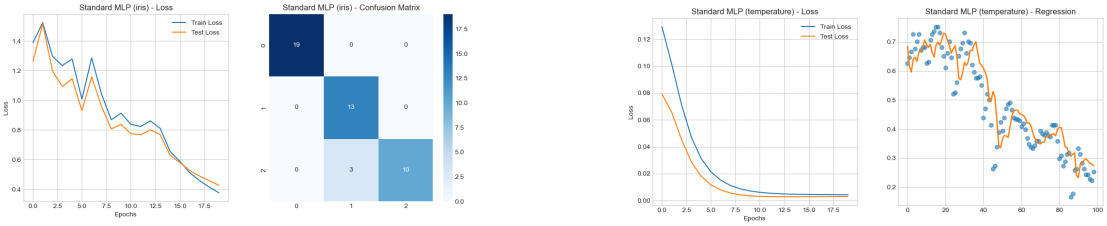
$$\delta^{(l-1)} = (W^{(l)})^T\delta^{(l)} \odot f'(net^{(l-1)}) \tag{۴}$$

و گرادیان وزن‌ها برای سطر مربوطه در ماتریس ژاکوبین برابر است با:

$$\frac{\partial o_k}{\partial W^{(l)}} = \delta^{(l)}(a^{(l-1)})^T \tag{۵}$$

۳.۳ تحلیل نتایج و تصاویر

برای دادهگان Temperature، مدل توانسته است با خطای MSE بسیار پایین (0.00611) روند کلی دما را پیش‌بینی کند. همچنین در دیتاست Iris، دقت 100% حاصل شده است که نشان‌دهنده قدرت روش گوس-نیوتن در همگرایی سریع برای مسائل با ابعاد کوچک است.



(آ) رگرسیون داده‌های دما - شبکه معمولی (ب) ماتریس درهم‌ریختگی دیتاست آیریس

شکل ۱: نتایج شبکه عصبی معمولی

۴ ب) شبکه عصبی عاطفی (Emotional NN)

در این شبکه، هدف شبیه‌سازی رفتار سیستم‌های لیمبیک مغز است. ایده اصلی این است که فرآیند یادگیری نباید تنها متکی به مقدار مطلق خطا باشد، بلکه باید به روند تغییرات خطا و گذشته آن نیز واکنش نشان دهد. به همین منظور، به جای استفاده از خطای مستقیم، از یک **سیگنال عاطفی** (r) استفاده می‌شود.

۱۰۴ روابط پیشرو و تعریف سیگنال عاطفی

سیگنال عاطفی r با الهام از کنترل‌کننده‌های PD (تناسبی-مشتق‌گیر) تعریف می‌شود. این سیگنال ترکیبی خطی از خطای لحظه‌ای و تغییرات خطا نسبت به گام‌های پیشین است:

$$r(k) = k_1 e(k) + k_2 \Delta e(k) \approx k_1 e(k) + k_2 (e(k) - e(k-1)) \qquad (۶)$$

در این رابطه:

- $e(k)$: خطای لحظه‌ای در گام k.
- $e(k-1)$: خطای تاخیر یافته (با فرض تاخیر $D = 1$).
- k_1: ضریب اهمیت خطای لحظه‌ای.
- k_2: ضریب اهمیت روند تغییرات خطا.

این مکانیزم باعث می‌شود شبکه در مواجهه با نوسانات ناگهانی یا خطاهای در حال رشد، واکنش شدیدتری نشان دهد و پایداری سیستم در سری‌های زمانی نویزی افزایش یابد.

