یادگیری ماشین برای بیوانفورماتیک

بهار ۱۴۰۲

استاد:علی شریفی زارچی

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایانترم - ۱۵۰ دقیقه | تاریخ برگزاری: ۲۳ خرداد

سوالات (۱۰۰ نمره)

۱. (۲۰ نمره) پاسخ کوتاه

به سوالات زیر به صورت کوتاه پاسخ دهید:

- یکی از استفاده های مرسوم Autoencoderها میتواند کاهش ابعاد باشد. تفاوت استفاده از Autoencoder و PCA را برای این هدف توضیح دهید. همچنین ذکر کنید آیا امکان دارد استفاده از نوع خاصی از تابع فعالسازی باعث شود که عملکرد این دو الگوریتم تقریبا مشابه شود؟
- یک شبکه عصبی fully connected را در نظر بگیرید که تابع فعالسازی تمام لایه‌ها تابع sigmoid است. برای مقداردهی اولیه وزن‌ها، همه وزن‌های شبکه را به صورت تصادفی از توزیع $\text{Uniform}[-۱۰۰۰, ۱۰۰۰]$ انتخاب میکنیم. آیا این ایده خوبی است؟ استفاده از این مقداردهی اولیه موجب چه پدیده‌ای می‌شود؟
- توضیح دهید این فرض که سعی میکنیم توزیع مقادیر latent layer در ساختار VAE به یک توزیع نرمال عادی(تک قله) نزدیک باشد باعث چه محدودیتی می شود؟ منظور از توزیع نرمال عادی single modality gaussian distribution است.
- به چه دلیل از batch normalization در شبکه های عصبی استفاده میشود؟ فرمول و روش استفاده از آن بر حسب آماره های یک mini-batch را بیان کنید.

۲. (۱۰ نمره) استفاده از momentum در بهینه سازی

یکی از تکنیک های متداول در آموزش شبکه های عصبی استفاده از momentum است؛ یعنی در هر قدم از gradient descent از قاعده زیر برای آپدیت وزن های شبکه(W) استفاده میکنیم: (β نرخ momentum و α نرخ آموزش است.)

$$V_{t+۱} \leftarrow \beta V_t + (۱ - \beta)\frac{\partial J}{\partial W_t}$$

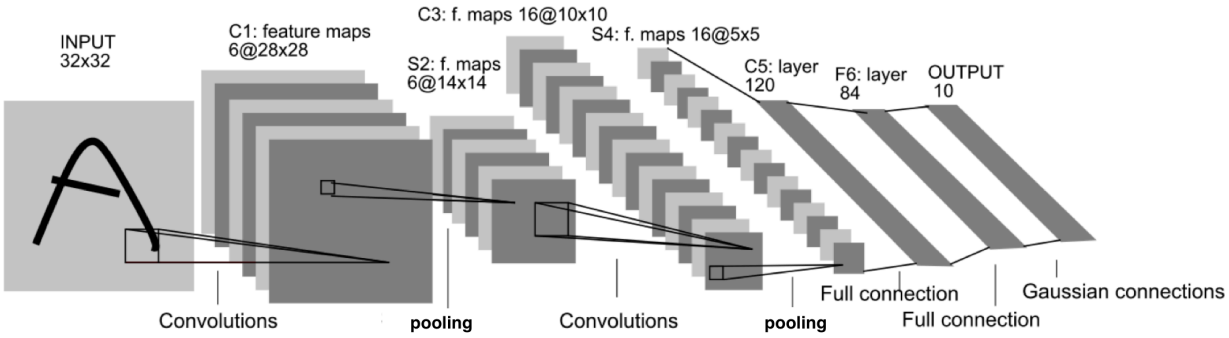
$$W_{t+۱} \leftarrow W_t - \alpha V_{t+۱}$$

الف) تفاوت این روش را با gradient descent عادی توضیح داده و ذکر کنید استفاده از momentum چگونه میتواند باعث افزایش پایداری و سرعت آموزش مدل شود؟
راهنمایی: یک قیاس مشابه میتواند اینطور باشد: به این فکر کنید که چگونه شخصی که از تپه پایین می رود ممکن است تحت تأثیر موانع کوچک یا تغییرات زمین قرار گیرد. چگونه این می تواند به رفتار یک الگوریتم بهینه سازی مانند gradient descent مرتبط باشد؟

ب) اثر هایپرپارامتر β را در این تکنیک توضیح دهید.

