امیرحسین احمدی ۹۷۵۲۲۲۹۲ – تمرین ششم

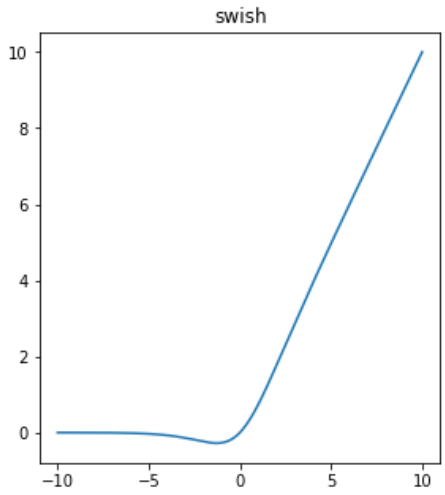
۱- الف) رابطه *swish* :

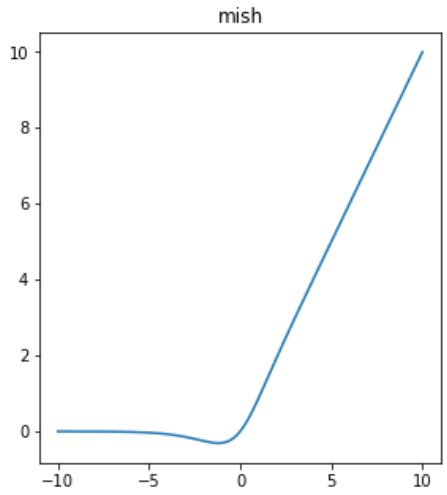
$$f(x) = x * sigmoid(x) = \frac{x}{1 + e^{-\beta x}}$$

رابطه *mish* :

$$f(x) = x * \tanh\big(softplus(x)\big) = x * \tanh\left(\ln\left(1 + e^{x}\right)\right)$$

نمودار ها:





ب) مشتق *swish* :

$$\begin{aligned} f'(x) = x'sig(x) + x\, sig'(x) &= sig(x) + x\, sig(x)\big(1 - sig(x)\big) \\ &= sig(x) + xsig(x) - xsig(x)^2 \\ &= x\, sig(x) + sig(x)\big(1 - x\, sig(x)\big) \\ &= f(x) + sigmoid(x)\,(1 - f(x)) \end{aligned}$$

مشتق *mish* :

$$\begin{aligned} f'(x) = \; & tanh\big(ln(1 + e^{x})\big) + \frac{xe^{x}}{1 + e^{x}}\,\mathrm{sech}^2\big(ln(1 + e^{x})\big) \\ = \; & \mathrm{sech}^2\big(softplus(x)\big) * x * sigmoid(x) + \frac{f(x)}{x} \\ = \; & \mathrm{sech}^2\big(softplus(x)\big) * swish(x) + \frac{f(x)}{x} \end{aligned}$$

ت) *relu* نسبت به *sigmoid* و *tanh* سریع تر است زیرا با مثبت بودن ورودی گرادیان میتواند بیشتر شود و بیشتر تغییر کند. از طرف دیگر حد بالا ندارد و این باعث میشود که گرادیان اشباع نشود.

در رابطه با *swish* و *mish* هر دو مانند *relu* از بالا نامحدود هستند و از پایین حد دارند، ولی به دلیل *non-monotonicity* هر دو مقادیر منفی کوچکی تولید میکنند که باعث جریان گرادیان در مقادیر زیر صفر شده که در *relu* این طور نیست و مقادیر زیر صفر از بین میروند. همچنین به دلیل *smooth* بودن وابستگی شبکه به *learning rate* و وزن های اولیه کم شده و سریع تر میتوان به نقطه بهینه رسید.

ث) میتوان با کنترل بتا خروجی تابع را کنترل کرد به طوری که با زیاد کردن این مقدار میتوان تابع را به *relu* و با کم کردن آن میتوان به تابع خطی میل داد.

ج) با وجود *Δ* گرادیان تابع *mish* نسبت به *swish*، *smooth* تر شده و بهینه سازی بهتر و سریع تر انجام میشود.

۲- الف) اگر مقادیر اولیه به صورت رندوم باشد به طور میانگین خروجی برابر با ۰.۵ میشود که با جایگذاری آن در فرمول دو تابع ضرر مد نظر مقادیر ۰.۷ و ۰.۲۵ بدست میاید و نتیجه دور از انتظاری نیست.

ب) خروجی تابع ضرر *mse* عددی بین صفر و یک است ولی در *bce* مقادیر بیشتر هستند، در نتیجه مقادیر میتوانند اختلافات بیشتری داشته باشند.

ت) در *bce* همانطور که مشاهده میکنیم، از حدود *epoch* ۶۰ به بعد تابع به سمت اورفیت شدن میرود و مقادیر برای دیتای ولیدیشن افزایش میابد، در نتیجه بهتر است آموزش را در همان لحظات متوقف کنیم، ولی در *mse* همانطور که مشاهده میشود تا *epoch* ها آخر همچنان شیب تابع منفی است و در حال کنم شدن است و حتی میتوان چند *epoch* دیگر آن را جلو برد و نتایج بهتری بدست آورد.

۳- در مدل با آلفای برابر با یک دقت به شدت پایین و چیزی بین ۰.۱۱ تا ۰.۱۷ است که اصلا خوب نیست. علت این موضوع این است که وقتی آلفا یک باشد تابع فعال ساز به صورت خطی خواهد شد و درواقع انگار شبکه ای با یک لایه داریم که طبیعتا دقت پایینی خواهد داشت. همچنین بهترین عملکرد مربوط به آلفای منفی یک است، زیرا هم دقت آن هم در ولیدیشن و هم تست بالا ۹۸ درصد اصن و عملکرد خیلی خوبی داشته است.