

تکلیف دوم درس شبکه های عصبی

وحید ملکی، پوریا دادستان

۱۷ آذر ۱۴۰۴

۱ مقدمه و تعریف مسئله

در این بخش از تمرین، یک معماری نوین برای شبکه عصبی پیشنهاد شده است که همزمان سه ویژگی پیشرفته را داراست:

۱. انعطاف پذیری (Flexibility): استفاده از توابع فعال ساز ترکیبی با پارامترهای یادگیرنده.
 ۲. رفتار عاطفی (Emotional Behavior): استفاده از تاریخچه خطا برای تعدیل فرآیند به روزرسانی وزن ها.
 ۳. آموزش پیشرفته: مقایسه الگوریتم Adam (مرتبه اول) و Levenberg-Marquardt (مرتبه دوم).
- هدف نهایی، بررسی کارایی این ساختار بر روی دو دسته مسئله رگرسیون (سری زمانی) و طبقه بندی (داده های آماری) است.

۲ مدل ریاضی و روابط حاکم

۱.۲ توابع فعال ساز منعطف (F_1 و F_2)

برخلاف شبکه های کلاسیک، در این مدل هر نورون دارای تابع فعال سازی است که ترکیبی خطی از چهار تابع پایه می باشد. پارامترهای شکل دهنده و ضرایب ترکیب، همگی در حین آموزش بهینه می شوند. تابع فعال ساز نوع اول (F_1): کامل ترین فرم (شامل بخش خطی):

$$F_1(x) = \alpha \cdot \sigma(\theta_1 x) + \beta \cdot \tanh(\theta_2 x) + \gamma \cdot \text{ReLU}(\theta_3 x) + \delta \cdot (\theta_4 x) \quad (1)$$

تابع فعال ساز نوع دوم (F_2): فرم فشرده (بدون بخش خطی):

$$F_2(x) = \alpha \cdot \sigma(\theta_1 x) + \beta \cdot \tanh(\theta_2 x) + \gamma \cdot \text{ReLU}(\theta_3 x) \quad (2)$$

در اینجا $\sigma(z) = \frac{1}{1+e^{-z}}$ است. مجموعه پارامترهای آموزش پذیر برای هر نورون عبارتند از: $\Omega = \{\alpha, \beta, \gamma, \delta, \theta_1, \dots, \theta_4\}$

۲.۲ تزریق منطق عاطفی (Emotional Logic)

در شبکه های عصبی استاندارد، هدف کمینه سازی خطای لحظه ای $e(k)$ است. اما در شبکه عاطفی پیشنهادی، یک سیگنال عاطفی $r(k)$ تعریف می شود که به تغییرات خطا نیز حساس است:

$$e(k) = y_{\text{target}}(k) - y_{\text{pred}}(k) \quad (3)$$

$$\Delta e(k) = e(k) - e(k-1) \quad (۴)$$

$$r(k) = k_1 e(k) + k_2 \Delta e(k) \quad (۵)$$

ضرایب k_1 و k_2 میزان اهمیت خطای لحظه‌ای و روند تغییرات خطا را تعیین می‌کنند (در این پیاده‌سازی $k_1 = 0.7, k_2 = 0.3$).

۳.۲ روابط پیشرو (Forward Propagation)

در این شبکه، محاسبات لایه به لایه انجام می‌شود. فرض کنیم $A^{(l-1)}$ بردار ورودی به لایه l ام باشد (که $A^{(0)} = X$ است). خروجی خالص (net) و خروجی فعال‌شده (A) برای هر نورون j در لایه l به صورت زیر محاسبه می‌شود:

$$net_j^{(l)} = \sum_i w_{ji}^{(l)} A_i^{(l-1)} + b_j^{(l)} \quad (۶)$$

سپس تابع فعال‌ساز ترکیبی F_1 (یا F_2) روی net اعمال می‌شود. رابطه کلی برای کامل‌ترین حالت (F_1) عبارت است از:

