**A Deep Understanding of Deep Learning**

**by Mike X Cohen**

فولدر 3

* در مدل های دیپ لرنینگ از ریاضیات ساده در حد دبیرستان یعنی استفاده از چهارعمل اصلی و تعدادی از توابع ساده مثل مثلثات یا لگاریتم استفاده می شود
* در لایه های شبکه از این توابع و عملیات ریاضی به تعداد بسیار زیاد استفاده می شود و همین باعث امکان انجام محاسبات پیچیده می گردد
* مدل کلی شبکه های عصبی عمیق به این صورت است که ورودی را گرفته و خروجی را پیش بینی می نماید
* برای پیش بینی در کاربردهای مختلف، استفاده از شبکه های عصبی متفاوتی توصیه شده است. مثلا برای پردازش و پیش بینی تصاویر از سی ان ان استفاده می شود و برای پردازش متن یا ترجمه ویا تحلیل سری های زمانی از آر ان ان. برای انجام محاسبات دسته بندی و خوشه بندی از ANN استفاده می شود
* نکته مهم این است که ساختار همه شبکه های عمیق مشابه هم است... مانند ساختمان ها که ممکن است از نظر کاربری و نما و امکانات متفاوت باشند ولی زیر ساخت همه آنها مشابه هم است و از آجر و سیمان و لوله و سیم وغیره تهیه شده اند.
* حرکت رو به جلو به معنی اختصاص وزن های پارامتر های ورودی به معادله می باشد و حرکت برگشت به معنی اخذ بازخورد و نیاز به اصلاح وزن پارامتر ها می باشد. Forward Propagation و Backward Propagation
* مثال حرکت رو به جلو و برگشت: تهیه ساندویچ کره و مربا : تعیین میزان کره و مربا در حرکت رو به جلو و اخذ بازخورد از مشتری در میزان شیرین یا چرب بودن ساندویچ و اصلاح میزان استفاده از آنها در حرکت رو به جلو بعدی
* مشابهت شبکه های عصبی با سلول های عصبی انسان، قیاس درستی نیست و مثل تشابه خودرو و اسب و یا هواپیما با پرنده است. تکنولوژی خصوصیات خود را دارد و قیاس آن با طبیعت درست نیست.

فولدر 5

* تفاوت Complicated , Complex :
  + اشیاء Complicated شامل تعداد زیادی از اجزا ساده هستند که در کنار یکدیگر کار کرده و مجموعه پیچیده ای را بوجود می آورند مثل خودرو یا گوشی موبایل
  + اشیاء Complex دارای تعداد اجزا کمی هستند ولی پیچیدگی آنها زیاد است. مثل علوم پزشکی و Conway’s game of Life
* تفاوت Dummy variables و One-hot encoding :
  + متغیرهای دامی در واقع بصورت صفر و یک وجود یا عدم وجود مورد خاصی را برای یک متغیر مشخص می کنند. مثلا شخص در امتحان قبول یا رد شده است. خروجی یک بردار یا ماتریس یک بعدی است.
  + در one-hot encoding چند متغیر داریم که می خواهیم وجود یا عدم وجود آن را بررسی کنیم. در واقع از چند متغیر دامی تشکیل می شود. مثلا شخص در درس ریاضی، تاریخ و علوم نمره قبول یا رد دارد. خروجی یک ماتریس چند بعدی است که از چند ماتریس یک بعدی (دامی) تشکیل شده است.
* تعریف تابع سافت مکس:
  + این تابع با محاسبه احتمال هر یک از اعداد ورودی، مجموعه ای از احتمال ها را بعنوان خروجی بر می گرداند. جمع خروجی ها برابر یک می باشد.
  + فرمول محاسبه احتمال برای هر i برابر است با :
  + بزرگترین عدد در این تابع بیشترین مقدار احتمال و کوچکترین آنها، کمترین احتمال را دارد.
* آنتروپی و کراس آنتروپی
  + به‌طور کلی در علوم و مهندسی، آنتروپی معیاری از میزان ابهام یا بی‌نظمی است.
  + آنتروپی به معنای این است که هر چه میزان عدم قطعیت یا پیش بینی پذیر بودن مقادیر بالا تر یا پایین تر باشد، مقدار آنتروپی به صفر نزدیکتر می شود. یعنی هرچه احتمال رخداد اتفاقی بیشتر باشد، آنتروپی آن کمتر است. از طرفی هر چقدر احتمال رخ دادن اتفاقی کمتر باشد، آنتروپی آن کمتر است. مثلا در پرتاب سکه، احتمال آنکه از 100 پرتاب هر صد مرتبه شیر بیاید نزدیک به صفر است و آنتروپی آن هم صفر است. آنتروپی احتمال آنکه 50 درصد شیر بیاید یک است. هر چقدر احتمال شیر آمدن سکه به ۰٫۵ نزدیکتر باشد ابهام در مورد نتیجه آن بیشتر است و اطلاع از نتیجه، به‌طور میانگین اطلاعات بیشتری دربردارد.

