به نام خدا

یادگیری عمیق

پاییز ۱۴۰۳

استاد: دکتر فاطمی‌زاده

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی برق

گردآورندگان: محمد جواد امین ، مهشاد مرادی

تمرین اول | **مفاهیم پایه** | مهلت ارسال: چهارشنبه ۲ آبان

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۵ روز و در مجموع ۱۵ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- همکاری و همفکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد. (دقت کنید در صورت تشخیص مشابهت غیرعادی برخورد جدی صورت خواهد گرفت.)
- در صورت همفکری و یا استفاده از هر منابع خارج درسی، نام همفکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- نتایج و پاسخ های خود را در یک فایل با فرمت zip به نام HW۱-Name-StudentNumber در سایت CW قرار دهید. برای بخش عملی تمرین نیز در صورتی که کد تمرین و نتایج خود را در گیت‌هاب بارگذاری می‌کنید، لینک مخزن مربوطه (repository) را در پاسخنامه خود قرار بدهید. دقت کنید هر سه فایل نوتبوک تکمیل شده بخش عملی را در گیت‌هاب قرار دهید. همچنین لازم است تا دسترسی های لازم را به دستیاران آموزشی مربوط به این تمرین بدهید.
- لطفا تمامی سوالات خود را از طریق صفحه درس در سایت Quera مطرح کنید (برای اینکه تمامی دانشجویان به پاسخ‌های مطرح شده به سوالات دسترسی داشته باشند و جلوی سوالات تکراری گرفته شود، به سوالات در بسترهای دیگر پاسخ داده نخواهد شد).
- دقت کنید کدهای شما باید قابلیت اجرای دوباره داشته باشند، در صورت دادن خطا هنگام اجرای کدتان، حتی اگه خطا بدلیل اشتباه تایپی باشد، نمره صفر به آن بخش تعلق خواهد گرفت.

سوالات نظری (۳۰۰ نمره)

۱. (۶۰ نمره)

(آ) در حالت قابل تفکیک خطی، اگر یکی از نمونه‌های آموزشی حذف شود، آیا مرز تصمیم به سمت نقطه حذف شده حرکت می‌کند، از آن دور می‌شود یا همان‌طور باقی می‌ماند؟ پاسخ خود را توجیه کنید. اکنون اگر مرز تصمیم را برای رگرسیون لجستیک در نظر بگیریم، آیا مرز تصمیم تغییر خواهد کرد یا همان‌طور باقی می‌ماند؟ پاسخ خود را توضیح دهید. (نیازی به ذکر جهت تغییر نیست)

(ب) i. از یادداشت‌های درس به یاد آورید که اگر ما اجازه دهیم برخی از طبقه‌بندی‌ها در داده‌های آموزشی اشتباه باشند، بهینه‌سازی SVM (حاشیه نرم) به شکل زیر است:

$$\min_{\omega,\xi_i} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{s.t.} \quad y_i(w^\top(x_i)) \geq 1-\xi_i, \quad \forall i\in\{1,\dots,n\} \tag{2}$$

$$\xi_i \geq 0, \quad \forall i\in\{1,\dots,n\} \tag{3}$$

که در آن $\xi_n , \dots , \xi_1$ به عنوان متغیرهای نرم نامیده می‌شوند. فرض کنید $\xi_n , \dots , \xi_1$ بهینه محاسبه شده‌اند. از ξ_i برای به‌دست آوردن یک حد بالا برای تعداد نمونه‌های طبقه‌بندی شده نادرست استفاده کنید.

ii. در بهینه‌سازی SVM، نقش ضریب C چیست؟ پاسخ خود را به طور مختصر با در نظر گرفتن دو حالت افراطی توضیح دهید، یعنی $C \to 0$ و $C \to \infty$.

iii. SVM سخت و رگرسیون لجستیک را هنگامی که دو کلاس قابل تفکیک خطی هستند، مقایسه کنید. هر تفاوت قابل توجهی را بیان کنید. (*راهنما* - به مرز تصمیم فکر کنید)

iv. SVM نرم و رگرسیون لجستیک را هنگامی که دو کلاس قابل تفکیک خطی نیستند، مقایسه کنید. هر تفاوت قابل توجهی را بیان کنید.

۲. (۶۰ نمره)

(آ) فرض کنید در PCA ، هر نقطه x_i را به $z_i = V_{1:k}^\top x_i$ پروجکت می‌کنیم، که در آن $V_{1:k} = [v_1, \dots , v_k]$، یعنی اولین k مؤلفه های اصلی. ما می‌توانیم x_i را از z_i به صورت زیر بازسازی کنیم: $\hat{x}_i = V_{1:k} z_i$.

i. نشان دهید که

$$\|\hat{x}_i - \hat{x}_j\| = \|z_i - z_j\|$$

می‌باشد.

ii. نشان دهید که خطا در بازسازی برابر است با:

$$\sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 = (n-1) \sum_{j=k+1}^{p} \lambda_j$$

که در آن $\lambda_p , \dots , \lambda_{k+1}$ به عنوان $p-k$ کوچک‌ترین مقادیر ویژه هستند. بنابراین، هرچه مؤلفه‌های اصلی بیشتری برای بازسازی استفاده کنیم، دقت آن بیشتر است.

