به نام خدا

مفاهیم پیشرفته در یادگیری ماشین

نیم‌سال دوم ۴۰۱ـ۱۴۰۰

مدرس: دکتر مهدیه سلیمانی

دانشکده‌ی مهندسی کامپیوتر

تمرین سری اول | عنوان تمرین | زمان تحویل: ۲۵ اسفندماه

لطفا نکات زیر را رعایت کنید:

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- در هر کدام از سوالات، اگر از منابع خاصی استفاده کرده‌اید باید آن را ذکر کنید.
- پاسخ ارسالی واضح و خوانا باشد.
- تمام پاسخ‌های خود را در یک فایل با فرمت HW#_[SID]_[Fullname].zip روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. علاوه بر آن، در هر تمرین می توانید تا سقف هفت روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید.
- برای کسب نمره کامل در این تمرین کافیست ۱۶۰ نمره را دریافت نمایید، ما بقی نمرات امتیازی می باشند (۴۰ نمره امتیازی).

سوال ۱: بهینه‌سازی در یادگیری چند وظیفه‌ای (۱۵ نمره)

مشکلی که متایادگیری مبتنی بر پروتوتایپ دارند این است که دسته‌بند ساده‌ای دارند و تعمیم‌پذیری دسته‌بند در فاز Meta-test کم می‌باشد. به همین دلیل در مقاله Bertinetto ۲۰۱۸ به بررسی دو دسته‌بند رایج در یادگیری ماشین (Logistic Regression, Ridge Regression) و نحوه استفاده موثر آن در مسئله متایادگیری پرداخته است. با مطالعه مقاله و راهنمایی های داده شده در زیر، به سوالات پاسخ دهید

(آ) نحوه استفاده از Ridge Regression برای دسته‌بندی را به طور کامل شرح دهید.

(ب) دو ماتریس $X \in ^{n\times d}$ و $Y \in ^{n\times l}$ را در نظر بگیرید که n تعداد داده‌ها و d اندازه بردار بازنمایی هر داده و l تعداد برچسب‌های موجود در دسته می‌باشد. ثابت کنید که دو رابطه زیر در صورتی که $\lambda > 0$ باشد، با هم برابر می‌باشند:

$$\left(X^T X + \lambda I\right)^{-1} X^T Y = X^T \left(X X^T + \lambda I\right)^{-1} Y$$

و بیان کنید که در ادبیات متایادگیری استفاده از کدام یک از دو رابطه بالا در حل مسئله بهتر می‌باشد و چرا؟

(ج) در روش Logistic Regression برای بدست آوردن دسته‌بند برخلاف روش قبلی نیازمند بهینه‌سازی می‌باشیم. در مقاله از Newton's Method برای بهینه‌سازی استفاده شده است. دلیل این امر را بیان کنید و همچنین در مورد خود Newton's Method تحقیق کنید و رابطه به روزرسانی و نحوه بدست آوردن این رابطه را بنویسید.

(د) با اعمال Newton's Method روی تابع هزینه این دسته‌بند، رابطه شماره ۷ مقاله که به‌روزرسانی وزن‌های مدل می‌باشد را بدست آورید.

پاسخ:

(آ) طبق توضیحات مقاله، تصویر ورودی توسط یک Feature Extractor به یک بردار نگاشت می‌شود و سپس با استفاده از ماژول Regressor Ridge پیش‌بینی مدل از روی فضای بازنمایی بدست می‌آید. مشکلی که ممکن است با آن مواجه شویم این است که خروجی مدل Regressor مناسب تابع هزینه دسته‌بندی نباشد. به همین یک تبدیل affine روی خروجی regressor اعمال می‌شود که پارامترهای آن می‌تواند به عنوان متاپارامتر در حلقه بیرونی الگوریتم متایادگیری بدست آید.