$$\begin{aligned} A_j^{(l)} = F(net_j^{(l)}, \Omega_j^{(l)}) = & \alpha_j^{(l)} \sigma(\theta_{1,j}^{(l)} net_j^{(l)}) \\ & + \beta_j^{(l)} \tanh(\theta_{2,j}^{(l)} net_j^{(l)}) \\ & + \gamma_j^{(l)} \text{ReLU}(\theta_{3,j}^{(l)} net_j^{(l)}) \\ & + \delta_j^{(l)} (\theta_{4,j}^{(l)} net_j^{(l)}) \end{aligned} \quad (۷)$$

که در آن مجموعه پارامترهای آموزش‌پذیر تابع فعال‌ساز برای نورون j در لایه l است.

۴.۲ روابط پسرو و اصلاح گرادیان (Backpropagation)

هدف فرآیند پسرو، محاسبه گرادیان تابع هزینه نسبت به تمامی پارامترهای شبکه ($\Theta = \{W, b, \Omega\}$) جهت استفاده در الگوریتم LM یا Adam است.

۱.۴.۲ ۱. تابع هزینه عاطفی و قاعده زنجیره‌ای اصلاح‌شده

در شبکه‌های عصبی عاطفی، تابع هزینه بر اساس سیگنال عاطفی r تعریف می‌شود، نه صرفاً خطای خام e .

$$E(k) = \frac{1}{2} r(k)^2, \quad r(k) = k_1 e(k) + k_2 [e(k) - e(k-1)] \quad (۸)$$

برای محاسبه گرادیان نسبت به یک پارامتر دلخواه Θ ، از قاعده زنجیره‌ای استفاده می‌کنیم:

$$\frac{\partial E}{\partial \Theta} = \frac{\partial E}{\partial r} \cdot \frac{\partial r}{\partial e} \cdot \frac{\partial e}{\partial \Theta} = r(k) \cdot (k_1 + k_2) \cdot \frac{\partial e}{\partial \Theta} \quad (۹)$$

از آنجا که $e = y_d - y_{net}$ ، داریم $\frac{\partial e}{\partial \Theta} = -\frac{\partial y_{net}}{\partial \Theta}$. بنابراین، تمام مشتقات استاندارد شبکه در ضریب $^{**}(k_1 + k_2)^{**}$ ضرب می‌شوند تا رفتار عاطفی را مدل‌سازی کنند.

۲.۴.۲. ۲. مشتقات پارامترهای انعطاف‌پذیر

برای هر نورون، مشتق خروجی فعال‌ساز (A) نسبت به پارامترهای شکل‌دهنده آن (Ω) باید محاسبه شود. این مشتقات ستون‌های ماتریس ژاکوبین را تشکیل می‌دهند.
(الف) مشتق نسبت به ضرایب ترکیب $(\alpha, \beta, \gamma, \delta)$:

$$\frac{\partial A}{\partial \alpha} = \sigma(\theta_1 net) \quad (10)$$

$$\frac{\partial A}{\partial \beta} = \tanh(\theta_2 net) \quad (11)$$

$$\frac{\partial A}{\partial \gamma} = \text{ReLU}(\theta_3 net) \quad (12)$$

$$\frac{\partial A}{\partial \delta} = \theta_4 net \quad (13)$$

(ب) مشتق نسبت به پارامترهای درونی $(\theta_1, \theta_2, \theta_3, \theta_4)$: در اینجا نیاز به مشتق‌گیری از تابع نسبت به آرگومان داخلی داریم:

$$\frac{\partial A}{\partial \theta_1} = \alpha \cdot net \cdot \sigma(\theta_1 net)[1 - \sigma(\theta_1 net)] \quad (14)$$

$$\frac{\partial A}{\partial \theta_2} = \beta \cdot net \cdot [1 - \tanh^2(\theta_2 net)] \quad (15)$$

$$\frac{\partial A}{\partial \theta_3} = \gamma \cdot net \cdot H(\theta_3 net) \quad (\text{Step Heaviside } H:) \quad (16)$$

$$\frac{\partial A}{\partial \theta_4} = \delta \cdot net \quad (17)$$

۳.۴.۲. ۳. حساسیت لایه‌ها (Sensitivities)