* + برای شرایطی که دو خروجی متصور هستیم مثل تشخیص تصویر گربه که در واقع خروجی درست یا غلط هست فرمول آنتروپی به شرح زیر است:

اگر احتمال p= 0.25 باشد، آنگاه آنتروپی یعنی ابهام در تشخیص نتیجه حدود 0.56 می باشد.

* + کراس آنتروپی احتمال رخ داد یک اتفاق به شرط اتفاق دیگر را بررسی می کند. در واقع در این زمینه دوتوزیع احتمال دخیل هستند.

* + در مثال تشخیص گربه اگر دسته بندی گربه بصورت صفر و یک باشد و احتمال دیده شدن گربه 0.25 و دیده نشدن 0.75 باشد، آنگاه فرمول محاسبه کراس آنتروپی باینری (صفر و یک) به شرح زیر است.

* گرادیان کاهشی (Gradient Descent) : برای پیدا کردن نقطه مینیموم تابع می توانیم از مشتق تابع استفاده کنیم و در نقطه ای که شیب (مشتق) برابر صفر می شود (نقطه ماکزیموم یا مینیموم) با توجه به مقدار مشتق در قبل و بعد از نقطه مشخص شده تشخیص دهیم که مینیموم یا ماکزیموم است. در نقطه مینیموم مقدار مشتق یا شیب خط قبل از نقطه منفی و بعد از آن مثبت می شود.
* گرادیان محو شونده (Vanishing Gradient) : در مواقعی ممکن است مقدار شیب خط یا مشتق تابع بدون اینکه در نقطه مینیموم یا ماکزیموم باشد برابر با صفر شود. این نقاط جایی هستند که مقدار تابع ثابت (Constant) می شود و شیب یا مشتق آن تغییر نمی کند و برابر صفر می شود. به این حالت گرادیان محو شونده (Vanish) گویند.

فولدر 6 : گرادیان کاهشی

یکی از مهمترین الگوریتم های شبکه های عصبی و یادگیری عمیق الگوریتم گرادیان کاهشی است و بدون آن بی معنی هستند

