به نام خدا

یادگیری عمیق

پاییز ۱۴۰۱

استاد: دکتر فاطمی‌زاده

گردآورندگان: -

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی برق

تمرین اول | **مفاهیم پایه** | مهلت ارسال: سه‌شنبه ۱۸آبان (با احتساب تاخیر)

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۶ روز و در مجموع ۲۰ 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 Quera قرار دهید. برای بخش عملی تمرین نیز لینگ گیت‌هاب که تمرین و نتایج را در آن آپلود کرده‌اید قرار بدهید. دقت کنید هر سه فایل نوتبوک تکمیل شده بخش عملی را در گیت‌هاب قرار دهید .
- لطفا تمامی سوالات خود را از طریق کوئرای درس مطرح بکنید (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. (۲۵ نمره)

(آ) با توجه به تعریف نرم ماتریسی رابطه زیر را برای $A \in R^{m*n}$ ثابت کنید:

$$||A||_۲ \leq ||A||_F \leq \sqrt{rank(A)}||A||_۲$$

راهنمایی: از تجزیه SVD ماتریس A استفاده کنید.

(ب) i. با در نظر گرفتن متغییر تصادفی نامنفی X ،نامساوي زیر را اثبات کنید (نامساوي ماركوف)

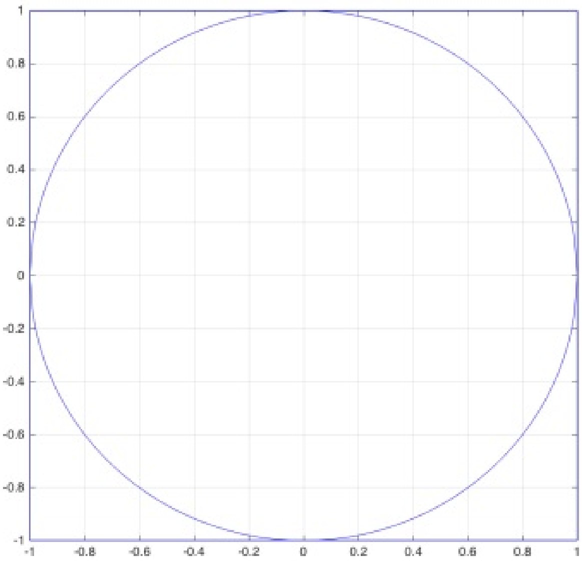
$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}$$

ii. با در نظر گرفتن نتیجه بخش الف، نشان دهید براي متغیر تصادفی دلخواه Z با امید ریاضی μ و واریانس $\sigma^۲$ نامساوی زیر برقرار است(نامساوی چبیشف):

$$P(|Z-\mu| \geq \varepsilon) \leq \frac{\sigma^۲}{\varepsilon^۲}$$

iii. می خواهیم مقدار عدد π را تخمین بزنیم. براي این کار روي صفحه مختصات دوبعدي، دایره اي به شعاع واحد و مربع محیطی آن را شبیه شکل زیر رسم می کنیم. مساحت این دایره π و مساحت مربع محیطی آن ۴ است. براي تخمین مقدار π تعدادي نقاط تصادفی داخل این مربع تولید کرده

و نسبت تعداد نقاطی که داخل دایره قرار می گیرند را به تعداد کل به عنوان مقدار عدد π در نظر می گیریم.



iv. با استفاده از نامساوي چبیشف تعداد عددهاي تصادفی اي که باید تولید کنیم تا با قطعیت ۹۵ درصد بدانیم که خطاي تخمین از ۱ درصد کمتر است را مشخص کنید.

۲. (۲۵ نمره)

- با فرض اینکه $x \in \mathbb{R}$ و A ماتریسی مربعی به ابعاد $n \times n$ باشد، دو مورد زیر را اثبات کنید.

$$\frac{\partial a^T x}{\partial x} = a^T$$
$$\frac{\partial x^T A x}{\partial x} = x^T(A + A^T)$$

- اگر درایه های ماتریس A تابع پارامتر اسکالر β باشند، رابطه‌ی زیر را اثبات کنید.

$$\frac{\partial A^{-1}}{\partial \beta} = -A^{-1} \frac{\partial A}{\partial \beta} A^{-1}$$

- نشان دهید

$$\nabla_A |A| = |A| A^{-T}$$

و به دنبال آن

$$\nabla_A log|A| = A^{-1}$$

۳. (۱۰ نمره) با فرض اینکه $\lambda_1, \lambda_2, \cdots, \lambda_n$ مقادیر ویژه ماتریس A باشند، موارد خواسته شده را اثبات کنید.