برای انتشار خطا به لایه‌های عقب‌تر، مشتق خروجی نسبت به ورودی خالص (net) محاسبه می‌شود:

$$F'(net) = \frac{\partial A}{\partial net} = \alpha \theta_1 \sigma' + \beta \theta_2 \text{sech}^2 + \gamma \theta_3 H + \delta \theta_4 \quad (18)$$

اگر $\delta^{(l)}$ بردار حساسیت لایه l باشد، رابطه بازگشتی عبارت است از:

$$\delta^{(l-1)} = (W^{(l)})^T \delta^{(l)} \odot F'(net^{(l-1)}) \quad (19)$$

که در نهایت گرادیان وزن‌ها برابر $\frac{\partial E}{\partial W^{(l)}} \propto \delta^{(l)} (A^{(l-1)})^T$ خواهد بود.

۳ تحلیل نتایج تجربی

آزمایش‌ها بر روی ۴ دیتاست انجام شد. دو سناریوی معماری بررسی گردید:

• سناریو A: تمام لایه‌ها از تابع کامل F_1 .

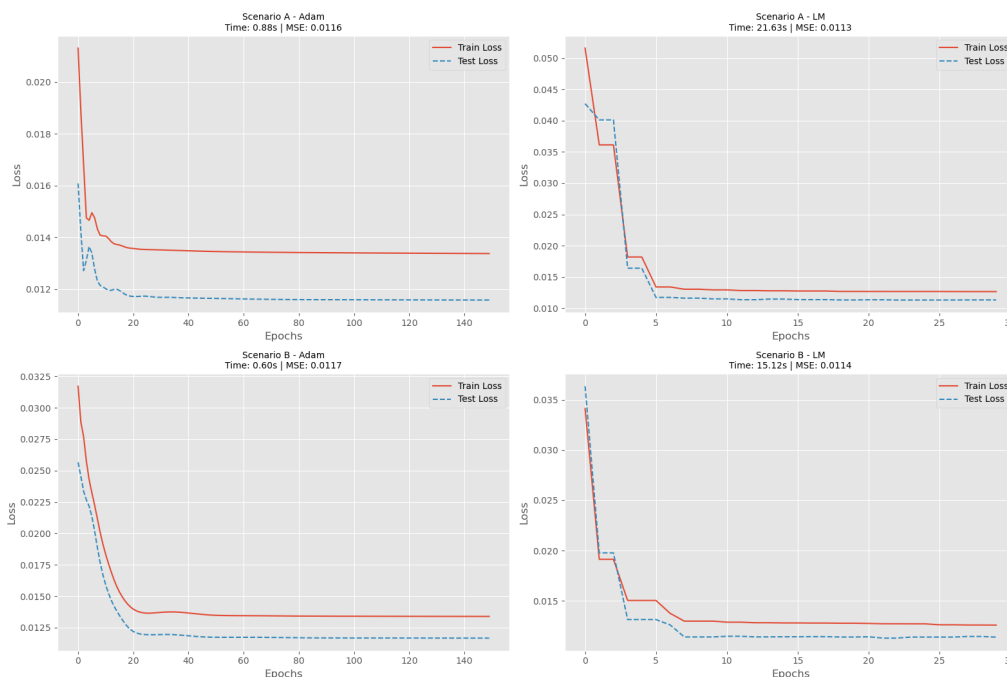
• سناریو B: لایه مخفی F_2 و لایه خروجی خطی (Pureline).

۱.۳ تحلیل مسائل رگرسیون (سری زمانی)

۱.۱.۳ ۰۱ دیتاست دمای ملبورن (Melbourne Temperature)

هدف: پیش بینی دما بر اساس ۵ روز گذشته.