۳. (۶۰ نمره)

فرض کنید معادله $Xw = y$ را داریم، که در آن $X \in \mathbb{R}^{m\times n}$ یک ماتریس داده‌ها غیرمربع است، w یک بردار وزن است و y بردار برچسب‌ها است که به هر داده در هر سطر X متناظر است.

فرض کنید $X = U\Sigma V^T$ (SVD کامل از X) باشد. در اینجا، U و V ماتریس‌های مربعی و متعامد هستند و Σ یک ماتریس $m \times n$ با مقادیر منفرد (σ_i) غیرصفر بر روی قطر است.

برای این مسئله، $\Sigma^\dagger$ به عنوان یک ماتریس $n \times m$ با وارون مقادیر منفرد ($\frac{1}{\sigma_i}$) در طول قطر تعریف می‌شود.

(الف) ابتدا، حالتی را در نظر بگیرید که $m > n$، یعنی ماتریس داده ما X دارای تعداد سطرهای بیشتری نسبت به ستون‌ها است (ماتریس $tall$) و سیستم $overdetermined$ است. چگونه وزن‌های w را پیدا کنیم که خطا بین Xw و y را به حداقل برساند به عبارت دیگر، می‌خواهیم $\min \|Xw - y\|^2$ را حل کنیم.

(ب) از $X = U\Sigma V^\top$ SVD استفاده کنید و ساده‌سازی کنید.

(ج) توجه خواهید کرد که راه‌حل کمترین مربعات به فرم $w^* = Ay$ است. چه اتفاقی می‌افتد اگر X را از سمت چپ در ماتریس A ضرب کنیم؟ به همین دلیل ماتریس A راه‌حل کمترین مربعات را معکوس چپ نام‌گذاری می‌کنند.

(د) حالا، حالتی را در نظر بگیرید که $m < n$، یعنی ماتریس داده X تعداد ستون‌های بیشتری نسبت به سطرها دارد و سیستم $underdetermined$ است. راه‌حل‌های بی‌نهایتی برای w وجود دارد، ولی ما به دنبال راه‌حل مینیمم‌نرم هستیم، یعنی می‌خواهیم $\min \|w\|^2$ به شرط $Xw = y$ را حل کنیم. راه‌حل مینیمم نرم چیست؟

(ه) از $X = U\Sigma V^\top$ SVD استفاده کنید و ساده‌سازی کنید.

(و) توجه خواهید کرد که راه‌حل مینیمم نرم به فرم $w^* = By$ است. چه اتفاقی می‌افتد اگر X را از سمت راست در ماتریس B ضرب کنیم؟ به همین دلیل ماتریس B راه‌حل مینیمم‌نرم را معکوس راست نام‌گذاری می‌کنند.

۴. (۶۰ نمره)

یک مسئله رگرسیون خطی را در نظر بگیرید که شامل n نقاط آموزشی و d ویژگی است. زمانی که $n = d$، ماتریس ویژگی $F \in \mathbb{R}^{n\times n}$ دارای بیشینه مقدار تکین α و کوچک‌ترین مقدار تکین بسیار کوچک است. ما مشاهدات نویزی $\mathbf{y} = F\mathbf{w}^* + \epsilon$ را داریم. اگر $\hat{\mathbf{w}}_{inv} = F^{-1}\mathbf{y}$ را محاسبه کنیم، به دلیل مقدار تکین کوچک F و وجود نویز مشاهده می‌کنیم که $\|\hat{\mathbf{w}}_{inv} - \mathbf{w}^*\|_2 = 10^{10}$.

به جای وارون‌سازی ماتریس، فرض کنید که از gradient descent استفاده می‌کنیم. ما k تکرار از gradient descent را اجرا می‌کنیم تا مقدار تابع زیان $\ell(w) = \frac{1}{2}\|\mathbf{y} - F\mathbf{w}\|_2^2$ را با شروع از $\mathbf{w}_0 = 0$ کمینه کنیم. ما از نرخ یادگیری η که به اندازه کافی کوچک است استفاده می‌کنیم تا gradient descent برای مسئله داده شده همگرا شود (این نکته مهم است.).

فرمول به‌روزرسانی gradient descent برای $t > 0$ به صورت زیر است:

$$\mathbf{w_t} = \mathbf{w_{t-1}} - \eta\left(F^T(F\mathbf{w_{t-1}} - \mathbf{y})\right)$$

ما به دنبال خطای $\|\mathbf{w_k} - \mathbf{w}^*\|_2^2$ هستیم. می‌خواهیم نشان دهیم که در بدترین حالت، این خطا می‌تواند به صورت خطی با تکرارهای k رشد کند و به طور خاص: $\|\mathbf{w_k} - \mathbf{w}^*\|_2 \leq k\eta\alpha\|\mathbf{y}\|_2 + \|\mathbf{w}^*\|_2$

به عبارت دیگر، خطا نمی‌تواند "از کنترل خارج شود"، حداقل نه خیلی سریع.

