دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده‌ی ریاضی و علوم کامپیوتر

جبرخطی برای شبکه‌های عصبی

ترجمه و تهیه گزارش: سینا مهدی پور

sinamahdipour@aut.ac.ir

استاد راهنما: فرهاد رحمتی

frahmati@aut.ac.ir

تابستان ۱۳۹۷

چکیده

شبکه‌های عصبی مدل‌های کمّیی هستند که با استفاده از الگوریتم‌های یادگیری آموزش می‌بینند تا الگوهای ورودی و خروجی را تطبیق داده و به هم مرتبط کنند. ما در این مقاله به بیان چهار مفهوم اصلی از جبر خطی که برای تحلیل این مدل‌ها ضروری هستند، می‌پردازیم: ۱- نمایش یک بردار، ۲- تجزیه‌ی مقدار ویژه و مقدار تکین، ۳- بردار گرادیانی و ماتریس هسین (Hessian) یک تابع برداری و ۴- بسط تیلور یک تابع برداری. ما این مفاهیم را با تحلیل قوانین هبین (Hebbian) و ویدرو-هوف (Widrow-Hoff) و با تحلیل برخی ساختارهای ساده‌ی شبکه‌های عصبی (مثل خود مرتبط‌ساز خطی (Autossociator: نوعی شبکه که از نگاشت چندخطی استفاده می‌کند.)، هترو مرتبط‌ساز خطی (نوعی دیگر از شبکه مرتبط‌ساز) و شبکه‌ی پس‌انتشار خطا) توضیح می‌دهیم. ما هم‌چنین نشان می‌دهیم که شبکه‌های عصبی معادل نسخه‌های تکرارشونده‌ی مدل‌های بهینه-سازی و آمار استاندارد مثل تحلیل رگرسیون چندگانه و تحلیل مؤلفه‌ی اصلی هستند.

فهرست

۱ مقدمه

جبر خطی به طور خاص بسیار مناسب تحلیل نوعی از شبکه‌های عصبی به نام مرتبط‌سازها است. این مدل‌های کمّی یاد می‌گیرند تا با استفاده از الگوریتم‌های یادگیری، ورودی و خروجی را به هم مربوط کرده و بهم تطبیق دهند. وقتی مجموعه‌ی ورودی با مجموعه‌ی خروجی متفاوت است، این مدل‌ها را هترو مرتبط‌ساز می‌نامیم؛ و وقتی الگوهای ورودی و خروجی دقیقا یکسان هستند، این مدل‌ها خود مرتبط‌ساز نامیده می‌شوند. مرتبط سازها از لایه‌هایی از واحدهای پایه به اسم نورون ساخته می‌شوند. اطلاعات از لایه‌ی اول (لایه‌ی ورودی) به لایه‌ی آخر (لایه‌ی خروجی) جریان می‌یابند. برخی ساختارها ممکن است شامل لایه‌های میانی یا مخفی باشند. به طور عادی نورون‌های یک لایه به نورون‌های یک لایه‌ی دیگر متصل هستند. عمل‌های جبر خطی تغییر و تحول اطلاعات در جریان انتقال از یک لایه به لایه‌ی دیگر را توصیف می‌کنند.

۲ مفاهیم پیش‌نیاز و نشانه‌ها

بردارها با حروف کوچک مثل x و ماتریس‌ها با حروف بزرگ مثل X نمایش داده شده اند. فرض شده است که با مفاهیم ابتدایی رو به رو آشنا هستیم: عمل ترانهاده (x^T)، نرم یک بردار ($||x||$)، ضرب اسکالر $x^T w$ و ضرب ماتریس‌ها (AB). هم‌چنین از ضرب مؤلفه به مؤلفه یا هادامار $A \otimes \mathrm{B}$ نیز استفاده خواهیم کرد.