Analysis Results for daily-minimum-temperatures-in-melbourne (1)



شکل ۱: روند آموزش در دیتاست ملبورن

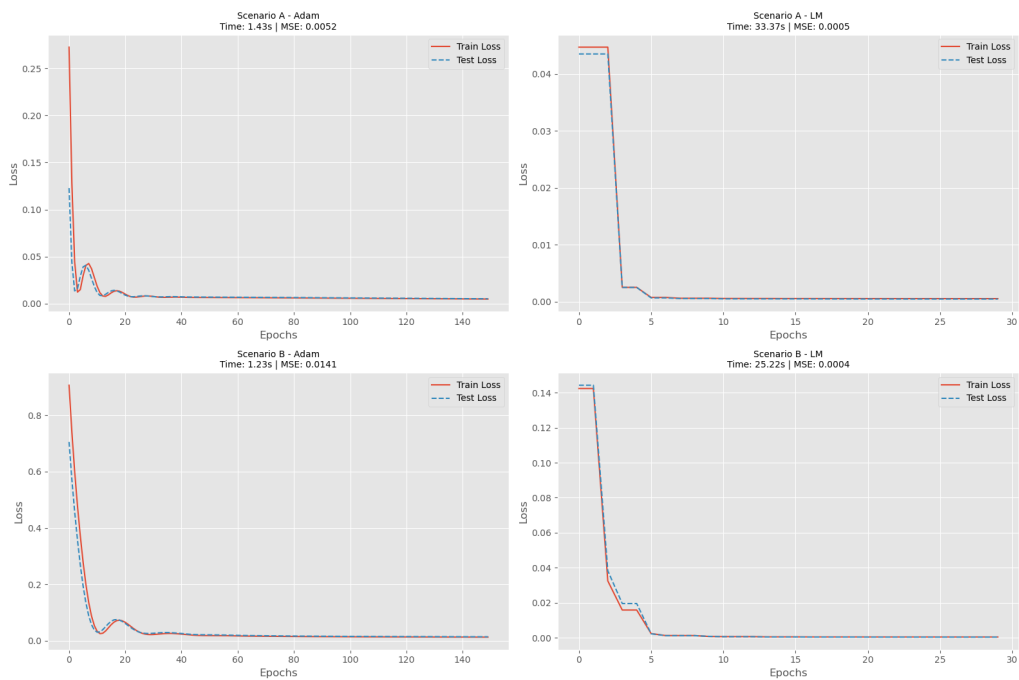
تحلیل نتایج:

- دقت نهایی: الگوریتم LM در سناریوی A به کمترین خطا ($MSE = 0.0113$) دست یافت. الگوریتم Adam نیز بسیار نزدیک عمل کرد ($MSE = 0.0116$).
- زمان اجرا: تفاوت اصلی در زمان اجراست. Adam تنها در 0.88 ثانیه اجرا شد، در حالی که LM حدود 21.6 ثانیه زمان برد.
- نتیجه گیری: برای دیتاست های ساده و هموار مانند دما، استفاده از Adam به صرفه تر است زیرا با دقت مشابه، ۲۰ برابر سریع تر است.

۲.۱.۳ ۰۲ دیتاست نظارت تصویری (Video Surveillance)

هدف: دنبال کردن الگوهای پیکسلی در فریم های ویدیو.