(ب) باتوجه به اینکه ماتریس Y در هردو طرف مساوی از سمت راست در معادله ضرب شده است، پس تنها کافی است که اثبات کنیم:

$$\left(X^T X + \lambda I\right)^{-1} X^T = X^T \left(X X^T + \lambda I\right)^{-1}$$

برای شروع اثبات در رابطه زیر را در نظر بگیرید:

$$\lambda X^T = \lambda X^T$$

ماتریس همانی را برای یک سمت از رابطه مساوی بالا از سمت چپ ماتریس X^T و یک بار از سمت راست ماتریس، ضرب می‌کنیم:

$$\lambda I_d X^T = \lambda X^T I_n$$

حال عبارت $X^T X X^T$ را به دو سمت مساوی اضافه می‌کنیم.

$$X^T X X^T + \lambda I_d X^T = X^T X X^T + \lambda X^T I_n$$

با فاکتورگیری داریم:

$$\left(X^T X + \lambda I_d\right) X^T = X^T \left(X X^T + \lambda I_n\right)$$

حال اگر در دو طرف تساوی، عبارت $\left(X^T X + \lambda I_d\right)^{-1}$ را از چپ و عبارت $\left(X X^T + \lambda I_n\right)^{-1}$ را از سمت راست ضرب کنیم داریم:

$$\left(X^T X + \lambda I_d\right)^{-1} \left(X^T X + \lambda I_d\right) X^T \left(X X^T + \lambda I_n\right)^{-1} = \left(X^T X + \lambda I_d\right)^{-1} X^T \left(X X^T + \lambda I_n\right) \left(X X^T + \lambda I_n\right)^{-1}$$

با ساده‌سازی داریم:

$$X^T \left(X X^T + \lambda I_n\right)^{-1} = \left(X^T X + \lambda I_d\right)^{-1} X^T$$

که مطلوب سوال می‌باشد.

این نکته لازم به ذکر می‌باشد که ماتریس‌هایی که معکوس آن را در اثبات بالا استفاده کردیم هردو ماتریس‌های مثبت نیمه‌معین می‌باشند و بنابراین حتما معکوس‌پذیر می‌باشند.

در ادبیات متایادگیری استفاده از رابطه $X^T \left(X X^T + \lambda I_n\right)^{-1}$ بهینه‌تر می‌باشد. چون در این رابطه نیاز به محاسبه معکوس یک ماتریس با ابعاد $n \times n$ داریم ولی در حالت دیگر نیاز به محاسبه معکوس ماتریسی به ابعاد $d \times d$ می‌باشد که مقدار d که بیانگر ابعاد بازنمایی در شبکه عصبی می‌باشد که به مراتب خیلی بزرگتر از تعداد نمونه‌ها در مسئله‌های متایادگیری می‌باشد که تعداد نمونه‌های آموزش خیلی کم می‌باشد.

(ج) به دلیل محدودیت تعداد به‌روزرسانی پارامترها در حلقه درونی الگوریتم،از Newton's Method برای همگرایی سریع‌تر به نسبت گرادیان گیری ساده استفاده شده است. در Newton's Method علاوه بر گرادیان، اطلاعات مشتق دوم نیز استفاده می‌شود. در این روش با نوشتن بسط Taylor حول یک نقطه داریم:

$$f\left(x+t\right) = f\left(x\right) + f'\left(x\right) t + \frac{f''\left(x\right) t^2}{2}$$

در رابطه بالا برای کمینه کردن مقدار تابع، نسبت به پارامتر t که جهت به‌روزرسانی و حرکت می‌باشد مشتق میگیریم که جهت بهینه به‌روزرسانی را پیدا کنیم:

$$\frac{dy}{dx}\left(f\left(x\right) + f'\left(x\right) t + \frac{f''\left(x\right) t^2}{2}\right) = f'\left(x\right) + f''\left(x\right) t = 0$$

$$t^* = -\frac{f'\left(x\right)}{f''\left(x\right)}$$

(د) هدف سوال بیشینه‌کردن بیشینه درست‌نمایی می‌باشد.

$$P\left(Y|\omega, X\right) = \prod_{i=1}^{N} \left(\sigma\left(\omega^T x_i\right)\right)^{y_i} \left(1 - \sigma\left(\omega^T x_i\right)\right)^{1-y_i}$$

$$L = -\prod_{i=1}^{N} \left(\sigma\left(\omega^T x_i\right)\right)^{y_i} \left(1 - \sigma\left(\omega^T x_i\right)\right)^{1-y_i} = -\sum_{i=1}^{N} \left[y_i \log\left(\sigma\left(\omega^T x_i\right)\right) + \left(1 - y_i\right) \log\left(1 - \sigma\left(\omega^T x_i\right)\right)\right]$$

