به نام خدا

دانشکده‌ی مهندسی کامپیوتر

مفاهیم پیشرفته در یادگیری ماشین

نیم‌سال دوم ۴۰۱ـ۱۴۰۰

مدرس: دکتر مهدیه سلیمانی

تمرین سری اول | عنوان تمرین | زمان تحویل: ۱۹ فروردین

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.
- نمره این تمرین از ۲۰۰ محاسبه می‌شود.

سوال ۱: (نظری) متاـ یادگیری مبتنی بر بهینه‌سازی دو سطحی (۲۵ نمره)

همانطور که در جلسات درس مشاهده کرده‌اید، یکی از روش‌های متاـ یادگیری [۱] ، خانواده بهینه‌سازی دو سطحی [۲] بوده که مهمترین کار در این زمینه روش MAML [۳] می‌باشد. در این خانواده از روش‌ها، متاـ پارامترها [۴] (پارامترهای آهسته) همبعد با پارامترهای مختص وظیفه [۵] (پارامترهای سریع) بوده و به عنوان یک نقطه شروع برای کل وزن‌های شبکه عمل می‌کنند. به طور دقیق‌تر، اگر توزیع وظایف [۶] را به صورت $p(\mathcal{T})$ در نظر بگیریم، می‌توان رابطه زیر را برای یادگیری متاـ پارامترها ارائه داد:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\mathcal{T}=(S,Q)\sim p(\mathcal{T})}[\mathcal{L}(\phi, Q)] \quad (۱آ)$$

$$Where \quad \phi = \mathcal{A}\text{lg}(\theta, S) \quad (۱ب)$$

که در این رابطه S مجموعه داده‌های پشتیبان [۷] و Q مجموعه داده‌های پرسمان [۸] مربوط به هر وظیفه [۹] را نشان می‌دهد. در این رابطه، محاسبه پارامترهای سریع ϕ توسط روش $\mathcal{A}\text{lg}$ انجام می‌شود که در مقالات مربوطه به طرق مختلفی انتخاب شده و بهینه‌سازی داخلی [۱۰] نامیده می‌شود. همچنین در رابطه فوق، بهینه‌سازی خارجی [۱۱] که روی پارامترهای θ صورت می‌پذیرد، به شکل $\arg\min$ نمایش داده شده است.
برای انجام بهینه‌سازی خارجی، لازم است تا از تابع معرفی شده نسبت به θ گرادیان را به صورت زیر محاسبه کرده و θ را از طریق آن به‌روزرسانی نماییم:

$$\nabla_\theta \mathbb{E}[\mathcal{L}(\phi, Q)] = \mathbb{E}[\nabla_\theta \mathcal{L}(\mathcal{A}\text{lg}(\theta, S), Q)] \quad (۲)$$

برای محاسبه عبارت فوق، از قاعده زنجیره‌ای مشتق استفاده می‌کنیم:

$$\nabla_\theta \mathcal{L}(\mathcal{A}\text{lg}(\theta, S), Q) = \nabla_\phi \mathcal{L}(\phi, Q)|_{\phi=\mathcal{A}\text{lg}(\theta,S)} \times \frac{d}{d\theta}\mathcal{A}\text{lg}(\theta, S) \quad (۳)$$

همانطور که مشاهده می‌کنید، رابطه فوق از دو جمله تشکیل شده است؛ محاسبه جمله اول نسبتاً راحت است. چرا که کافیست تا از $\mathcal{L}$ مشتق گرفته و مقدار $\mathcal{A}\text{lg}(\theta, S)$ را در آن جایگذاری کنیم و در این صورت گرادیانی از $\mathcal{A}\text{lg}(\theta, S)$ عبور نمی‌کند. این در حالیست که برای محاسبه جمله دوم،

[1] Meta-Learning
[2] Bi-Level Optimization
[3] Model Agnostic Meta-Learning
[4] Meta-Parameters
[5] Task-Specific Parameters
[6] Tasks
[7] Support
[8] Query
[9] Task
[10] Inner-Level Optimization
[11] Outer-Level Optimization

لازم است تا عملیات مشتق‌گیری را از داخل الگوریتم $\mathcal{A}\text{lg}(\theta, S)$ عبور دهیم. این مسئله می‌تواند مشکل‌زا باشد چرا که ممکن است $\mathcal{A}\text{lg}(\theta, S)$ اصلاً قابلیت عبور گرادیان را نداشته باشد یا اگر دارد ممکن است منجر به محاسبات پرهزینه شود. به عنوان مثال، در روش MAML داریم:

$$\mathcal{A}\text{lg}(\theta, S) = \theta - \alpha \nabla_\theta \mathcal{L}(\theta, Q) \Rightarrow \frac{d}{d\theta}\mathcal{A}\text{lg}(\theta, S) = I - \alpha \nabla^2_\theta \mathcal{L}(\theta, S) \quad (۴)$$