۳ تصویر

۳ – ۱ کسینوس بین دو بردار

کسینوس بردارهای x و y کسینوس زاویه ساخته شده توسط مبدا فضا با نقاط تعریف شده با مختصات این دو بردار است. پس:

$$\cos(x, y) = \frac{x^T y}{||x|| \, ||y||}$$

کسینوس نشان‌دهنده‌ی شباهت بین بردارهاست. وقتی دو بردار نسبت به هم متناسب هستند (یعنی در یک جهت هستند)، کسینوس آن‌ها یک خواهد بود؛ در حالی که وقتی دو بردار عمود بر هم هستند، کسینوس آن‌ها صفر می‌شود.

۳ – ۲ فاصله‌ی بین بردارها

در بین خانواده‌ی بزرگ فواصل بین بردارها، متداول‌ترین آن‌ها فاصله‌ی اقلیدسی است. این فاصله به کسینوس بین بردارها مرتبط است و به صورت زیر تعریف می‌شود:

$$\begin{aligned}\mathrm{d}^2(\boldsymbol{x},\boldsymbol{y}) &= (\boldsymbol{x}-\boldsymbol{y})^\mathsf{T}(\boldsymbol{x}-\boldsymbol{y}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{x} + \boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{x}^\mathsf{T}\boldsymbol{y} \\ &= \|\boldsymbol{x}\| + \|\boldsymbol{y}\| - 2\left[\cos(\boldsymbol{x},\boldsymbol{y}) \times \|\boldsymbol{x}\| \times \|\boldsymbol{y}\|\right] = \sum (x_i - y_i)^2\end{aligned}$$

۳ – ۳ تصویر بردار بر یک بردار دیگر

تصویر (عمودی) بردار x بر بردار w به صورت زیر تعریف می‌شود:

$$\mathrm{proj}_{\boldsymbol{w}}\boldsymbol{x} = \frac{\boldsymbol{x}^\mathsf{T}\boldsymbol{w}}{\boldsymbol{w}^\mathsf{T}\boldsymbol{w}}\boldsymbol{w} = \cos(\boldsymbol{x},\boldsymbol{w}) \times \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{w}\|}\boldsymbol{w}$$

نرم $\mathrm{proj}_w x$ فاصله‌ی آن با مبدا فضا و برابر مقدار زیر است:

$$\|\mathrm{proj}_{\boldsymbol{w}}\boldsymbol{x}\| = \frac{|\boldsymbol{x}^\mathsf{T}\boldsymbol{w}|}{\|\boldsymbol{w}\|} = |\cos(\boldsymbol{x},\boldsymbol{y})| \times \|\boldsymbol{x}\|$$

۳ – ۴ مثالی از قوانین یادگیری هبین و ویدرو–هوف

یک شبکه‌ی عصبی از سلول‌های متصل به هم با وزن اتصالات قابل تنظیم و تغییر به اسم سیناپس تشکیل شده است. یک شبکه عصبی با I سلول ورودی و تنها یک سلول خروجی را در نظر بگیرید. اطلاعات به وسیله‌ی سیناپس‌ها از مجموعه‌ی سلول‌های خارجی به سلول خروجی متنقل می‌شوند که به توجه به وضعیت فعالیتش یک پاسخ می‌دهد. اگر الگوی ورودی با یک بردار I-بعدی به نام x و مجموعه‌ی وزن‌های سیناپسی با یک بردار I-بعدی به نام w داده شوند، فعالیت سلول خروجی از رابطه‌ی زیر به دست می‌آید:

$$a = \boldsymbol{x}^\mathsf{T}\boldsymbol{w}$$

پس فعالیت متناسب با نرم تصویر بردار ورودی بر روی بردار وزن‌ها است. پاسخ یا خروجی سلول به o نشان داده می‌شود. برای یک سلول خطی پاسخ متناسب با فعالیت است (برای سادگی فرض کنید که ثابت تناسب یک