حال با گرادیان گیری مقادیر گرادیان اول و دوم را حساب می‌کنیم:

$$\nabla L = \sum_{i=1}^{N} \left(\sigma\left(\omega^T x_i\right) - y_i\right) {x_i}^T$$

$$\nabla\nabla L = \sum_{i=1}^{N} \sigma\left(\omega^T x_i\right) \left(1 - \sigma\left(\omega^T x_i\right)\right) {x_i}^T x_i$$

ابتدا چند ماتریس تعریف می‌کنیم:

$$A_t = diag(q_t)$$
$$q_t^{(i)} = \sigma\left(\omega_t^T x^{(i)}\right)\left(1 - \sigma\left(\omega_t^T x^{(i)}\right)\right)$$
$$B_t^{(i)} = \sigma\left(\omega_t^T x^{(i)}\right) - y^{(i)}$$

که $A \in {}^{n\times n}$ یک ماتریس قطری و $B \in {}^{n}$ می‌باشد. حال اگر سیگماهای بالا را به ضرب ماتریسی تبدیل کنیم، طبق رابطه به‌روزرسانی داریم:

$$\begin{aligned}\omega_{t+1} &= \omega_t - H^{-1}\nabla L \\ &= \omega_t - \left(X^T A_t X + \lambda I\right)^{-1}\left(X^T B_t + \lambda\omega_t\right) \\ &= \left(X^T A_t X + \lambda I\right)^{-1}\left(\left(X^T A_t X + \lambda I\right)\omega_t - X^T B_t - \lambda\omega_t\right) \\ &= \left(X^T A_t X + \lambda I\right)^{-1}\left(X^T A_t X\omega_t - X^T B_t\right)\end{aligned}$$

سوال ۲: (نظری) تنظیم توزیع برای یادگیری چندنمونه‌ای (۴۰ نمره)

یکی از ریسک‌های احتمالی در یادگیری چندنمونه‌ای احتمال بیش برازش بر روی دادگان کم‌تعداد آموزشی است. در این مقاله روشی پیشنهاد شده است تا به کمک استخراج مشخصات آماری کلاس‌های حاضر در متاآموزش بتوان توزیع دادگان کلاس‌های حاضر در متاتست را تنظیم کرد. این مقاله را به دقت خوانده و به سوالات زیر به طور کامل پاسخ دهید:

(آ) از آنجایی که ممکن است توزیع دادگان هر کلاس حاضر در متاآموزش گاوسی نباشد و دارای مقداری کشیدگی باشد؛ در نظر گرفتن این توزیع‌ها به عنوان توزیع گاوسی و استخراج میانگین و کواریانس از آن‌ها میتواند اشتباه باشد. توضیح دهید این مقاله چه روشی را برای حل مشکل کشیدگی توزیع دادگان متاآموزش اتخاذ کرده است و چگونه این روش موجب حل مشکل کشیدگی توزیع می‌شود؟

(ب) پس از استخراج میانگین و کواریانس کلاس‌های حاضر در متاآموزش، مدل ارائه شده اقدام به تنظیم توزیع دادگان حاضر در متاتست می‌کند. به صورت کامل و با نوشتن روابط ریاضی مربوطه بیان کنید این تنظیم توزیع به چه صورت انجام می‌پذیرد و وجود کلاس‌های مشابه در متاترین به کلاس منظور در متاتست چه کمکی به تنظیم توزیع می‌کند؟

(ج) در تنظیماتی که در هنگام متاتست از هر کلاس بیش از یک نمونه آموزش داشته باشیم این مدل به جای میانگین‌گیری از نمونه‌ها، برای هر کدام از k نمونه آموزش اقدام به تنظیم توزیع جداگانه می‌کند. توضیح دهید توزیع تنظیم جداگانه چه مزیتی نسبت به میانگین گیری نمونه‌ها و سپس یک توزیع تنظیم دارد؟

