یادگیری عمیق

بهار ۱۴۰۰

دانشکده مهندسی کامپیوتر

مدرس: حمید بیگی

تمرین سری چهارم | یادگیری عمیق | زمان تحویل: ۱۸ خرداد

به نکات زیر توجه کنید:

فایل گزارش را به همراه تمامی کدها در یک فایل فشرده و با عنوان #HW4_STD بارگذاری نمایید.

برای هر یک از سوالات عملی، پوشه ای مجزا در نظر بگیرید و کدها را درون آن قرار دهید و از شماره سوال برای نام پوشه استفاده کنید.

در صورتی که در جواب از مقاله خاصی استفاده شده است لازم است که به آن اشاره شود.

نمره کل تمرین ۷۴ است. ۲۲ نمره از کل تمرین امتیازی می باشد.

برای بخش عملی دو نوتبوک در اختیار شما قرار خواهد گرفت و نمره عملی از آن محاسبه می‌شود که جمعا ۶۰ نمره است.

مسئله اول- KL-Divergence (۱۰ نمره)

۱) kl-divergence بین دو توزیع گاوسی $N(\mu_1, \Sigma_1)$ و $N(\mu_2, \Sigma_2)$ را بدست آورید(ماتریس کواریانس قطری است).

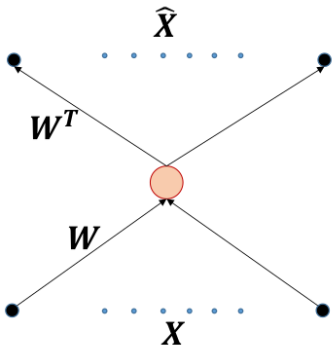
۲) توزیع $p(x)$ را یک توزیع ثابت دلخواه در نظر بگیرید که می‌خواهیم آن را با استفاده از توزیع $q(x) = N(x|\mu, I)$ تقریب بزنیم (ماتریس کواریانس ماتریس همانی است). با نوشتن kl-divergence بین دو توزیع و مشتق گرفتن نسبت به μ نشان دهید که میانگین بهینه برابر است با:

$$\mu^* = \underset{\mu}{\operatorname{argmin}} \mathbb{KL}(p||q) = \mathbb{E}_p[x] \quad (1)$$

مسئله دوم- autoencoders (۱۰ نمره)

۱) مزایای استفاده از autoencoder چیست و چرا از آن استفاده می‌شود؟

۲) نشان دهید که linear autoencoder شکل ۱که تلاش به کمینه کردن L۲ divergence می‌کند همان PCA است.



شکل ۱: linear autoencoder

مسئله سوم-marginal likelihood estimation (۱۰ نمره)

یک latent variable model در نظر بگیرید که در آن x مقدار مشاهده شده و z مقدار نهان[۱] است. با استفاده از تابع بیشینه حاشیه‌ای[۲]:

$$p_\theta(x) = \int_z p_\theta(x|z)p(z)dz \tag{۲}$$

با استفاده از توزیع پیشنهاد شده $q(z|x)$ یک importance sampling estimator را در نظر بگیرید:

$$\hat{L}(x) = \log(\frac{۱}{M}\Sigma_{i=۱}^{M}\frac{p_\theta(x|z_i)p(z)}{q(z_i|x)}) \tag{۳}$$

نشان دهید که $\hat{L}$ یک biased estimator از $\log(p_\theta(x))$ است اما در صورتی که $M \to \infty$ آنگاه به طور تقریبی unbiased است.

$$\mathbb{E}_{z_i \sim q(.|x)}[\hat{L(x)}] \leq \log(p_\theta(x)) \tag{۴}$$

$$\lim_{M\to\infty} \hat{L(x)} = \log(p_\theta(x)) \tag{۵}$$

این ایده استفاده شده در مقاله importance weighted autoencoder است.

مسئله چهارم (امتیازی ۱۰ نمره)

۱) در صورتی که latent variable در VAE ها به جای پیوسته گسسته باشد . چه تغییراتی لازم است داده شود.

۲) در beta-VAE یک ضریب β پشت جمله KL اضافه شده است. با انجام چه محاسباتی این عمل انجام شده است و

[۱] latent

[۲] likelihood marginal

چه تاثیری در عملکرد VAE می‌گذارد؟

مسئله پنجم شبکه‌های تخاصمی مولد[۳] (۱۲+۲۲)

یک) تابع هدف شبکه‌های تخاصمی مولد در حالت پایه به صورت زیر تعریف می‌شود:

$$\mathcal{V}(G,D) = \mathbb{E}_{x\sim p_{Data}}[logD(x)] + \mathbb{E}_{x\sim p_G}[log(۱ - D(x))] \quad (۶)$$

که G شبکه مولد و D شبکه تمیزدهنده است و رابطه فوق نسبت به D بیشینه و نسبت به G کمینه می‌گردد.