است). شبکه‌های هترو مرتبط‌ساز و خود مرتبط‌ساز خطی از سلول‌های خطی ساخته می‌شوند. به طور کلی خروجی یک سلول، یک تابع (معمولا اما نه الزاما پیوسته) از فعالیت آن سلول است که تابع انتقال خوانده می‌شود:

$$o = f\left(a\right) \qquad (١)$$

برای مثال در شبکه‌های پس انتشار خطا، تابع (غیر خطی) انتقال معمولا تابع لجستیک است:

$$o = f\left(a\right) = \operatorname{logist} \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x} = \frac{1}{1 + \exp\{-a\}}$$

اغلب یک شبکه‌ی عصبی طراحی شده است تا به یک ورودی داده شده، یک پاسخ خاص که هدف خوانده شده و با t نمایش داده می‌شود را نسبت دهد. یادگیری به معنای تعریف یک قانون است که مشخص می‌کند چگونه یک مقدار کوچک را به وزن‌های سیناپسی در هر دور آموزش اضافه کنیم. یادگیری خروجی شبکه را به هدف نزدیک‌تر می‌کند.

قوانین یادگیری در دو گروه اصلی تقسیم‌بندی می‌شوند: ۱- با سرپرستی (مثل ویدرو-هوف) که خطا یا فاصله‌ی بین خروجی نورون و هدف را در نظر می‌گیرند؛ و ۲- بدون سرپرستی (مثل هبین) که به چنین بازخوردی نیاز ندارند. قانون یادگیری هبین بردار وزن‌ها را در دور n+1 از آموزش به صورت زیر تغییر می‌دهد:

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \eta t \boldsymbol{x}$$

که اتا در آن یک مقدار ثابت و کوچک مثبت است که ثابت یادگیری خوانده می‌شود. پس یک دور آموزش هبین بردار وزن‌ها را با یک مقدار متناسب با هدف در جهت بردار ورودی حرکت می‌دهد.

قانون ویدرو-هوف از خطا و مشتق تابع انتقال برای محاسبه‌ی اصلاح وزن‌ها استفاده می‌کند:

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \eta f'(a)(t - o)\boldsymbol{x} \qquad (٢)$$

پس یک دور آموزش ویدرو-هوف بردار وزن‌ها با یک مقدار متناسب با خطا در جهت بردار ورودی حرکت می‌دهد.

برای شبکه‌هایی با تعداد بیشتری سلول (مثلا J سلول) در لایه‌ی خروجی، الگوی فعالیت، خروجی و هدف بردار J-بعدی خواهند بود و وزن‌های سیناپسی در یک ماتریس J*I-بعدی به اسم W ذخیره می‌شوند. معادلات یادگیری به صورت زیر تغییر خواهند کرد:

$$\boldsymbol{W}_{[n+1]} = \boldsymbol{W}_{[n]} + \eta \boldsymbol{x}\boldsymbol{t}^{\mathsf{T}} \text{ (Hebbian) and}$$
$$\boldsymbol{W}_{[n+1]} = \boldsymbol{W}_{[n]} + \eta (f'(\boldsymbol{a}) \circledast \boldsymbol{x})(\boldsymbol{t} - \boldsymbol{o})^{\mathsf{T}} \text{ (Widrow-Hoff)}$$

توجه شود که مشتق تابع انتقال به صورت مؤلفه‌ای اعمال شده است.

به طور کلی، چند (مثلا K) زوج ورودی و هدف برای آموزش وجود دارند. بنابراین مجموعه‌ی الگوهای ورودی در یک ماتریس I*K-بعدی به اسم X ذخیره می‌شود. الگوهای فعالیت و هدف نیز به همین شکل در ماتریس‌های I*K-بعدی A و T ذخیره می‌شوند. در این حالت که حالت دسته‌ای خوانده می‌شود مقدار فعالیت و یک مرحله‌ی آموزش می‌تواند برای همه‌ی الگوهای ورودی به صورت یکجا انجام شود. ماتریس خروجی در این حالت به شکل زیر محاسبه می‌شود:

$$\boldsymbol{O} = f(\boldsymbol{A}) = f\left(\boldsymbol{W}\boldsymbol{X}^{\mathsf{T}}\right)$$