۳. (۲۰ نمره) معماری ساده LeNet-5

در شکل زیر میتوانید معماری شبکه ساده و تاریخی LeNet-5 را مشاهده نمایید که برای یک طبقه بندی ۱۰ کلاسه طراحی شده است:



با توجه به این ساختار و ذکر این نکته که در لایه های .Conv از padding استفاده نشده و همچنین stride=1 است، به سوالات زیر پاسخ دهید:

الف) با توجه به ابعادی که برای ورودی و feature mapها بر روی شکل ذکر شده است، مشخص کنید که سایز فیلترهای .Conv مورد استفاده در لایه C1 چه اندازه است؟

ب) چه تعداد پارامتر(وزنها و بایاس ها) در لایه C1 وجود دارد؟

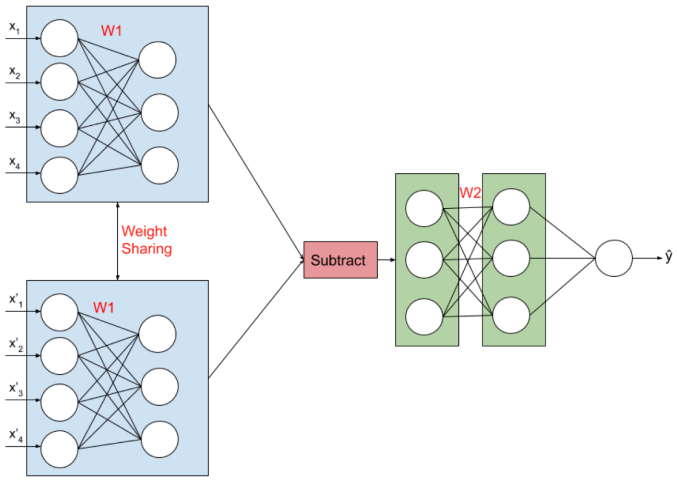
پ) چه تعداد پارامتر(وزنها و بایاس ها) در لایه C3 وجود دارد؟

ت) چه تعداد پارامتر(وزنها و بایاس ها) در لایه F6 وجود دارد؟

ث) تفاوت اثر استفاده از Max-Pooling و Average-Pooling را ذکر کنید.

۴. (۳۵ نمره) شبکه عصبی Siamese و back-propagation

شبکه عصبی زیر را در نظر بگیرید که به نام شبکه عصبی Siamese شناخته میشود و از دو شبکه مشابه (با ساختار و وزنهای یکسان) تشکیل شده که ورودی های متفاوتی دریافت میکند ولی وزنهای دو شبکه با یکدیگر به اشتراک گذاشته شده است. معمولا تفاوت خروجی این دو شبکه نیز در ادامه از یک شبکه عصبی دیگر عبور میکند.



معمولا از این شبکه در مواردی همچون سیستم های تشخیص هویت و یا مواردی که دیتای کمی برای آموزش داریم استفاده میشود.

فرض کنید که یک شبکه Siamese دولایه داریم که معادلات زیر مسیر forward شبکه را توصیف میکنند:

$$
\begin{aligned}
z_1 &= W_1 x^{(i)} + b_1 \\
a_1 &= ReLU(z_1) \\
z_2 &= W_1 x'^{(i)} + b_1 \\
a_2 &= ReLU(z_2) \\
a &= a_1 - a_2 \\
z_3 &= W_2 a + b_2 \\
\hat{y}^{(i)} &= \sigma(z_3) \\
L^{(i)} &= y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \\
J &= -\frac{1}{m} \sum_{i=1}^{m} L^{(i)}
\end{aligned}
$$

توجه داشته باشید که $x^{(i)}$ و $x'^{(i)}$ نشان دهنده یک زوج ورودی نمونه می باشند که هر یک به سایز $\mathbb{R}^{D_x \times ۱}$ هستند. همچنین $z_۱, z_۲ \in \mathbb{R}^{D_A \times ۱}$ هستند. خروجی این شبکه نیز به صورت اسکالر است و در نهایت به تعداد m نمونه آموزشی داریم.