* چگونگی یادگیری عمیق توسط مدل:
  + تخمین پارامترها
  + محاسبه خطا از پارامترهای تخمینی
  + یادگیری از خطاها و اصلاح پارامترها (گرادیان کاهشی)
* در واقع در یادگیری عمیق ابتدا وزن هر متغیر بصورت رندوم انتخاب می شود و سپس بر اساس آنها مقادیر محاسبه و نیز خطا محاسبه می گردد. سپس بر اساس میزان خطا و اختلاف آن با مقادیر واقعی، سعی می کنیم مینموم خطا را بدست آوریم که در اینجا از گرادیان کاهشی (مشتق تابع) استفاه می کنیم.
* روش کار به این صورت است که در یک نقطه اتفاقی (رندوم) مقدار تابع و خطا محاسبه می شود. اگر مقدار خطا زیاد باشد، نقطه انتخاب شده باید اصلاح شود تا به نقطه مینیموم تابع خطا برسیم. برای اصلاح مقدار **شیب نقطه ضربدر نرخ یادگیری (Learning Rate) را محاسبه** و از نقطه کم می کنیم. بدین صورت به نقطه مینیموم نزدیک می شویم.
* زمانیکه مقدار محاسبه شده بالا مساوی صفر و یا خیلی نزدیک به صفر باشد، گرادیان محو شونده (Vanishing) اتفاق می افتد.
* نکته مهم اینکه گرادیان کاهشی ممکن است دقیقا روی نقطه مینیموم قرار نگیرد. دلیل این امر این است که میزان حرکت از نقطه انتخابی اول با نرخ یادگیری مشخص می شود و ممکن است جایی در اطراف و نزدیک به نقطه مینیموم قرار گیرد. لذا انتخاب نرخ یادگیری در مدل اهمیت ویژه ای دارد.
* در انتخاب مقدار مینیموم و بدست آوردن بهترین نقطه، ممکن است مینیموم های محلی(Local) باعث ایجاد خطا در بدست آوردن مینیموم کلی (Global) شوند. به دلیل زیر در یادگیری عمیق این مورد اهمیت زیادی ندارد:
  + در شبکه های یادگیری عمیق معمولا تعداد متغیرها بسیار زیاد است و لذا معادلات دارای ابعاد (بعدها) زیادی هستند. این موضوع باعث می شود که نقطه مینیموم متغیرها در یک نقطه محلی اتفاق نیافتد و جایی که یک متغیر در مینیموم قرار گرفته است سایر متغیرها لزوما در مینیموم نباشند. مثلا نموداری که مشابه زین اسب باشد را در نظر بگیریم. در یکی از بعدها یعنی یک متغیر در مینیموم است ولی در متغیرهای دیگر در ماکزیموم قرار گرفته است.
  + بطور خلاصه می توان گفت که در شبکه های یادگیری عمیق اتفاق افتادن مینیموم محلی تقریبا ناممکن است.
* زمانی که عملکرد مدل خوب است، می توان با انجام کارهای زیر از درست بودن عملکرد و اطمینان از انتخاب نقطه مینیموم:
  + چندین بار مدل را آموزش دهیم و نتیجه را بررسی نماییم تا مطمئن شویم که عملکرد مدل درست است. در واقع در هر بار شروع چون از نقطه جدیدی بصورت اتفاقی کار آغاز می گردد، می تواند نقطه مینیموم را با دقت متفاوتی تخمین بزند.
  + ابعاد مساله را زیاد کنیم تا مطمئن شویم که نقطه مینیموم محلی وجود ندارد.
* **گرادیان** مجموعه ای از دو یا چند مشتق تابع از متغیرهای مختلف است. اگر متغیرهای معادله x, y, z باشند و از معادله نسبت به تک تک آنها مشتق گرفته شود و در یک لیست قرار گیرد، به این لیست گرادیان گفته می شود.
* در محاسبه گرادیان کاهشی، با تغییر نرخ یادگیری و نیز تعداد مراحل اجرای محاسبه مشتق و محاسبه نقطه جدید (epoch) می توان به جواب بهینه و نقطه مینیموم کلی رسید. البته اگر مدل دچار نقطه مینیموم محلی نشود (مدل هایی با تعداد متغیرهای کم).
* نرخ یادگیری می تواند بر اساس زمان (epoch) ، بر اساس شیب خط (مشتق تابع) متغیر باشد و یا بصورت ثابت در نظر گرفته شود.

گرادیان انفجاری (Exploding Gradient):

زمانی که شیب تایع (مشتق) خیلی زیاد شود، زمانیکه حالت عمودی به محور ایکس داشته باشد، مقدار محاسبه شده میزان تغییر شیب ضربدر نرخ تغییرات بسیار بزرگ می شود و وقتی این عدد را از نقطه ای که در آن قرار داریم کم یا زیاد کنیم، ممکن است از روی نقطه بهینه مینیموم پرش کنیم و آن نقطه را نبینیم. در این حالت گرادیان انفجاری رخ داده است.