تابع f به صورت مؤلفه‌ای اعمال شده است. معادلات آموزش به فرم زیر خواهند بود:

$$\boldsymbol{W}_{[n+1]} = \boldsymbol{W}_{[n]} + \eta \boldsymbol{X}\boldsymbol{T}^{\mathsf{T}} \text{ (Hebb) and}$$
$$\boldsymbol{W}_{[n+1]} = \boldsymbol{W}_{[n]} + \eta (f'(\boldsymbol{A}) \circledast \boldsymbol{X})(\boldsymbol{T} - \boldsymbol{O})^{\mathsf{T}} \text{(Widrow-Hoff)}$$

۴ تجزیه‌ی مقادیر ویژه، بردارهای ویژه و مقدار تکین

بردارهای ویژه یک ماتریس مربعی داده شده‌ی W (که از تجزیه‌ی ویژه آن حاصل شده است)، بردارهای ثابت در ضرب با W هستند. تجزیه‌ی ویژه برای یک زیر گروه خاص از ماتریس‌ها به نام ماتریس‌های نیمه قطعی مثبت بهتر تعریف می‌شود. یک ماتریس X نیمه قطعی مثبت است اگر یک ماتریس دیگر Y وجود داشته باشد به طوری که $X = YY^{\mathsf{T}}$. این مورد برای اکثر ماتریس‌های به کار رفته در شبکه‌های عصبی صادق است و به همین دلیل در اینجا تنها به همین مورد می‌پردازیم.

به طور دقیق و فرموله شده، یک بردار غیر صفر u یک بردار ویژه از ماتریس مربعی W است اگر

$$\lambda \boldsymbol{u} = \boldsymbol{W}\boldsymbol{u}$$

ا سکالر λ مقدار ویژه‌ی مربوط به بردار u ا ست. پس u یک بردار ویژه‌ی W ا ست اگر جهت آن با ضرب در W تغییر نکند (تنها اگر λ یک نباشـــد، طول آن تغییر می‌کند). به طور کلی چند بردار ویژه برای یک ماتریس داده شـده وجود دارد (حداکثر به تعداد بعد ماتریس W). این بردارهای ویژه عموماً با کاهش مقدار ویژه‌هایشـان مرتب می‌شـــوند. پس اولین بردار ویژه (u_1) بزرگ‌ترین مقدار ویژه (λ_1) را دارد. تعداد بردارهای ویژه با مقدار ویژه‌ی غیر صفر برابر رتبه‌ی ماتریس است.

مقادیر ویژه‌ی ماتریس‌های نیمه قطعی مثبت همواره بزرگ‌تر یا مسـاوی صـفر اسـت (یک ماتریس با مقادیر ویژه‌ی فقط مثبت، قطعی مثبت است). هم‌چنین هر دو بردار ویژه با مقادیر ویژه‌ی متفاوت عمود بر هم هستند:

$$\boldsymbol{u}_\ell^\mathsf{T} \boldsymbol{u}_{\ell'} = 0 \qquad \forall \quad \ell \neq \ell'$$

علاوه بر این، مجموعه‌ی بردارهای ویژه‌ی یک ماتریس، یک پایه‌ی متعامد برای سـطرها و سـتون‌های آن ماتریس خواهد بود. این م سئله با تعریف دو ماتریس زیر ن شان داده می شود: ماتریس بردار ویژه (U) و ماتریس قطری مقادیر ویژه (Λ). تجزیه‌ی ویژه‌ی W با رتبه‌ی L به صورت زیر است:

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\mathsf{T} = \sum_{\ell}^{L} \lambda_\ell \boldsymbol{u}_\ell \boldsymbol{u}_\ell^\mathsf{T}, \text{ or equivalently:} \quad \boldsymbol{\Lambda} = \boldsymbol{U}^\mathsf{T}\boldsymbol{W}\boldsymbol{U}$$

تجزیه‌ی مقدار تکین (SVD) تجزیه‌ی ویژه را به ماتریس‌های مسـتطیلی تعمیم می‌دهد. اگر X یک ماتریس I*K-بعدی باشد، SVD آن به صورت زیر تعریف می‌شود:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^\mathsf{T}$$

که در آن Δ یک ماتریس قطری است و

$$\boldsymbol{U}^\mathsf{T}\boldsymbol{U} = \boldsymbol{V}^\mathsf{T}\boldsymbol{V} = \boldsymbol{I}$$

(I ماتریس همانی است).

مؤلفه‌های قطری Δ مقادیر حقیقی مثبت هسـتند که مقادیر تکین X خوانده می‌شـوند. ماتریس‌های U و V ماتریس‌های بردارهای تکینِ چپ و راسـت (که همان بردارهای ویژه هسـتند) هسـتند. SVD کاملا مرتبط با تجزیه‌ی ویژه است زیرا U، V و Δ را می‌توان از تجزیه‌ی ویژه‌ی ماتریس‌های $X^\mathsf{T}X$ و XX^T به دست آورد:

$$\boldsymbol{X}\boldsymbol{X}^\mathsf{T} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\mathsf{T}, \quad \boldsymbol{X}^\mathsf{T}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\mathsf{T}, \text{ and } \boldsymbol{\Delta} = \boldsymbol{\Lambda}^{\frac{1}{2}}$$

توجه شود که X^TX و XX^T مقادیر ویژه‌ی یکسان دارند.

تجزیه‌ی مقدار ویژه و تکین در بیشتر زمینه‌های ریاضیات کاربردی از جمله آمار، پردازش تصویر، مکانیک و سیستم‌های مکانیکی مورد استفاده قرار می‌گیرد. در شبکه‌های عصبی این تجزیه‌ها برای مطالعه‌ی دینامیک خود مرتبط‌سازهای و هترو مرتبط‌سازها الزامی هستند.

۴ – ۱ فرآیند تکراری

در یک شبکه‌ی هترو مرتبط ساز خطی که از قانون ویدرو-هوف استفاده می‌کند، آموزش تنها مقادیر ویژه‌ی ماتریس وزن‌ها را تغییر می‌دهد. به خصوص اگر الگوهای آموزشی در یک ماتریس I*K (X) ذخیره شده اند و تجزیه‌ی مقدار تکین آن‌ها به صورت

$$X = U\Delta V^{\mathsf{T}}$$

است، آنگاه معادله‌ی یادگیری به صورت زیر خواهد بود:

$$W_{[n+1]} = W_{[n]} + \eta X(T-A)^{\mathsf{T}} = U\left\{\Delta^{-1}\left[I-\left(I-\eta\Delta^{2}\right)^{n+1}\right]\right\}V^{\mathsf{T}}T^{\mathsf{T}} \qquad ۳$$

(زیرا برای هترو مرتبط‌ساز خطی داریم: O=A و f′(A) = I).

ماتریس وزن هبین همان دور اول الگوریتن است:

$$W_{[1]} = U\left\{\Delta^{-1}\left[I-\left(I-\eta\Delta^{2}\right)^{1}\right]\right\}V^{\mathsf{T}}T^{\mathsf{T}} = U\left\{\eta\Delta\right\}V^{\mathsf{T}}T^{\mathsf{T}} = \eta X T^{\mathsf{T}} \qquad ۴$$