برای این تکلیف، تنها لازم است ایده کلیدی را اثبات کنید، زیرا ادامه آن با استفاده از استقرا و نامساوی مثلثی نتیجه‌گیری می‌شود.

نشان دهید که برای $t > 0$:

$$\|\mathbf{w_t}\|_2 \leq \|\mathbf{w_{t-1}}\|_2 + \eta\alpha\|\mathbf{y}\|_2.$$

اگر gradient descent نتواند واگرا شود در مورد $(I - \eta F^T F)$ چه میتوان گفت؟ مقادیر ویژه آن چه شکلی دارند؟

۵. (۶۰ نمره)

(الف) نشان دهید که می‌توان خطای مربعی مورد انتظار را به سه بخش بایاس، واریانس و خطای غیر قابل کاهش σ^2 تجزیه کرد:

$$Error = Bias^2 + Variance + \sigma^2$$

به طور رسمی، فرض کنید یک مجموعه آموزشی $\mathcal{D}$ به صورت تصادفی نمونه‌برداری شده داریم (که به طور مستقل از داده‌های آزمون ما گرفته شده‌اند)، و یک برآوردگر $\theta = \hat{\theta}(\mathcal{D})$ انتخاب می‌کنیم (برای مثال، با استفاده از کمینه‌سازی ریسک تجربی). خطای مربعی مورد انتظار برای یک ورودی آزمون x به صورت زیر تجزیه می‌شود:

$$\mathbb{E}_{Y\sim p(y|x),\mathcal{D}}\left[(Y - \hat{f}_{\hat{\theta}(\mathcal{D})}(x))^2\right] = Bias\left(\hat{f}_{\hat{\theta}(\mathcal{D})}(x)\right)^2 + Var\left(\hat{f}_{\hat{\theta}(\mathcal{D})}(x)\right) + \sigma^2$$

تعاریف فرمولی واریانس و بایاس را که در زیر آمده است ممکن است مغید باشد به یاد آورید:

$$Bias(\hat{f}_{\hat{\theta}(\mathcal{D})}(x)) = \mathbb{E}_{Y\sim p(Y|x),\mathcal{D}}\left[\hat{f}_{\hat{\theta}(\mathcal{D})}(x) - Y\right]$$

$$Var(\hat{f}_{\hat{\theta}(\mathcal{D})}(x)) = \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\hat{\theta}(\mathcal{D})}(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\hat{\theta}(\mathcal{D})}(x)]\right)^2\right]$$

(ب) فرض کنید مجموعه آموزشی ما شامل $\mathcal{D} = \{(x_i, y_i)\}_{i=۱}^{n}$ است، جایی که تنها تصادفی بودن از بردار برچسب‌ها $Y = X\theta^* + \epsilon$ ناشی می‌شود که θ^* مدل خطی واقعی است و هر متغیر نویز ϵ_i به طور مستقل و یکسان توزیع شده است با میانگین صفر و واریانس ۱. ما از حداقل مربعات معمولی برای برآورد $\hat{\theta}$ از این داده‌ها استفاده می‌کنیم.

خطا و کوواریانس تخمین $\hat{\theta}$ را محاسبه کنید و از آن برای محاسبه خطا و واریانس پیش‌بینی در ورودی‌های تست خاص x استفاده کنید. به خاطر بیاورید که راه‌حل OLS به صورت زیر داده شده است:

$$\hat{\theta} = (X^\top X)^{-۱} X^\top Y,$$

که $X \in \mathbb{R}^{n\times d}$ ماتریس داده‌های غیرتصادفی ما است و $Y \in \mathbb{R}^n$ بردار (تصادفی) اهداف آموزشی است. برای سادگی، فرض کنید که $X^\top X$ قطری است.

سوالات عملی (۳۵۰ نمره)

۱. (۱۵۰ نمره) فایل نوتبوکی که در اختیارتان قرار داده شده است راکامل کنید. در این تمرین الگوریتم CART را یه صورت کامل پیاده سازی کرده و سپس از آن برای طبقه بندی MNIS اسفاده می‌کنیم.

۲. (۱۰۰ نمره) فایل نوتبوکی که در اختیارتان قرار داده شده است راکامل کنید. در این تمرین به کمک SVM تشخیص می‌دهیم کدام بیماران در معرض بیماری هستند. توضیحات تکمیلی در فایل نوتبوک موجود است.

۳. (۱۰۰ نمره) فایل نوتبوکی که در اختیارتان قرار داده شده است راکامل کنید. در این تمرین یک kNN classifier را پیاده سازی خواهید کرد.