Analysis Results for Video surveillance Dataset



شکل ۲: مقایسه عملکرد LM و Adam در دیتاست ویدیو

تحلیل نتایج (نقطه قوت LM):

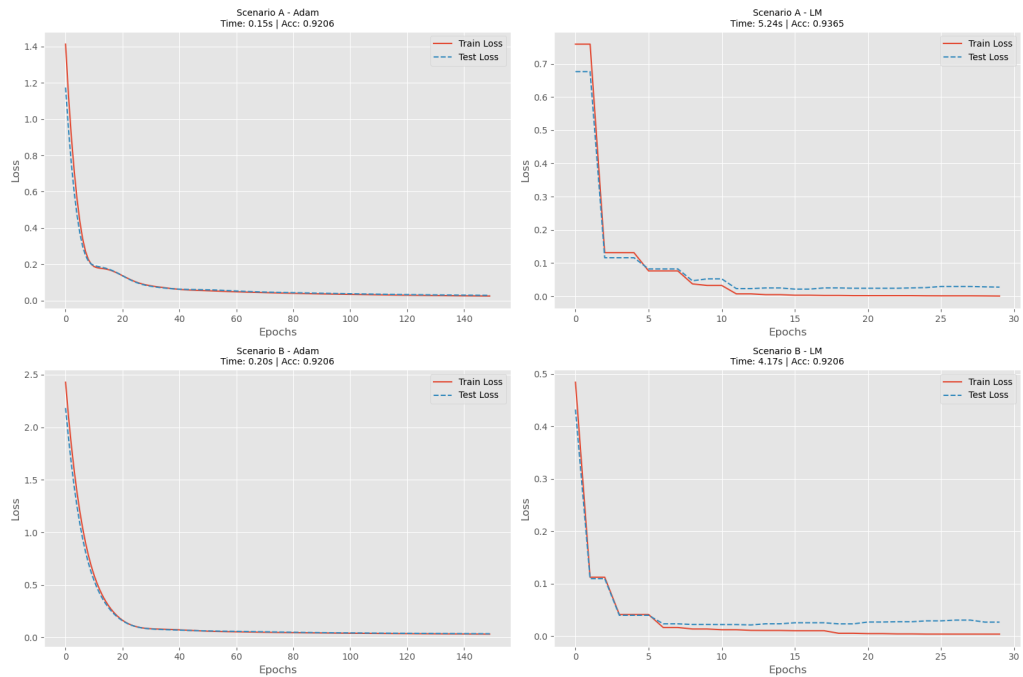
- برتری قاطع LM: در سناریوی B، الگوریتم LM به خطای خیره‌کننده 0.0004 رسید. این در حالی است که Adam در بهترین حالت خطای 0.0052 را ثبت کرد (بیش از ۱۰ برابر اختلاف).
- تحلیل عاطفی-انعطاف‌پذیر: ترکیب خروجی خطی (سناریوی B) با بهینه‌ساز مرتبه دوم، به شبکه اجازه داد تا تغییرات ظریف در سیگنال ویدیو را مدل‌سازی کند. مکانیزم عاطفی نیز به پایداری همگرایی در این سطح خطای پایین کمک کرده است.

۲.۳ تحلیل مسائل طبقه‌بندی

۱۰.۲.۳ ۰.۳ دیتاست دانه‌ها (Seeds)

هدف: تشخیص نوع گندم (۳ کلاس).