معادله‌ی ۳ مقادیر اتا را برای هم‌گرایی فرآیند تکراری آموزش مشخص می‌کند. اگر بزرگ‌ترین مقدار تکین X را δ_{max} بنامیم اگر اتا به گونه‌ای باشد که

$$0 < \eta < 2\delta_{\max}^{-2} \qquad ۵$$

آنگاه می‌توان نشان داد که:

$$\lim_{n\to\infty}\left(I-\eta\Delta^{2}\right)^{n} = 0, \text{ and } \lim_{n\to\infty} W_{[n]} = U\Delta^{-1}V^{\mathsf{T}}T^{\mathsf{T}} = X^{+}T$$

ماتریس $X^{+} = U\Delta^{-1}V^{\mathsf{T}}$ شبه معکوس X است. این روش یک راه حل بهینه با حداقل مربع برای انطباق بین ورودی و هدف را ارائه می‌دهد. بنابراین هترو مرتبط‌ساز خطی معادل رگرسیون چندگانه‌ی خطی خواهد بود. اگر اتا خارج محدوده‌ی اشاره شده در معادله‌ی ۵ قرار بگیرد، آنگاه مقادیر تکین و درایه‌های ماتریس

وزن‌ها در هر دور از آموزش بزرگ‌تر می‌شـــوند. در کاربرد چون شـــبکه‌های عصـــبی با کامپیوترهای دیجیتالی شبیه‌سازی می‌شوند، ماتریس وزن‌ها نهایتاً به محدودیت دقت در کامپیوتر می‌رسد و از آن عبور می‌کند.

وقتی بردارهای هدف با بردارهای ورودی یکســـان هســـتند (وقتی هر بردار ورودی به خودش نســـبت داده می شود)، هترو مرتبط ساز خطی به یک خود مرتبط ساز خطی تبدیل می شود. روش پیشین نشان می‌دهد که حالا ماتریس وزن هبین برابر ماتریس حاصل ضرب متقابل خواهد بود:

$$\boldsymbol{W}_{[1]} = \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathsf{T}}$$

با آموزش ویدرو-هوف وقتی به هم‌گرایی برســـیم، همه‌ی مقادیر ویژه‌ی غیر صـــفر ماتریس وزن‌ها برابر یک خواهند بود. در آن حالت می‌گوییم ماتریس وزن‌ها کروی شده است و برابر با

$$\boldsymbol{W}_{[\infty]} = \boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}$$

است. از آنجا که روش آماری تحلیل مؤلفه‌ی اصلی (PCA) تجزیه‌ی ویژه‌ی یک ماتریسِ ضربِ متقابلِ مشابه W را محاسبه می‌کند، خود مرتبط‌ساز خطی یک شبکه عصبی معادل PCA در نظر گرفته می‌شود.

۵ بهینه‌سازی، مشتق و ماتریس‌ها

شبکه‌های عصبی معمولاً برای بهینه‌سازی یک تابع از وزن‌های سیناپسی استفاده می‌شوند. مشتق‌گیری از یک تابع، اصلی‌ترین مسئله برای بررسی مسائل بهینه‌سازی است و در شبکه‌های عصبی شامل مشتق‌گیری از توابع برداری یا ماتریس می‌شود. در این مقاله نیاز داریم تا تابع انتقال را تابعی از بردار وزن‌ها در نظر بگیریم. این موضوع را با بازنوشتار معادله‌ی ۱ به صورت زیر نشان می‌دهیم:

$$o = f(\boldsymbol{w})$$

مشتق f(w) را که در آن w بردار I*1-بعدی است با $\nabla_{f(w)}$ نمایش می‌دهیم. این مقدار هم‌چنین گرادیان f خوانده می‌شود:

$$\boldsymbol{\nabla}_{\boldsymbol{f}(\boldsymbol{w})} = \frac{\partial f}{\partial \boldsymbol{w}} = \left[\frac{\partial f}{\partial w_1}, \ldots, \frac{\partial f}{\partial w_i}, \ldots, \frac{\partial f}{\partial w_I}\right]^{\mathsf{T}}$$