با توجه به اطلاعات بالا به سوالات زیر پاسخ دهید:

الف) سایز $W_۱, W_۲, b_۱, b_۲$ را مشخص کنید.

ب) $\delta_۱^{(i)} = \frac{\partial J}{\partial z_۳}$ را محاسبه نمایید.

پ) $\delta_۲^{(i)} = \frac{\partial z_۳}{\partial a}$ را محاسبه نمایید.

ت) $\delta_۳^{(i)} = \frac{\partial a}{\partial W_۱}$ را محاسبه نمایید.

ث) با استفاده از محاسبات قسمت های قبل $\frac{\partial J}{\partial W_۱}$ را محاسبه نمایید.

ج) با استفاده از محاسبات قسمت های بالا قاعده آپدیت $W_۱, W_۲, b_۱, b_۲$ را بنویسید. فرض کنید نرخ آموزش برابر با α است.

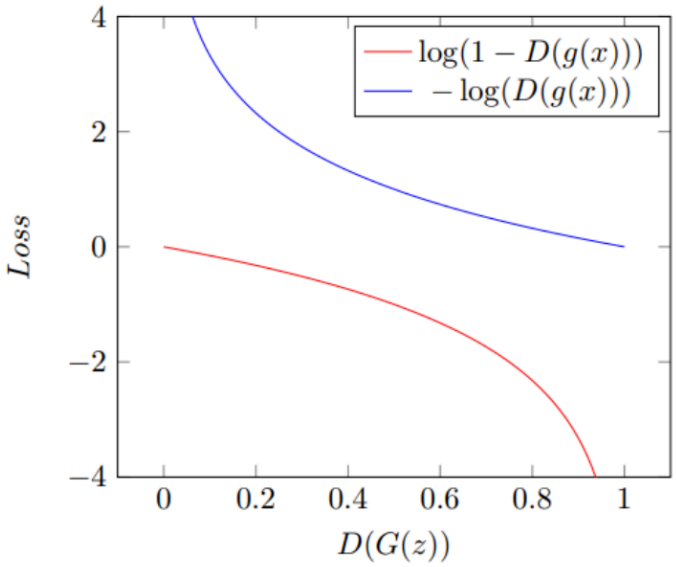
راهنمایی:

$$ReLU(x) = max\{۰, x\}$$

$$\sigma(x) = \frac{۱}{۱ + e^{-x}}$$

۵. (۲۰ نمره) شبکه های مولد تخاصمی (GAN)

الف) کدام یک از توابع هزینه $log(۱ - D(G(z))$ و $-log(D(G(z))$ که در شکل زیر نمایش داده شده اند برای آموزش قسمت Generator در GAN مناسب تر است؟ علت انتخاب خود را توضیح دهید.



ب) دو نمونه از مشکلاتی که در آموزش GANها وجود دارد را ذکر کنید.

پ) فرض کنید که دو دیتاست داریم؛ به طوریکه یکی از آنها دیتای mRNA gene expression و دیگری microRNA expression است که هر دو از یک بافت گرفته شده اما این دو دیتا به صورت unpaired هستند. یک شبکه عصبی بر پایه GAN پیشنهاد دهید که بتواند این دو دیتاست را به هم تبدیل کند؟ به عبارتی توضیح دهید چگونه میتوان با داشتن یکی از این دو دیتاست، دیتاست دیگر را با آن تطابق داد.

۶. (۲۰ نمره) مدل های زبانی

الف) روش Word2Vec را به صورت خلاصه توضیح دهید.

ب) فرض کنید در استفاده از روش Word2Vec به جای ورودی دادن به صورت One-hot ورودی خود را به صورت یک توزیع احتمال بدهیم؛ به طوریکه ورودی ها به شکل $\frac{M(i,j)}{\sum_k M(i,k)}$ باشند که $M(i,j)$ نشان میدهد دو کلمه i و j چند بار با یکدیگر اتفاق افتاده اند. توضیح دهید این نوع ورودی دادن باعث چه مزایایی در embedding میتواند شود.

پ) تفاوت بین RNN و GRU را به صورت خلاصه توضیح دهید.

ت) تفاوت بین GRU و LSTM را به صورت خلاصه توضیح دهید.