سوال ۳: (عملی) یادگیری چندنمونه‌ای از طریق یادگیری متر (۴۰ نمره)

در این سوال قصد داریم تا مدل یادگیرنده شبکه Prototypical را مورد پیاده‌سازی و بررسی قرار دهیم. به این منظور هر دو زیر مجموعه‌های آموزش و تست دادگان CIFAR100 را دریافت کرده و سپس آن‌ها را به یکدیگر الحاق نمایید. سپس این دادگان را به سه زیر مجموعه متاآموزش، متااعتبارسنجی و متاتست تقسیم کنید. به این صورت که دادگان آموزش شامل دادگان ۷۰ کلاس، دادگان اعتبارسنجی شامل ۲۰ کلاس و دادگان کلاس تست شامل ۱۰ دیگر باشند. در گام بعدی بایستی یک Sampler پیاده‌سازی کنید که با گرفتن پارامترهای Way و Shot بتواند دادگان اتکا و پرسمان برای یک وظیفه را تولید کنند (در واقع این ماژول با هر فراخوانی دو مجموعه داده به اندازه $Way * Shot$ برای اتکا و پرسمان خروجی میدهد). این ماژول در واقع هر بار یک اپیزود را تولید میکند. در طی آزمایش‌های زیر از ماژول ProtoNetBack که در فایل‌های پیوستی قرار داده شده است به عنوان شبکه تکیه استخراج ویژگی استفاده کنید و در جلوی آن دو لایه تمام متصل با اندازه دلخواه قرار دهید.

(آ) دسته‌بند را با تنظیمات 8-shot, 10-way آموزش دهید و سپس دقت مدل را بر روی دادگان متاتست گزارش دهید. انتظار می‌رود دقت در این بخش بیشتر از ۵۰ درصد باشد.

(ب) به ازای هر یک از تنظیمات $shot \in \{1, 2, 4, 8, 16\}$ و 10-way آزمایش بالا را تکرار کرده و نمودار دقت متاتست بر حسب shot را رسم نمایید.

(ج) حال به ازای هر یک از تنظیمات $way \in \{2, 4, 8, 16, 32\}$ و 5-shot آزمایش را تکرار کرده و نمودار دقت متاتست بر حسب way را رسم نمایید. (دقت کنید که در هنگام متاتست از تنظیمات 5-shot, 10-way استفاده نمایید)

(د) حال در هنگام متاآموزش با تنظیمات 10-shot, 10-way دسته‌بند را آموزش دهید. در هنگام متاتست اما متاتست را به ازای هر یک از تنظیمات $shot \in \{1, 5, 10, 15, 20\}$ انجام دهید و نمودار دقت آن را بر حسب shot رسم نمایید.

سوال ۴: (عملی) متایادگیری براساس بهینه‌سازی (۴۰ نمره)

در این سوال قصد داریم تا مدل معروف دسته متایادگیری براساس بهینه‌سازی،MAML، را پیاده‌سازی نماییم. مقاله مرتبط با این کار، این مقاله می‌باشد. در Notebook داده شده تمام پارامترهای مسئله و مراحل حل به صورت گام به گام تشریح شده است. سوال از دو بخش اصلی تشکیل

شده است که در بخش اول به دلیل کاهش هزینه آموزش بخش عمده شبکه به صورت pretrained شده در اختیار شما قرار داده شده است و شما تنها روی بخش مشخص شده شبکه فرایند متایادگیری را انجام خواهید داد. در بخش اول قرار است تاثیر تعداد گام‌های به‌روزرسانی مدل در حلقه داخلی الگوریتم، مورد بررسی قرار گیرد. از شما خواسته شده است که به ازای مقادیر ۱ تا ۳ این مورد را انجام دهید و نتیجه هر حالت را مقایسه و گزارش کنید. در بخش دوم نیز از شما خواسته شده است که حال با یک گام به‌روزرسانی حلقه داخلی، کل ساختار مدل (مدل متایادگیری بخش اول + ساختار مدل pretrained داده شده) را به صورت متاپارامتر در نظر بگیرید و متایادگیری را روی آن انجام دهید. در نهایت نتایج بدست آمده از هردو بخش را تحلیل و گزارش نمایید.