برای مثال مشتق خروجی یک نورون خطی به صورت زیر است:

$$\frac{\partial f}{\partial \boldsymbol{w}} = \left[\frac{\partial \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}}{\partial w_1}, \ldots, \frac{\partial \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}}{\partial w_i}, \ldots, \frac{\partial \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}}{\partial w_I}\right]^{\mathsf{T}} = [x_1, \ldots, x_i, \ldots, x_I]^{\mathsf{T}} = \boldsymbol{x}$$

وقتی یک تابع دو بار مشتق‌پذیر باشد، مشتق‌های مرتبه دوم در یک ماتریس ذخیره می‌شوند و به آن ماتریس هسین آن تابع گفته می‌شود. معمولا آن را با H نشان می‌دهند و به صورت دقیق به شکل زیر تعریف می‌شود:

$$\boldsymbol{H} = \boldsymbol{\nabla}_{\boldsymbol{f}}^{\boldsymbol{2}} = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 w_I} \\ \frac{\partial^2 f}{\partial w_2 w_1} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 w_I} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_I w_1} & \frac{\partial^2 f}{\partial w_I w_2} & \cdots & \frac{\partial^2 f}{\partial w_I^2} \end{bmatrix}$$

۵ – ۱ شرایط کمینگی

یک مسئله‌ی استاندارد این است که با داشتن یک قانون یادگیری نشان دهیم که آن قانون یک راه حل بهینه برای مسئله پیدا می‌کند. به عبارت دیگر یک تابع از ماتریس وزن‌ها که تابع خطا نامیده می‌شود در هنگام هم‌گرایی یادگیری به مقدار حداقل خود می‌رسد. معمولا تابع خطا به صورت جمع مربعات خطا برای همه‌ی الگوهای ورودی تعریف می‌شود.

هنگامی که گرادیان تابع خطا می‌تواند ارزیابی شود، یک شرط لازم برای بهینه بودن (چه حداقل، چه حداکثر) این است که یک بردار وزن $\widetilde{W}$ به صورت زیر یافت شود:

$$\boldsymbol{\nabla}_{\boldsymbol{f}(\tilde{\boldsymbol{w}})} = \boldsymbol{0}$$

این شرط هم‌چنین کافی است تا H قطعی مثبت باشد.

۵ – ۲ بسط تیلور

بسط تیلور یک روش استاندارد برای تقریب خطی یا درجه دوم یک تابع یک متغیره است. یادآوری می‌کنیم که بسط تیلور یک تابع پیوسته f(x) به صورت زیر است:

$$f(x) = f(a) + (x-a)\frac{f'(a)}{1!} + (x-a)^2\frac{f''(a)}{2!} + \ldots (x-a)^n\frac{f^{[n]}(a)}{n!} + \ldots$$

$$= f(a) + (x-a)\frac{f'(a)}{1!} + (x-a)^2\frac{f''(a)}{2!} + \mathcal{R}_2 \ .$$

که R_2 در آن همه‌ی عبارات با درجه‌ی بیشتر از ۲ را نشان می‌دهد و a یک مقدار مناسب برای ارزیابی f حول آن نقطه است.

این روش را می‌توان برای توابع برداری یا ماتریسی توسعه داد که شامل مفاهیم گرادیان و هسین می‌شود. به این صورت یک تابع f(x) به صورت زیر بیان خواهد شد:

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + f(\boldsymbol{x}-\boldsymbol{a})^{\mathsf{T}}\boldsymbol{\nabla}_{\boldsymbol{f(a)}} + f(\boldsymbol{x}-\boldsymbol{a})^{\mathsf{T}}\boldsymbol{\nabla}^{\boldsymbol{2}}_{\boldsymbol{f(a)}} f(\boldsymbol{x}-\boldsymbol{a}) + \mathcal{R}_2$$

