سوال ۱ –

قسمت آ) ابتدا جدول صحت گیت NAND را مینویسیم:

x1	x2	Y
-1	-1	+1
-1	+1	+1
+1	-1	+1
+1	+1	-1

حال روابط را تشکیل میدهیم:

$$sign(X1w1 + x2w2 + bw0) = Y$$

$$\rightarrow sign(-w1 - w2 + bw0) = +1$$

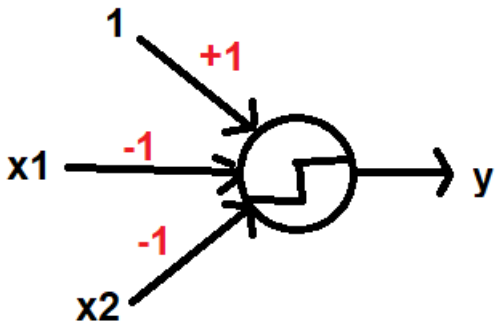
$$sign(-w1 + w2 + bw0) = +1$$

$$sign(w1 - w2 + bw0) = +1$$

$$sign(w1 + w2 + bw0) = -1$$

حال یک جواب برای معادلات بالا:

$$w1 = -1\,, w2 = -1\,, b = 1, w0 = 1$$



سپس جدول صحت گیت NOR را مینویسیم:

x1	x2	Y
-1	-1	+1
-1	+1	-1
+1	-1	-1
+1	+1	-1

حال روابط را تشکیل میدهیم:

$$sign(X1w1 + x2w2 + bw0) = Y$$

$$\rightarrow sign(-w1 - w2 + bw0) = +1$$

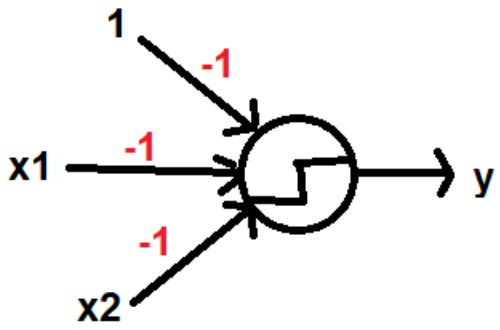
$$sign(-w1 + w2 + bw0) = -1$$

$$sign(w1 - w2 + bw0) = -1$$

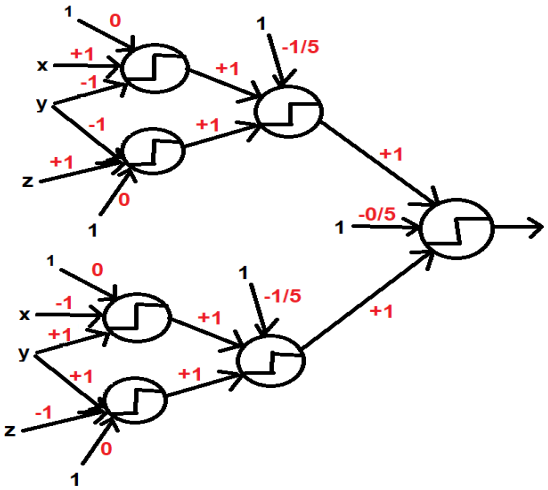
$$sign(w1 + w2 + bw0) = -1$$

حال یک جواب برای معادلات بالا:

$$w1 = -1\,, w2 = -1\,, b = 1, w0 = -1$$

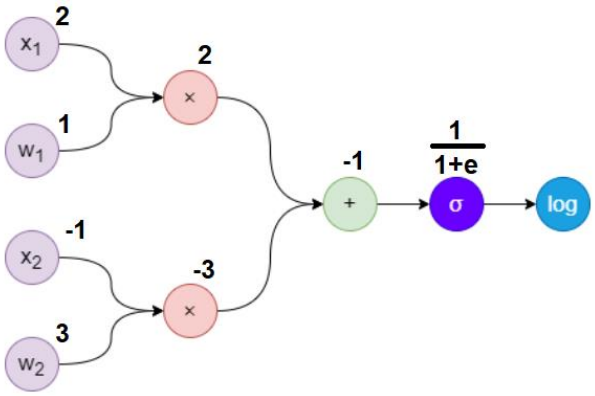


قسمت ب)



سوال ۲ –

قسمت الف)