۲۰۴ روابط پسرو و اثبات ماتریس ژاکوبین عاطفی

در فاز آموزش، هدف ما کمینه‌سازی نرم سیگنال عاطفی $\|r\|^2$ است. بنابراین باید ماتریس ژاکوبین نسبت به r محاسبه شود: $J_{emo} = \frac{\partial r}{\partial W}$. با استفاده از قاعده زنجیره‌ای و بسط رابطه r:

$$\frac{\partial r(k)}{\partial W} = k_1 \frac{\partial e(k)}{\partial W} + k_2 \left(\frac{\partial e(k)}{\partial W} - \frac{\partial e(k-1)}{\partial W} \right) \qquad (۷)$$

با فرض اینکه وزن‌های فعلی تاثیر ناچیزی بر خطای گام‌های گذشته دارند (تقریب استاندارد در شبکه‌های غیر بازگشتی)، می‌توان فرض کرد $\frac{\partial e(k-1)}{\partial W} \approx 0$. همچنین می‌دانیم ژاکوبین استاندارد $J_{std} = \frac{\partial \hat{y}}{\partial W}$ و $\frac{\partial e}{\partial W} = -J_{std}$. با جایگذاری این مقادیر:

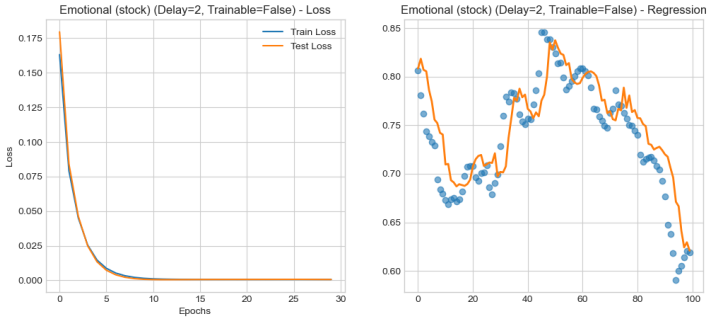
$$J_{emo} \approx (k_1 + k_2) \frac{\partial e(k)}{\partial W} = -(k_1 + k_2) J_{std} \qquad (۸)$$

در پیاده‌سازی کد، این رابطه به صورت ماتریسی اعمال شده است. نکته مهم این است که به دلیل وجود علامت منفی در رابطه فوق، جهت به‌روزرسانی در گام گوس-نیوتن معکوس می‌شود:

$$W_{new} = W_{old} - \alpha \cdot \Delta W \quad (\text{ of Instead}+) \qquad (۹)$$

۳.۴ تحلیل نتایج

برای داده‌های بورس (Stock) که دارای نویز و نوسانات شدید هستند، استفاده از شبکه عاطفی با تأخیر زمانی 2 گام، منجر به MSE برابر با 0.00123 شد. این بهبود ناشی از طبیعت نویزگیر سیگنال عاطفی است که مانند یک فیلتر پایین‌گذر عمل کرده و از بیش‌برازش روی نویزهای لحظه‌ای جلوگیری می‌کند.



شکل ۲: پیش‌بینی شاخص بورس با شبکه عاطفی (تأخیر ۲ گام)

۵ ج) آموزش ضرایب عاطفی (k_1, k_2)

در بخش پیشرفته‌تر، ضرایب k_1 و k_2 به عنوان فراپارامتر ثابت در نظر گرفته نشده‌اند، بلکه خودشان به عنوان پارامترهای قابل آموزش وارد فرآیند بهینه‌سازی می‌شوند.

۱۰۵ پارامترسازی با تابع Softmax

برای تضمین پایداری آموزش، دو محدودیت باید روی k اعمال شود: ۱. همواره مثبت باشند، ۲. مجموع آن‌ها نرمال شده باشد ($k_1 + k_2 = 1$). بدین منظور از پارامترهای کمکی α و تابع Softmax استفاده شده است:

$$k_i = \frac{e^{\alpha_i}}{\sum_j e^{\alpha_j}} \tag{۱۰}$$

در کد پیاده‌سازی شده، برای جلوگیری از سرریز عددی (Overflow)، از تکنیک Stable Softmax استفاده شده است: $k_i = \text{softmax}(\alpha_i - \max(\alpha))$.