۵ – ۳ به حداقل رساندن تکراری

می‌توان نشان داد که یک قانون یادگیری به یک مقدار بهینه هم‌گرا می‌شود اگر که مقدار تابع خطا را در هر دور از تکرار کاهش دهد. وقتی که گرادیان تابع خطا قابل ارزیابی است، روش گرادیان (یا تند ترین شیب) بردار وزن‌ها را با حرکت دادن آن در جهت خلاف گرادیان تابع خطا تنظیم می‌کند. به طور دقیق اصلاح وزن‌ها در مرحله‌ی (n+1) از آموزش به صورت زیر خواهد بود:

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \boldsymbol{\Delta} = \boldsymbol{w}_{[n]} - \eta\boldsymbol{\nabla}_{\boldsymbol{f(w)}}$$

به عنوان یک مثال نشان می‌دهیم که برای یک هترو مرتبط‌ساز خطی، قانون ویدرو-هوف مقدار مربع خطا بین خروجی و هدف را در هر دور از آموزش کاهش می‌دهد و به حداقل می‌رساند. تابع خطا به صورت زیر است:

$$e^2 = (t-o)^2 = t^2 + o^2 - 2to = t^2 + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{w}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} - 2t\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$$

گرادیان تابع خطا به صورت زیر است:

$$\frac{\partial e}{\partial \boldsymbol{w}} = 2(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x})\boldsymbol{x} - 2t\boldsymbol{x} = -2(t - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x})\boldsymbol{x}$$

بردار وزن‌ها با حرکت در خلاف جهت گرادیان اصلاح می‌شود. این مهم با اضافه کردن یک بردار کوچک Δ_w در خلاف جهت گرادیان به دست می‌آید. در نتیجه اصلاح وزن برای دور n+1 به صورت زیر خواهد بود:

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \boldsymbol{\Delta_w} = \boldsymbol{w}_{[n]} - \eta\frac{\partial e}{\partial \boldsymbol{w}} = \boldsymbol{w}_{[n]} + \eta(t - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x})\boldsymbol{x} = \boldsymbol{w}_{[n]} + \eta(t - o)\boldsymbol{x}$$

که همان قانون مربوط به معادله‌ی ۲ است.

روش گرادیان به خوبی عمل می‌کند زیرا گرادیان $w_{[x]}$ یک تقریب تیلور درجه اول از گرادیان بردار وزن بهینه ($\widetilde{W}$) است. این یک روش مورد توجه در شبکه‌های عصبی است زیرا روش معروف پس انتشار خطا یک روش گرادیانی است.

روش نیوتن یک تقریب تیلور درجه دو است که از معکوس ماتریس هسین w (به فرض وجود داشتن آن) استفاده می‌کند. این روش یک تقریب عددی بهتر ارائه می‌کند اما به محاسبات بیشتر و پیچیده‌تری نیاز دارد. در این روش اصلاح وزن‌ها در دور n+1 از آموزش به صورت زیر خواهد بود:

$$\boldsymbol{w}_{[n+1]} = \boldsymbol{w}_{[n]} + \boldsymbol{\Delta} = \boldsymbol{w}_{[n]} - (\boldsymbol{H}^{-1})(\boldsymbol{\nabla}_{\boldsymbol{f(w)}})$$

منابع

Abdi et al. (1999), Bishop (1995) Ellacot and Bose (1996), Haggan, Demuth, and Beale (1996), Haykin (1999), Reed and Marks (1999), Ripley (1996), and Rojas (1996)

[۱] ABDI, H. (1994a) Les r´eseaux de neurones. Grenoble, France: PUG.

[۲] ABDI, H., valentin, d., & edelman, b. (1999) Neural networks. Thousand Oak, CA: Sage.

[۳] BISHOP, C.M. (1995) Neural network for pattern recognition. Oxford, UK: Oxford University Press.

[۴] ELLACOTT, s., & bose, d. (1996) Neural networks: Deterministic methods of analysis. London: ITC.

[۵] HAGAN, M. T., demuth, h. b., & beale, m. (1996) Neural networks design. Boston: PWS.