



مفاهیم پیشرفته در یادگیری ماشین

نیم سال دوم ۱۴۰۰-۴۰۱

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 [Fullname]_[SID]_HW#.zip روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. علاوه بر آن، در هر تمرین می‌توانید تا سقف هفت روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید.
- برای کسب نمره کامل در این تمرین کفایت تا ۱۶۰ نمره از ۲۰۰ نمره را کسب نمایید (۴۰ نمره امتیازی).

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

همانطور که در جلسات درس مشاهده کرده‌اید، یکی از روش‌های متا-یادگیری^۱، خانواده بهینه‌سازی دو سطحی^۲ بوده که مهمترین کار در این زمینه روش MAML^۳ می‌باشد. در این خانواده از روش‌ها، متا-پارامترها^۴ (پارامترهای آهسته) همبند 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 $p(T)$ در نظر بگیریم، می‌توان رابطه زیر را برای یادگیری متا-پارامترها ارائه داد:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{T=(S,Q) \sim p(T)} [\mathcal{L}(\phi, Q)] \quad (1)$$

$$\text{Where } \phi = \text{Alg}(\theta, S) \quad (1b)$$

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 Alg 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:

$$\nabla_{\theta} \mathbb{E}[\mathcal{L}(\phi, Q)] = \mathbb{E}[\nabla_{\theta} \mathcal{L}(\text{Alg}(\theta, S), Q)] \quad (2)$$

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

$$\nabla_{\theta} \mathcal{L}(\text{Alg}(\theta, S), Q) = \nabla_{\phi} \mathcal{L}(\phi, Q)|_{\phi=\text{Alg}(\theta, S)} \times \frac{d}{d\theta} \text{Alg}(\theta, S) \quad (3)$$

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

¹Meta-Learning²Bi-Level Optimization³Model Agnostic Meta-Learning⁴Meta-Parameters⁵Task-Specific Parameters⁶Tasks⁷Support⁸Query⁹Task¹⁰Inner-Level Optimization¹¹Outer-Level Optimization

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

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

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 $\text{Alg}(\theta, S)$ 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 $\text{Alg}(\theta, S)$ به صورت زیر پیشنهاد شده است:

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

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 Alg 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 میانگین‌گیری 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 راهنمایی‌های 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)

¹²Iterative

¹³Representation Space

¹⁴Base-Learner

¹⁵Meta-Test

¹⁶Meta-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 برجسب‌های One-hot همان داده‌ها را در نظر بگیرید که n تعداد داده‌ها و d اندازه بردار بازنمایی هر داده و l تعداد برجسب‌های 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)

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

(د) توضیح دهید که برای انجام متا-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:

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

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):

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

(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 متا-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 متا-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 متا-پارامترها و پارامترهای مختص وظیفه به کار ببرید.

¹⁷State-of-the-Art

¹⁸Few-Shot Learning

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

یکی از ریسک‌های احتمالی در یادگیری چندمنونه‌ای احتمال بیش برآش^{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 میانگین‌گیری نمونه‌ها و سپس یک توزیع تنظیم دارد؟ (۵ نمره)

سوال ۵: (عملی) یادگیری چندمنونه‌ای 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 کند (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 **Backbone** استخراج ویژگی استفاده کنید و در جلوی آن دو لایه تمام متصل^{۲۱} 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 هر دو بخش را تحلیل و گزارش نمایید.

¹⁹Overfitting

²⁰Meta-Validation

²¹Fully Connected