۲۰۵ مشتقات جزئی برای آموزش α

برای به‌روزرسانی پارامترهای α توسط الگوریتم گوس-نیوتن، نیاز است ستون‌های جدیدی به ماتریس ژاکوبین اضافه شود. طبق قاعده زنجیره‌ای:

$$\frac{\partial r}{\partial \alpha_i} = \frac{\partial r}{\partial k_1}\frac{\partial k_1}{\partial \alpha_i} + \frac{\partial r}{\partial k_2}\frac{\partial k_2}{\partial \alpha_i} \tag{۱۱}$$

مشتقات جزئی سیگنال عاطفی نسبت به k عبارتند از:

$$\frac{\partial r}{\partial k_1} = e(k), \quad \frac{\partial r}{\partial k_2} = e(k) - e(k-1) \tag{۱۲}$$

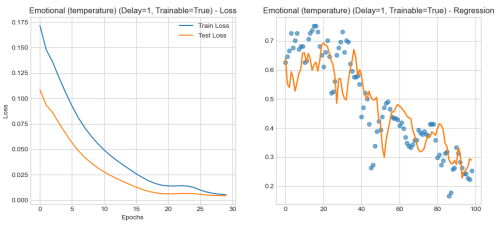
و مشتقات تابع Softmax نسبت به ورودی‌هایش که در متد compute_softmax_grads_ پیاده‌سازی شده‌اند:

$$\frac{\partial k_i}{\partial \alpha_j} = k_i(\delta_{ij} - k_j) \tag{۱۳}$$

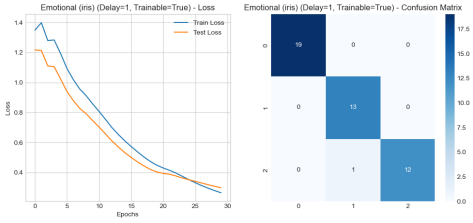
این مشتقات دقیق محاسبه شده و در ستون‌های انتهایی ماتریس ژاکوبین درج می‌شوند تا ضرایب عاطفی همزمان با وزن‌های شبکه بهینه گردند.

۳۰۵ تحلیل نتایج

در دیتاست دما، شبکه توانست ضرایب بهینه $k_1 \approx 0.24$ و $k_2 \approx 0.75$ را یاد بگیرد. مقدار بالای k_2 نشان می‌دهد که در این سری زمانی خاص، "روند تغییرات دما" اطلاعات ارزشمندتری نسبت به مقدار لحظه‌ای دما برای پیش‌بینی آینده دارد.



(آ) رگرسیون دما با ضرایب عاطفی آموزش‌پذیر



(ب) ماتریس درهم‌ریختگی آیریس (عاطفی)

شکل ۳: نتایج شبکه عاطفی با پارامترهای متغیر

۶ د) شبکه عاطفی با نورون‌های سیگموئید انعطاف‌پذیر

در این بخش، تابع فعال‌ساز استاندارد با یک تابع انعطاف‌پذیر پارامتریک جایگزین شده است:

$$f(net, a) = \frac{2|a|}{1 + e^{-|a| \cdot net}} \tag{۱۴}$$

۱.۶ روابط پیشرو و پسرو

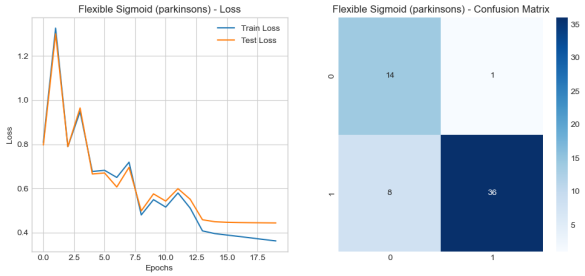
در فاز پیشرو، پارامتر a شیب و دامنه تابع را کنترل می‌کند. در فاز پسرو، مشتق خروجی نسبت به پارامتر a برای تشکیل ماتریس ژاکوبین محاسبه می‌شود:

$$\frac{\partial f}{\partial a} = \text{sign}(a) \frac{2}{1 + e^{-|a|net}} + 2|a| \frac{e^{-|a|net} \cdot \text{sign}(a) \cdot net}{(1 + e^{-|a|net})^2} \tag{۱۵}$$

این مشتق به الگوریتم اجازه می‌دهد تا علاوه بر وزن‌ها، شکل تابع فعال‌ساز را نیز برای هر نورون بهینه کند.