اگر ما خروجی ضرب بالایی را p و خروجی ضرب پایینی را q و خروجی جمع را k و خروجی سیگموید را m و خروجی نهایی را g بنامیم:

$$p = x1 * w1 = 2$$

$$q = x2 * w2 = -3$$

$$k = p + q = -1$$

$$m = \frac{1}{1 + e^{-k}} = 0.27$$

$$g = \ln m = -1.3$$

حال داریم(فرض میکنیم منظور از log همان ln است)(توجه شود در این جواب ها، مقدار عدد نپر با ۲.۷ جایگذاری شده است):

$$\frac{\partial g}{\partial g} = 1$$

$$\frac{\partial g}{\partial m} = \frac{1}{m} = 3.7$$

$$\frac{\partial g}{\partial k} = \frac{\partial g}{\partial m}\frac{\partial m}{\partial k} = \frac{1}{m} * \frac{e^{-k}}{(1 + e^{-k})^2} = 3.7 * 0.2 = 0.74$$

$$\frac{\partial g}{\partial q} = \frac{\partial g}{\partial m}\frac{\partial m}{\partial k}\frac{\partial g}{\partial k}\frac{\partial k}{\partial q} = \frac{1}{m} * \frac{e^{-k}}{(1 + e^{-k})^2} * 1 = 0.74$$

$$\frac{\partial g}{\partial p} = \frac{\partial g}{\partial m}\frac{\partial m}{\partial k}\frac{\partial k}{\partial p} = \frac{1}{m} * \frac{e^{-k}}{(1+e^{-k})^2} * 1 = 0.74$$

$$\frac{\partial g}{\partial x_1} = \frac{\partial g}{\partial m}\frac{\partial m}{\partial k}\frac{\partial k}{\partial p}\frac{\partial p}{\partial x_1} = \frac{1}{m} * \frac{e^{-k}}{(1+e^{-k})^2} * 1 * 1 = 0.74$$

$$\frac{\partial g}{\partial w_1} = \frac{\partial g}{\partial m}\frac{\partial m}{\partial k}\frac{\partial k}{\partial p}\frac{\partial p}{\partial w_1} = \frac{1}{m} * \frac{e^{-k}}{(1+e^{-k})^2} * 1 * 2 = 1.48$$

$$\frac{\partial g}{\partial x_2} = \frac{\partial g}{\partial m}\frac{\partial m}{\partial k}\frac{\partial k}{\partial q}\frac{\partial q}{\partial x_2} = \frac{1}{m} * \frac{e^{-k}}{(1+e^{-k})^2} * 1 * 3 = 2.22$$

$$\frac{\partial g}{\partial w_2} = \frac{\partial g}{\partial m}\frac{\partial m}{\partial k}\frac{\partial k}{\partial q}\frac{\partial q}{\partial w_2} = \frac{1}{m} * \frac{e^{-k}}{(1+e^{-k})^2} * 1 * -1 = -0.74$$

قسمت ب) گزینه هایی که در صورت سوال آورده شده اند، به ترتیب GD سپس SGD و سپس BGD هستند.لذا ازین پس آن هارا با این نام ها مقایسه میکنیم.

زمانی که تعداد نمونه های آموزشی خیلی خیلی زیاد باشند، استفاده از GD زمان بر است زیرا باید برای هر iteration کل داده های آموزشی بررسی شوند و استفاده از BGD و SGD سریع تر هستند زیرا آن ها یا یک نمونه آموزشی را بررسی میکنند(SGD) یا یک زیر مجموعه از همه ی نمونه های آموزشی را(BGD).

روش SGD سریع تر از BGD و آن سریع تر از GD همگرا میشوند اما کمینه کردن تابع هزینه(دقت)، در GD بهتر از BGD و آن بهتر از SGD انجام میشود.

هزینه محاسباتی SGD کمتر از BGD و آن کمتر از GD است.

روش SGD میتواند از کمینه های محلی آسان تر از BGD و آن آسان تر از GD عبور کند.

سوال ۳ –

قسمت الف) باید توجه شود که L2 Regularization و Weight Decay یک چیز نیستند اما با اعمال روش SGD و پارامتر گذاری دوباره فاکتور weight decay بر اساس نرخ یادگیری،میتوان آن هارا معادل در نظر گرفت.

معادله weight decay با فاکتور لاندا، به شکل زیر است:

$$w = (1-\lambda)w - \alpha\Delta E_0$$

که ***C0*** برابر است با:

$$E_0 = \left(y - \sum_i w_i x_i\right)^2$$

حال رابطه ***L2 Regularization*** را نیز به همین شکل بازنویسی میکنیم:

$$E = E_0 + \lambda \|w\|_2^2$$

هدف آن است که پارامتر های رابطه بالا را جوری تغییر دهیم که معادل رابطه ***weight decay*** شود.