الف) (۲ نمره) اگر ظرفیت تمیزدهنده نامحدود باشد، نقاط بیشینه رابطه ۶ نسبت به تمیز دهنده را بیابید. (بر حسب p_{Data} و p_G بدست آورید)

ب) (۳ نمره) حال اگر در رابطه ۶، گام بهینه سازی نسبت به D به‌صورت بهینه انجام گیرد، نشان دهید کمینه کردن $\mathcal{V}(G,D^*)$ که D^* نقطه بهینه آن نسبت به D است، معادل کمینه کردن فاصله $JS(p_{Data}||p_G)$ (Shannon Jensen) خواهد بود.

پ) (۲ نمره) دیدگاه فوق به لحاظ نظری مطرح می‌شود. آنچه در عمل اتفاق می‌افتد، پارامترهای G و D در هر دسته داده[۴]، با توجه به رابطه ۶ یک یا چند بار به‌روزرسانی می‌شوند. برای مثال در یک دسته داده می‌توان سه بار پارامترهای تمیزدهنده و سپس یکبار پارامترهای مولد را به‌روزرسانی کرد. حال تعیین نسبت به آموزش مولد به تمییر دهنده و سپس یکبار پارامترهای مولد را به‌روزرسانی کرد. حال تعیین نسبت آموزش مولد به تمییزدهنده یکی از مشکلات اساسی آموزش این خانواده از شبکه‌ها است. با توجه به دو بخش قبل، برای نزدیک شدن به دیدگاه نظری مطرح شده، چه نسبت آموزشی را بین مولد و تمییزدهنده پیشنهاد می‌کنید؟

۲) فرض کنید، دامنه توزیع داده اصلی و توزیع شبکه مولد همپوشانی نداشته باشند و همچنین D نزدیک به تمییز دهنده بهینه باشد.

الف) (۲ نمره) گرادیان $log(۱ - D(x))$ را نسبت به logit های شبکه D بدست آورید (اگر تمییز دهنده به صورت $\sigma(a)$ که $a = f(x)$ تعریف شده باشد، a، logit خواهد بود.)

ب) (۲ نمره) در این حالت چه گرادیانی به شبکه G می‌رسد؟ چه مشکلی ایجاد می‌شود؟

پ) (۲ نمره) حال اگر از $-logD(x)$ به عنوان تابع هزینه مولد شبکه مولد استفاده شود، چطور به حل این مشکل کمک می‌کند؟

۳) (امتیازی) رویکرد دیگری که در نحوه به‌روزرسانی پارامترهای دو شبکه G و D می‌توان متصور بود، آن است که تابع هدف ۶، به صورت همزمان نسبت به پارامترهای هر دو شبکه به روز رسانی شود. فرض کنید تابع هدف شما xy که نسبت به x بیشینه و y کمینه می‌گردد.

الف) (۱ نمره) نقاط زینی این رویه را بیابید.

ب) (۳ نمره) اگر از نقطه (۱،۱) شروع کنیم و پارامترهای x و y را به صورت همزمان به‌روزرسانی کنیم، مسیر حرکت روی

[۳] Generative Adversarial Networks:
[۴] Batch

فضای xy را بررسی کنید (با استفاده از معادله دیفرانسیل و یا شبیه سازی). آیا این روش همگرا می‌شود؟

۴) (۴ نمره) از آنجا که در هر دو روش بیشینه درست‌نمایی و تخاصمی به دنبال آموزش شبکه مولد هستیم، در این بخش می‌خواهیم روش تخاصمی دیگری را بیابیم که در عمل با روش بیشینه درست‌نمایی یکسان باشد. می‌دانیم در روش بیشینه درست‌نمایی تابع هزینه زیر مورد استفاده قرار می‌گیرد.

$$\mathcal{L}_{MLE}(\theta) = \mathbb{E}_{x\sim p_{Data}}[-logp_G(x)] \quad (۷)$$

که θ پارامترهای شبکه مولد است.

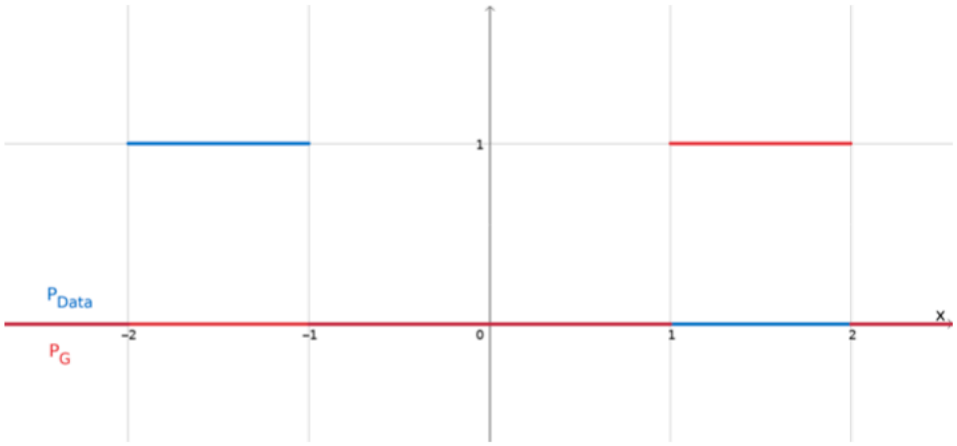
حال روش تخاصمی را به این صورت تعریف می‌کنیم که تابع هزینه تمییزدهنده تغییری نکرده (رابطه ۶) اما تابع هزینه شبکه مولد به صورت زیر باشد.