Analysis Results for classification-seeds



شکل ۳: دقت دسته‌بندی در دیتاست Seeds

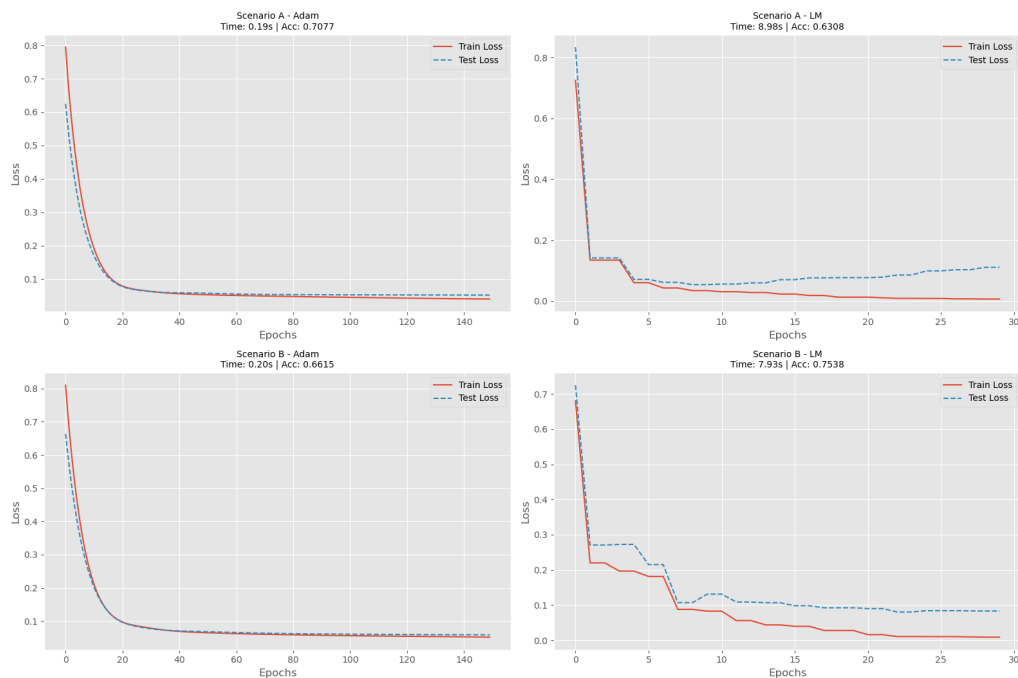
تحلیل:

- بهترین دقت توسط LM در سناریوی A بدست آمد (93.65%).
- الگوریتم Adam در هر دو سناریو روی عدد 92.06% ثابت ماند. این نشان می‌دهد LM توانسته است مرزهای تصمیم‌گیری دقیق‌تری را با تنظیم پارامترهای انعطاف‌پذیر پیدا کند.

۲۰۲۰۳ ۰۴ دیتاست شیشه (Glass)

هدف: تشخیص نوع شیشه (دیتاست دشوار و نامتوازن).

Analysis Results for classification-glass



شکل ۴: تاثیر سناریوی B در بهبود دقت دیتاست Glass

تحلیل مهم:

- شکست سناریوی A: سناریوی A با الگوریتم LM تنها به دقت 63.08% رسید. دلیل این امر احتمالاً Overfitting یا گیر افتادن در بهینه محلی به دلیل پیچیدگی بیش از حد توابع فعال‌ساز در تمامی لایه‌هاست.
- پیروزی سناریوی B: با ساده‌سازی لایه خروجی به Pureline (سناریوی B) دقت LM به طور چشمگیری افزایش یافت و به 75.38% رسید که بهترین نتیجه بین تمام مدل‌هاست. این نشان می‌دهد برای داده‌های نویزی یا نامتوازن، ترکیب لایه مخفی منعطف با خروجی خطی، تعمیم‌پذیری بهتری دارد.

۴ نتیجه‌گیری نهایی

جدول زیر خلاصه بهترین نتایج بدست آمده در این آزمایش را نشان می‌دهد:

جدول ۱: خلاصه عملکرد نهایی مدل پیشنهادی

Score	Metric	Scenario Best	Optimizer Best	Dataset
0.0113	MSE	A	LM	(Reg) Melbourne
0.0004	MSE	B	LM	(Reg) Video
93.65%	Accuracy	A	LM	(Cls) Seeds
%75.38	Accuracy	B	LM	(Cls) Glass

جمع‌بندی: پیاده‌سازی شبکه عاطفی-انعطاف‌پذیر نشان داد که استفاده از سیگنال عاطفی به همراه توابع فعال‌ساز ترکیبی، قدرت یادگیری شبکه را افزایش می‌دهد. اگرچه الگوریتم Adam سریع‌تر است، اما الگوریتم LM (به

خصوص در سناریوی (B) برای دستیابی به دقت‌های بسیار بالا و خطای نزدیک به صفر در مسائل رگرسیون پیچیده، بی‌رقیب است.