دانشکده‌ی مهندسی کامپیوتر

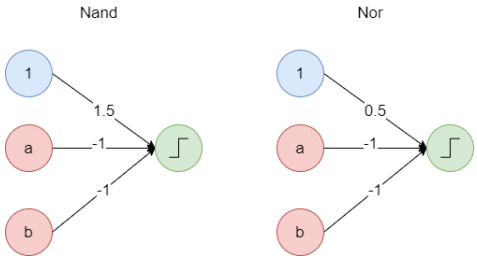
یادگیری ماشین بهار ۱۴۰۰

پاسخ تمرین سری چهارم

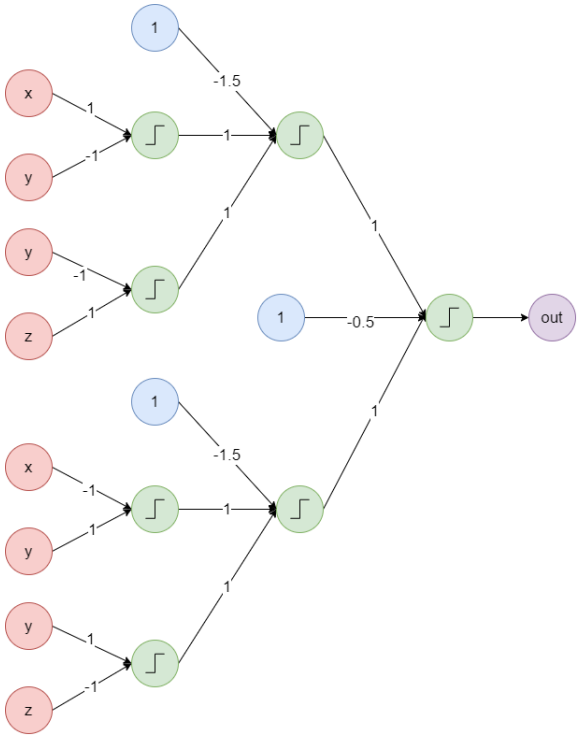
مدرس: دکتر محمّدحسین رهبان

پاسخ سوال ۱

آ)

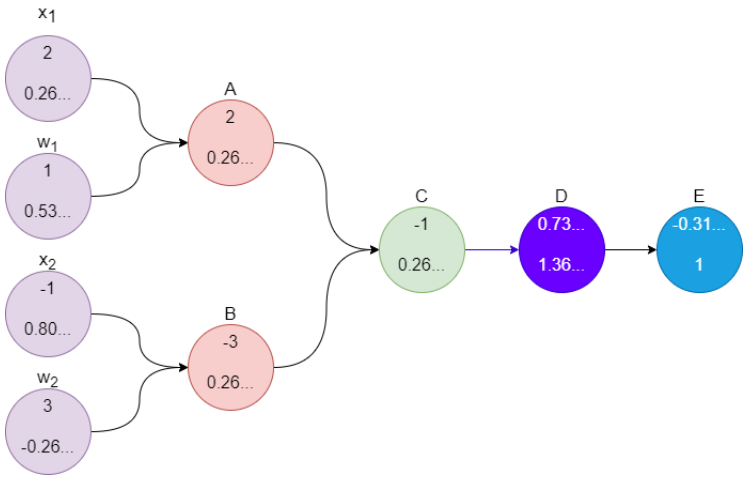


ب)



پاسخ سوال ۲

آ) برای هر راس، مقدار خود راس در بالا و مقدار مشتق نسبت به آن در پایین راس آمده است. دقت کنید که نمره به راه حل و انجام درست الگوریتم بستگی دارد نه مقادیر به دست آمده.



$$\frac{\partial E}{\partial E} = 1 \qquad \frac{\partial E}{\partial D} = \frac{1}{D} = 1.36\ldots$$

$$\frac{\partial E}{\partial C} = \frac{\partial E}{\partial D}\frac{\partial D}{\partial C} = \frac{\partial E}{\partial D} \times (1-C)\,C = 0.26\ldots \qquad \frac{\partial E}{\partial A} = \frac{\partial E}{\partial C}\frac{\partial C}{\partial A} = \frac{\partial E}{\partial C} = \frac{\partial E}{\partial B} = 0.26\ldots$$