که در رابطه فوق، محاسبه مشتق‌های مرتبه دوم (Hessian) می‌تواند بسیار سخت یا پرهزینه باشد چرا که لازم است تا کل گراف محاسباتی در طول مسیر محاسبه $\mathcal{A}\text{lg}(\theta, S)$ نگه داشته شود تا بتوان گرادیان را از روی آن عبور داده و به مراحل قبلی رساند.
در این سوال قصد داریم تا تکنیکی معرفی کنیم که این مشکلات مرتفع شوند. برای این منظور، فرض کنید $\mathcal{A}\text{lg}(\theta, S)$ به صورت زیر پیشنهاد شده است:

$$\phi = \mathcal{A}\text{lg}(\theta, S) = \arg\min_{\phi'} \mathcal{L}(\phi', S) + \frac{\lambda}{2}||\phi' - \theta||^2 \quad (۵)$$

که λ یک هایپرپارامتر است و هر چه مقدار آن بیشتر باشد، باعث می‌شود تا جواب بهینه‌سازی درونی، به نقطه شروع خود یعنی θ نزدیک‌تر بماند. بهینه‌سازی ۵ را می‌توان با انجام چندین گام بهینه‌سازی تکرار شونده [۱۲] حل کرد اما مشکل آن جاست که امکان عبور گرادیان از چنین محاسباتی وجود ندارد. با در نظر گرفتن این مسئله، به پرسش‌های زیر پاسخ دهید:

(آ) فرض کنید ما قادر هستیم تا بهینه‌سازی ۵ را به صورت کامل حل کنیم و ϕ را به عنوان جواب بهینه دقیق آن به دست آورده‌ایم. از این مسئله استفاده کنید و $\frac{d\phi}{d\theta}$ را محاسبه کنید (راهنمایی: در نقطه بهینه دقیق، مشتق $\mathcal{L}(\phi', S) + \frac{\lambda}{2}||\phi' - \theta||^2$ نسبت به ϕ' صفر خواهد بود). (۱۰ نمره)

(ب) اگر مراحل پرسش قبل را به درستی طی کرده باشید، در جواب خود به یک عبارت حاوی مشتق مرتبه دوم $\nabla^2 \mathcal{L}$ (یا همان ماتریس Hessian) می‌رسید. محاسبه این عبارت چه تفاوتی با مشتق مرتبه دوم موجود در رابطه ۴ دارد؟ استفاده از این تابع $\mathcal{A}\text{lg}$ پیشنهادی چه مزیتی نسبت به MAML دارد؟ (۱۰ نمره)

(ج) به عنوان جمع‌بندی، الگوریتمی که متا_یادگیر در هر اپیزود طی می‌کند را به صورت گام گام شرح دهید. (۵ نمره)

سوال ۲: (نظری) روش‌های مبتنی بر یادگیری متریک (۵۵ نمره)

روش‌های یادگیری متریک یکی دیگر از روش‌هایی هستند که در درس به عنوان یکی از اعضای خانواده متا_یادگیری با آن‌ها آشنا شدید. در این روش‌ها هدف آن است که یک شبکه استخراج ویژگی مثل $f_\theta(x)$ یاد گرفته شود تا داده‌ها با برچسب یکسان را در فضای نمایش مخفی در کنار یکدیگر تصویر کند. پارامترهای θ متا_پارامترهای مدل در نظر گرفته شده و در طول متا_یادگیری آموزش داده می‌شوند. از آنجایی که $f_\theta(x)$ صرفاً یک فضای نمایشی [۱۳] فراهم می‌کند، برای کامل کردن شبکه نیاز به یک دسته‌بند (یادگیر پایه [۱۴]) پارامتریک یا نان_پارامتریک داریم که روی این فضای نمایشی قرار گرفته، از داده‌های پشتیبان برای آماده‌سازی خود استفاده کرده و با کمک آن‌ها عمل دسته‌بندی نمونه‌های پرسمان را انجام دهد. پارامترهای دسته‌بند معرفی شده را به عنوان پارامترهای سریع می‌شناسند و آن‌ها را با ϕ نمایش می‌دهند. این پارامترها مختص هر وظیفه به دست آمده و برای وظیفه بعدی تغییر می‌کنند.
در این روش‌ها یکی از تصمیمات مهم در زمینه طراحی الگوریتم انتخاب مناسب همین دسته‌بند می‌باشد. از جمله انتخاب‌های موجود برای این خانواده، انتخاب روش Nearest Neighbour می‌باشد. همچنین روش دیگری که در مقاله ProtoNet معرفی شد استفاده از دسته‌بندهای مبتنی بر پروتوتایپ دسته‌ها (با میانگین گیری از داده‌های موجود در مجموعه پشتیبان از هر کلاس) می‌باشد. مشکلی که متا_یادگیری مبتنی بر پروتوتایپ دارند این است که دسته‌بند ساده‌ای دارند و تعمیم‌پذیری دسته‌بند در فاز متا_ارزیابی [۱۵] کم می‌باشد. به همین دلیل در مقاله Bertinetto 2018 به بررسی دو دسته‌بند رایج در یادگیری ماشین (Logistic Regression, Ridge Regression) و نحوه استفاده موثر آن در مسئله متا_یادگیری پرداخته شده است. با مطالعه مقاله و راهنمایی های داده شده در زیر، به سوالات پاسخ دهید:

(آ) توضیح دهید که در نگاه اول، استفاده از این دسته‌بندها چه مشکلی می‌تواند برای فرایند متا_آموزش [۱۶] ایجاد کند؟ چرا استفاده از رویکردهای پروتوتایپی یا KNN این مشکل را ایجاد نمی‌کند؟ (راهنمایی: پاسخ این مورد غیر مرتبط با پرسش قبل نیست) (۱۰ نمره)

(ب) توضیح دهید که در مقاله معرفی شده، چگونه مشکل معرفی شده را حل می‌کند؟ تفاوت رویکردی که برای Ridge Regression و Logistic Regression به کار گرفته می‌شود را توضیح دهید. (۵ نمره)

[12] Iterative
[13] Representation Space
[14] Base-Learner
[15] Meta-Test
[16] Meta-Training

(ج) دو ماتریس $X \in \mathbb{R}^{n \times d}$ به عنوان داده‌های پشتیبان تعبیه شده در فضای نمایش و $Y \in \mathbb{R}^{n \times l}$ به عنوان برچسب‌های One-hot همان داده‌ها 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 و نحوه استفاده آن در دسته‌بند Ridge Regression را توضیح دهید : (۱۰ نمره)

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

(د) توضیح دهید که برای انجام متاـ یادگیری استفاده از کدام یک از دو رابطه بالا بهتر می‌باشد و چرا؟ (۵ نمره)

(ه) در مقاله معرفی شده، برای محاسبه وزن‌های دسته‌بند Logistic Regression از بهینه‌سازی با روش Newton 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. (۱۰ نمره)

(و) تعاریف زیر را در نظر بگیرید:

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

که $A \in \mathbb{R}^{n \times n}$ یک ماتریس قطری و $B \in \mathbb{R}^n$ می‌باشد.

با اعمال Newton's Method روی تابع هزینه این دسته‌بند، به رابطه به‌روزرسانی زیر برسید (۱۵ نمره) :

$$\omega_{t+1} = \left(X^T A_t X + \lambda I\right)^{-1} X^T \left(A_t X \omega_t - B_t\right)$$

(رابطه فوق مشابه رابطه ۷ مقاله می‌باشد با این تفاوت که به نظر می‌رسد روابط موجود در مقاله در برخی جزئیات از نظر Notation ایراد دارند.)

سوال ۳: (نظری) استفاده از دسته‌بند SVM در رویکرد یادگیری متریک (۳۰ نمره)

در این سوال قصد داریم مقاله MetaOptNet را مورد بررسی قرار دهیم که برای مدت قابل توجهی SOTA [۱۷] در زمینه یادگیری با تعداد نمونه کم [۱۸] به شمار می‌رفت. این مقاله شباهت بسیار زیادی به مقاله معرفی شده در پرسش قبل دارد با این تفاوت که قصد دارد به جای دسته‌بندهای معرفی شده، دسته‌بند SVM را به عنوان لایه آخر روی شبکه استخراج ویژگی سوار کند. رویکردهایی که در پرسش قبل برای Ridge Regression و Logistic Regression معرفی شدند را نمی‌توان برای SVM به کار برد. لذا این مقاله به دنبال ارائه روشی است تا این مشکل را برطرف نماید. برای این منظور، رویکردی شبیه به پرسش اول را دنبال می‌کند. با این تفاوت که در پرسش اول کل وزن‌های شبکه (اعم از دسته‌بند و Backbone) در دستگاه بهینه‌سازی قرار می‌گرفتند اما در این سوال دستگاه بهینه‌سازی فقط روی وزن‌های دسته‌بند SVM نوشته می‌شود. برای درک بهتر روش، مقاله مورد نظر را بررسی نموده و به پرسش‌های زیر پاسخ دهید:

(آ) توضیح دهید که پارامترهای سریع چگونه به کمک داده‌های پشتیبان و پارامترهای آهسته ساخته می‌شوند و دستگاه بهینه‌سازی معرفی شده در این مقاله که از حل آن پارامترهای سریع ساخته می‌شوند را به همراه دوگان آن به صورت دقیق و با ذکر جزئیات نمادگذاری معرفی کنید. (۵ نمره)

(ب) با مطالعه صفحه ۴ این مقاله، توضیح دهید که استفاده از قضایای KKT و Implicit Function Theorem چه کمکی در راستای محاسبه گرادیان می‌کنند و آپدیت شبکه Backbone با استفاده از چه گرادیانی انجام می‌شود؟ (برای این قسمت اثبات دقیق ریاضی مد نظر نیست و به شرطی که به صورت شفاف کاربرد این دو قضیه و نحوه استفاده آن‌ها را بیان کنید نمره کامل را دریافت می‌کنید) (۱۰ نمره)

(ج) با مطالعه روابط این مقاله ضمن نوشتن دستگاه بهینه‌سازی مربوطه، توضیح دهید که چگونه می‌توان دسته‌بند Ridge Regression مطرح شده در سوال قبل را ذیل همین رویکرد جای داد. (۵ نمره)

(د) بعد از به دست آوردن وزن‌های بهینه w برای دسته‌بند SVM یا Ridge Regression ، تابع متاـ هزینه به چه صورتی نوشته می‌شود؟ از این تابع هزینه برای به‌روزرسانی کدام پارامترها استفاده می‌شود؟ (از رابطه ۱۲ مقاله کمک بگیرید، ولی جزئیات به دست آوردن آن را به صورت شفاف بیان کنید) (۱۰ نمره)

توجه: در ادبیات این مقاله، نمادگذاری رایج متاـ یادگیری رعایت نشده است و جای نماد θ و ϕ با هم عوض شده است. برای یکسان شدن جواب‌ها، شما از نمادگذاری معرفی شده در پرسش اول استفاده کرده و θ و ϕ را به ترتیب برای متاـ پارامترها و پارامترهای مختص وظیفه به کار ببرید.

[17] State-of-the-Art
[18] Few-Shot Learning

سوال ۴: (نظری) تنظیم توزیع برای یادگیری چندنمونه‌ای (۱۵ نمره)

یکی از ریسک‌های احتمالی در یادگیری چندنمونه‌ای احتمال بیش برازش [۱۹] بر روی دادگان کم‌تعداد آموزشی است. در این مقاله روشی پیشنهاد شده است تا به کمک استخراج مشخصات آماری کلاس‌های حاضر در متا_آموزش بتوان توزیع دادگان کلاس‌های حاضر در متا_ارزیابی 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 متا_آموزش 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 متا_آموزش اتخاذ کرده است و چگونه این روش موجب حل مشکل کشیدگی توزیع می‌شود؟(۵ نمره)

(ب) پس از استخراج میانگین و کواریانس کلاس‌های حاضر در متا_آموزش، مدل ارائه شده اقدام به تنظیم توزیع دادگان حاضر در متا_ارزیابی 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 متا_آموزش به کلاس منظور در متا_ارزیابی چه کمکی به تنظیم توزیع می‌کند؟ (۵ نمره)

(ج) در تنظیماتی که در هنگام متا_ارزیابی 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؟(۵ نمره)

سوال ۵: (عملی) یادگیری چندنمونه‌ای از طریق یادگیری متریک (۲۵ 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 متا_آموزش، متا_اعتبارسنجی [۲۰] و متا_ارزیابی 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 متا_ارزیابی 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 متا_ارزیابی 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 متا_ارزیابی بر حسب way را رسم نمایید. (دقت کنید که در هنگام متا_ارزیابی از تنظیمات 5-shot, 10-way استفاده نمایید)

(د) حال در هنگام متا_آموزش با تنظیمات 10-shot, 10-way دسته‌بند را آموزش دهید. در هنگام متا_ارزیابی اما متا_ارزیابی 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 ۶: (عملی) متا_یادگیری براساس بهینه‌سازی (۵۰ نمره)

در این سوال قصد داریم تا مدل معروف دسته متا_یادگیری 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 متا_یادگیری 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 متا_یادگیری بخش اول + ساختار مدل pretrained داده شده) را به صورت متاپارامتر در نظر بگیرید و متا_یادگیری را روی آن انجام دهید. در نهایت نتایج بدست آمده از هردو بخش را تحلیل و گزارش نمایید.

[19] Overfitting
[20] Meta-Validation