در ابتدا گرادیان تابع هزینه ***L2 Regularization*** را بر اساس ***w*** محاسبه میکنیم:

$$\Delta E = \frac{\partial E}{\partial w} = \frac{\partial E_0}{\partial w} + 2\lambda w$$

حال رابطه بالا را در رابطه ***SGD*** جایگذاری میکنیم:

$$w = w - \alpha \Delta E$$
$$w = w - \alpha(\Delta E_0 + 2\lambda w)$$
$$w = (1 - 2\alpha\lambda)w - \alpha \Delta E_0$$

همانطور که مشاهده میشود رابطه بدست آمده همان رابطه ***weight decay*** است با تفاوت اینکه در این رابطه فاکتور آن به جای λ ، $2\alpha\lambda$ است. لذا برای معادل سازی آن ها:

$$\lambda' = \frac{\lambda}{2\alpha}$$

لذا رابطه ***L2 regularization*** به شکل زیر میشود :

$$w = (1 - \lambda')w - \alpha \Delta E_0 \qquad \lambda' = \frac{\lambda}{2\alpha}$$

که میبینیم با رابطه ***weight decay*** معادل شد.البته توجه شود که با ***SGD*** توانستیم به این معادل بودن برسیم و با روش های ***adaptive*** این نتیجه حاصل نمیشود.

منبع: ***medium***

قسمت ب) حال رابطه ***L1 Regularization*** را نیز به همین شکل بازنویسی میکنیم:

$$E = E_0 + \lambda \|w\|_1$$

در ابتدا گرادیان تابع هزینه ***L1 Regularization*** را بر اساس w محاسبه میکنیم:

$$\Delta E = \frac{\partial E}{\partial w} = \frac{\partial E_0}{\partial w} + \lambda$$

حال رابطه بالا را در رابطه ***SGD*** جایگذاری میکنیم:

$$w = w - \alpha \Delta E$$
$$w = w - \alpha(\Delta E_0 + \lambda)$$
$$w = w - \alpha(\Delta E_0 + \lambda)$$

اگر ***feature*** ای کم اهمیت باشد، وزن آن کوچک میشود و گرادیان اش محو میشود یعنی وزن آن ***feature*** به سمت صفر میل میکند.بله این روش مفید است زیرا برای ***feature selection*** به کار میرود به طوری که میتوان ***feature*** های کم اهمیت را حذف کرد.

سوال ۴ –

قسمت آ)داده های *validation* داده هایی هستند که بعد از آموزش شبکه، از آن ها برای بهینه سازی پارامتر ها و یا مثلا برای مشخص کردن تعداد بهینه واحد های مخفی(در *MLP*) یا مثلا برای پیدا کردن نقطه توقف بهینه در *back propagation* استفاده میشود.یعنی یه جورایی استفاده از داده های *validation* جزئی از مرحله آموزش است.

اما داده های *test* داده هایی هستند که پس از آموزش کامل شبکه برای ارزیابی عملکرد آن و تخمین مدل انتخاب شده استفاده میشود و در مرحله ای که از آن استفاده میکنیم، آموزشی رخ نمیدهد یعنی پارامتری بهینه سازی نمیشود.

از داده های *test* برای *validation* استفاده نمیکنیم زیرا در صورتی که اینکار را انجام دهیم، نرخ خطا و کارایی مدل به عبارتی بایاس میشود(کمتر از مقدار واقعی) زیرا هنگام استفاده از داده *validation* ، در اصل آموزش متوقف نشده و مدل طبق داده های *test* آموزش داده میشود و از واقع نگری برای داده های جدید دور میشویم.

قسمت ب) خیر تخمین ما از مدلی که روی داده های *training* به عبارتی *fit* شده است، زمانی که از *validation* استفاده میکنیم *unbiased* کامل نیست.هرچه در آموزش و ارزیابی(همزمان) شبکه بوسیله داده های *validation* جلو برویم، تخمینگر بیشتر و بیشتر *biased* میشود زیرا استفاده از یک مجموعه داده *validation* از پیش تعیین شده، در اصل تخمین *unbiased* ای به ما ارائه نمیکند زیرا *uncertainty* خروجی را به ما ارائه نمیدهد.البته توجه شود استفاده از *validation* بهتر از زمانی است که از آن استفاده نکنیم زیرا آموزش بوسیله فقط داده های *training* یک تخمینگر کاملا *biased* به ما میدهد.اما خب نکته اینجاست که حتی با استفاده از *validation* نیز نمیتوانیم یک تخمینگر *unbiased* داشته باشیم.

وقتی از *k-fold cross validation* نیز استفاده کنیم، تخمینگر ما بیشتر از حالت قبل *unbiased* میشود اما باز هم یک تخمینگر کاملا *unbiased* نداریم زیرا این روش هم نمیتواند بحث *uncertainty* خروجی را به ما ارائه کند. راه حل این است که از *naïve estimator* ها استفاده کنیم.

منبع: سایت *machinelearningmastery* و مقاله *no biased estimator of the variance of k-fold cross-validation*

سوال ۵– در نوت بوک جواب داده شده است.