$$\frac{\partial E}{\partial x1} = \frac{\partial E}{\partial A}\frac{\partial A}{\partial x1} = \frac{\partial E}{\partial A} \times w1 = 0.26\ldots \qquad \frac{\partial E}{\partial w1} = \frac{\partial E}{\partial A}\frac{\partial A}{\partial w1} = \frac{\partial E}{\partial A} \times x1 = 0.53\ldots$$

$$\frac{\partial E}{\partial x2} = \frac{\partial E}{\partial B}\frac{\partial B}{\partial x2} = \frac{\partial E}{\partial B} \times w2 = 0.80\ldots \qquad \frac{\partial E}{\partial w2} = \frac{\partial E}{\partial B}\frac{\partial B}{\partial w2} = \frac{\partial E}{\partial B} \times x2 = -0.26\ldots$$

ب) برداشتن یک گام به ازای همه داده‌ها برای مساله‌هایی که دارای سطح *loss* نسبتا هموار هستند (به مساله *convex* نزدیک‌تر است) بهتر است و به خوبی به نقطه کمینه موضعی همگرا می‌شود. همچنین اگر قابلیت موازی‌سازی داشته باشیم، هزینه محاسباتی کمتری دارد.

برداشتن یک گام به ازای هر داده از همه داده‌ها در حالتی که سطح *loss* ناهموار است بهتر عمل می‌کند چرا که ما را از گیرکردن در کمینه‌های موضعی ناشی از ناهمواری ها نجات می‌دهد ولی این روش هزینه محاسباتی بالایی دارد و سخت‌تر همگرا می‌شود و از *noise* بالایی برخوردار است. در نهایت بخش کردن داده‌ها اگر با نسبت‌های خوبی انجام شود، روشی بینابین دو روش قبلی است و از بدی‌های هر کدام کاسته و از خوبی‌های آن‌ها استفاده می‌کند.

پاسخ سوال ۳

آ) ابتدا مشتق را حساب می کنیم:

$$\frac{\partial E}{\partial w_i} = -2(y - \Sigma w_i x_i)x_i + 2\lambda w_i$$

حال بروزرسانی را اعمال می کنیم:

$$w_i' = w_i - \eta \frac{\partial E}{\partial w_i}$$

$$w_i' = (1 - 2\eta\lambda)w_i + 2\eta(y - \Sigma w_i x_i)x_i$$

همانطور که می‌بینید در هر گام بروزرسانی، وزن در عددی بین صفر و یک ضرب شده و سپس مقدار بهینه‌سازی روی آن اعمال می‌شود؛ به همین علت به این روش *weight decay* نیز می‌گویند.

ب) حال مراحل مشابه را برای عبارت جدید تکرار می کنیم، در نهایت به قانون بروززسانی زیر می رسیم:

$$w_i' = (w_i - 2\eta\lambda\, sign(w_i)) + 2\eta(y - \Sigma w_i x_i)x_i$$

همانطور که می‌بینید اگر وزنی مثبت باشد از مقدار آن کاسته شده و اگر منفی باشد به آن اضافه می‌شود و سپس بروزرسانی می‌شود. این باعث می‌شود که وزن‌های خیلی کوچک و بی‌تاثیر در طول آموزش به صفر میل کنند و حذف شوند.

پاسخ سوال ۴

آ) داده‌های *validation* در واقع بخشی از داده‌های آموزش ما هستند و ما از روش‌های *validation* به این خاطر استفاده می‌کنیم که پارامترهای مدل خود را به خوبی تنظیم کنیم. داده‌های *test* برای سنجیدن *generalization* مدل به کار می‌روند و فرض بر این است که ما در هنگام طراحی و انتخاب پارامترهای مدل به آن‌ها دسترسی نداریم؛ بنابراین اگر از آن‌ها برای *validation* استفاده کنیم نمی‌توانند سنجش *generalization* مدل را به درستی انجام دهند.

ب) در حالت کلی داده‌هایی که کنار گذاشته‌ایم تخمینگری نااریب هستند، اما اگر با استفاده از آن‌ها *early stopping* انجام دهیم یا پارامترهای مدل را انتخاب کنیم (که در *K-fold cross validation* این کار را انجام می‌دهیم) حتی با وجود این که مدل را روی آن‌ها آموزش نداده‌ایم دچار *bias* شده و تخمین نسبتا بهتری از واقعیت ارائه می‌دهیم.

پاینده باشید