$$\mathcal{L}_{MLE-GAN}(\theta) = \mathbb{E}_{x\sim p_G}[f(x)] \quad (۸)$$

تابع f چه باشد تا گرادیان $\mathcal{L}_{MLE-GAN}$ نسبت به θ، با گرادیان $\mathcal{L}_{MLE}$ نسبت به θ یکسان شود؟ این تابع را برحسب D^* و logit آن به دست آورید. چند نکته:

- D^* تمیز دهنده بهینه است.
- فرض کنید تابع f تابع مستقیمی از پارامترهای θ نباشد. به عبارت دیگر مشتق f نسبت به θ صفر است.
- logit را با a نشان دهید.

۵) اگر ساختار تخاصمی استاندارد (رابطه ۶) مورد استفاده باشد و توزیع داده آموزشی و شبکه مولد به‌صورت شکل زیر باشد، به سوالات زیر پاسخ دهید.



شکل ۲: توزیع داده اصلی (p_{Data}) و توزیع شبکه مولد (p_G)

الف) (۲ نمره) فواصل $JS(p_{Data}||p_G)$، $KL(p_{Data}||p_G)$ و $KL(p_G||p_{Data})$ را بیابید.
ب) (۱ نمره) در این حالت چه گرادیانی به شبکه مولد می‌رسد.
پ) (۲ نمره) یکی از راه‌های پایدارتر کردن آموزش شبکه‌های تخاصمی اضافه کردن نویز به تصاویر آموزشی ورودی تمییز دهنده است. با توجه به جواب بخش الف (فاصله JS)، اضافه کردن نویز، چطور به بهینه سازی کمک خواهد کرد؟
۶) (امتیازی) فاصله دیگری به نام فاصله واسرشتاین[۵] وجود دارد که آشنایی با آن به دلیل محبوبیتش در حوزه شبکه‌های تخاصمی مولد، خالی از لطف نیست. به زبان ساده، فاصله واسرشتاین بین دو توزیع کمترین هزینه جابجایی یک نوزیع به توزیع دیگر است. فاصله (واسرشتاین _۱) و به اختصار با W_1 نشان داده می‌شود، بین دو توزیع $p(x)$ و $q(x)$ را به صورت زیر تعریف می‌کنیم:

$$T(\gamma, c) = \int c(x,y)\gamma(x,y)dxdy \tag{۹}$$

$$W_1(p,q) = \inf_{\gamma \in \Gamma(p,q) T(\gamma,c)} \tag{۱۰}$$

که $c(x,y)$ هزینه جابجایی بین دو نقطه x و y (دارای خاصیت متر)، و $\Gamma(p,q)$ خانواده تمام روش‌های مختلف جابجایی[۶] بین این دو توزیع است. مقدار $\gamma(x,y)$ نشان‌دهنده مقداری است که می‌خواهید از نقطه x به y (از توزیع p به توزیع q) منتقل کنید. بنابراین $T(\gamma, c)$ هزینه جابجایی بین دو توزیع p و q تحت روش جابجایی $\gamma(x,y)$ خواهد بود.
الف) (۳ نمره) ضمن مطالعه این فاصله و یا مقاله‌های مرتبط و بلاگ‌های مرتبط با روش شبکه تخاصمی مولد واسرشتاین بیان کنید چه توابعی می‌توانند نشان‌دهنده یک روش جابجایی باشد؟ چه شروطی باید در این دسته توابع صدق کنند؟ ۲ شرط باید ذکر کنید.
ب) (۳ نمره) اگر $c(x,y) = |x-y|$ باشد، ضمن ذکر یک روش جابجایی بین دو توزیع (لزومی به کمینه بودن هزینه روش جابجایی نیست) p_G و p_{Data} سوال قبل (شش)، هزینه جابجایی بین آن‌ها را محاسبه کنید.
پ) (۲ نمره) با توجه به جواب قسمت ب، این فاصله چه مزیتی نسبت به فواصل JS و KL دارد؟

موفق باشید.

[۵] Wasserstein
[۶] Transport Plan