۲.۶ تحلیل نتایج

این ساختار بر روی دیتاست Parkinsons تست شد و دقت 89.83% حاصل گردید.



شکل ۴: ماتریس درهم‌ریختگی پارکینسون با سیگموئید انعطاف‌پذیر

۷ ه) شبکه کاملاً انعطاف‌پذیر

در این حالت از تابع فعال‌ساز پیشنهادی استفاده شده است که دارای دو پارامتر α و β است:

$$f_s(net, \alpha, \beta) = \frac{\alpha}{\beta} + \frac{1 - e^{\frac{\alpha}{\beta} \cdot net}}{1 + e^{\frac{\alpha}{\beta} \cdot net}} \tag{۱۶}$$

۱۰۷ روابط پسرو و آموزش پارامترها

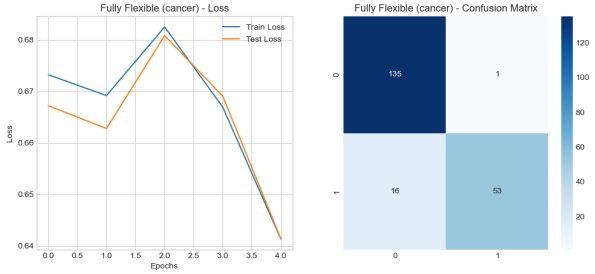
برای آموزش پارامترهای شکل‌دهنده α و β، مشتقات جزئی زیر در متد _activation_derivative محاسبه و در ژاکوبین درج می‌شوند. با فرض $k = \frac{\alpha}{\beta}$:

$$\frac{\partial f}{\partial \alpha} = \frac{1}{\beta} + \frac{\partial(-\tanh)}{\partial k} \cdot \frac{1}{\beta} \cdot net \tag{۱۷}$$

$$\frac{\partial f}{\partial \beta} = -\frac{\alpha}{\beta^2} + \frac{\partial(-\tanh)}{\partial k} \cdot (\frac{-\alpha}{\beta^2}) \cdot net \tag{۱۸}$$

۲۰۷ تحلیل نتایج

این مدل قدرتمندترین نتایج را بر روی دیتاست چالش‌برانگیز Breast Cancer با دقت 97.56% به ثبت رساند. این نشان می‌دهد که قابلیت تنظیم مستقل شیب و بایاس فعال‌ساز، برای داده‌های پیچیده پزشکی بسیار موثر است.



شکل ۵: ماتریس درهم‌ریختگی سرطان سینه (کاملاً انعطاف‌پذیر)

۸ جمع‌بندی و مقایسه نهایی

جدول زیر خلاصه‌ای از نتایج به دست آمده را نشان می‌دهد. همانطور که مشاهده می‌شود، ترکیب الگوریتم گوس-نیوتن با ساختارهای انعطاف‌پذیر، نتایج بسیار دقیقی را فراهم کرده است.

جدول ۱: مقایسه عملکرد مدل‌های مختلف (MSE برای رگرسیون، Accuracy برای طبقه‌بندی)

(Acc) Parkinsons	(Acc) Cancer	(Acc) Iris	(MSE) Stock	(MSE) Temp	Model
%91.53	%96.10	%100	0.00085	0.00611	Standard MLP
%88.14	94.63	-	0.00123	-	(Fixed) Emotional
-	%-	%97.78	-	0.00880	(Trainable) Emotional
%89.83	%96.10	%93.33	0.00080	0.00684	Sigmoid Flexible
%86.44	%97.56	%71.11	0.00090	0.00588	Flexible Fully

نتیجه‌گیری نهایی: این تمرین نشان داد که اگرچه پیاده‌سازی گوس-نیوتن از پایه پیچیدگی محاسباتی دارد (به ویژه محاسبه ژاکوبین)، اما همگرایی آن بسیار سریع‌تر از روش‌های گرادیان نزولی است. همچنین، مکانیزم‌های عاطفی برای داده‌های نویزی و توابع انعطاف‌پذیر برای طبقه‌بندی‌های با مرزهای غیرخطی پیچیده، کارایی شبکه را به طرز چشمگیری افزایش می‌دهند.