- $\lambda_1 + \lambda_2 + \cdots + \lambda_n = trace(A)$
- $\lambda_1 \lambda_2 \cdots \lambda_n = det(A)$

۴. (۲۰ نمره) فرض کنید ماتریس A تجزیه SVD به شکل $A = U\Sigma V^T$ داشته باشد. برای این ماتریس عملگر Pseudo-Inverse به شکل $A^\dagger = V\Sigma^{-1}U^T$ تعریف می‌شود. حال اثبات کنید:

- اگر ماتریس A رنک کامل ستونی داشته باشد، $A^\dagger = (A^T A)^{-1} A^T$
- اگر ماتریس A رنک کامل سطری داشته باشد، $A^\dagger = A^T (A A^T)^{-1}$

۵. (۴۰ نمره) فرض کنید ماتریس M را بصورت $M = \begin{bmatrix} A_{n*n} & B_{n*k} \\ C_{k*n} & D_{k*k} \end{bmatrix}$ تعریف کنیم، برای این ماتریس اثبات کنید :

- اگر A وارون پذیر باشد آنگاه:

$$\det \begin{bmatrix} A_{n*n} & B_{n*k} \\ C_{k*n} & D_{k*k} \end{bmatrix} = \det(A)\det(D - CA^{-۱}B)$$

- اگر D وارون پذیر باشد آنگاه:

$$\det \begin{bmatrix} A_{n*n} & B_{n*k} \\ C_{k*n} & D_{k*k} \end{bmatrix} = \det(D)\det(A - BD^{-۱}C)$$

- با کمک دو بخش قبل اثبات کنید :

$$(A + BDC)^{-۱} = A^{-۱} - A^{-۱}B(D + CA^{-۱}B)^{-۱}CA^{-۱}$$

- اگر A یک ماتریس وارون‌پذیر و u, v بردارهای ستونی باشند، با استفاده از تکنیک‌های استفاده شده در بخش‌های قبل ثابت کنید :

$$\det(A - uv^T) = \det(A)(۱ - v^T A^{-۱}u)$$

(نکته: مشخصا هر کجا از ماتریسی وارون گرفته شده است آن ماتریس را وارون‌پذیر در نظر بگیرید.)

۶. (۷۰ نمره) در این سوال میخواهیم یک ویژگی جالب درباره ارتباط مقادیر ویژه دو ماتریس مرتبط با هم را کشف کنیم که شاید بعدها در رابطه با کاربرد آن بیشتر با هم صحبت کنیم، اما قبل اثبات حکم کلی چند لم و ابزار جرئی را با هم اثبات می‌کنیم .

- فرض کنید x و y دو بردار ستونی با درایه‌های در حالت کلی مختلط، با بعد n باشند، یعنی $x, y \in \mathbb{C}^n$، $a \in \mathbb{C}$ یک اسکالر مختلط باشد، $B \in M_n$ ماتریس اولیه ما باشد و به کمک بردارها و اسکالر تعریف شده ماتریس جدیدی بصورت $A = \begin{bmatrix} B & x \\ y^* & a \end{bmatrix} \in M_{n+۱}$ تولید بکنیم، حال اثبات کنید که چندجمله‌ای مشخصه ماتریس جدید تولید شده بصورت زیر است :

$$p_A(t) = \det(tI - A) = (t - a)p_B(t) - y^*(adj(tI - B))x$$

- فرض کنید ماتریس‌های $A, B \in M_n$ ماتریس های هرمیتی باشند و مقادیر ویژه این ماتریس‌ها به ترتیب از کوچک به بزرگ مرتب شده باشند یعنی رابطه زیروند مقادیر ویژه و بزرگی آنها بصورت : $\lambda_۱ \leq \lambda_۲ \leq \cdots \leq \lambda_n$ باشد، حال برای این دو ماتریس و ماتریس حاصل جمع آنها رابطه زیر را اثبات بکنید :

$$\lambda_i(A + B) \leq \lambda_{i+j}(A) + \lambda_{n-j}(B) \quad j = ۰, ۱, \cdots, n - i$$

برای هر $i = ۱, \cdots, n$

- حال میخواهیم حکم کلی را اثبات کنیم، فرض کنید متغیرها و ماتریس هایی به شرح روبرو داشته باشیم: $B \in M_n, y \in \mathbb{C}^n, a \in \mathbb{R}, A = \begin{bmatrix} B & y \\ y^* & a \end{bmatrix}$ که ماتریس B یک ماتریس هرمیتی است، حال به کمک بخش‌های قبل اثبات بکنید:

$$\lambda_۱(A) \leq \lambda_۱(B) \leq \lambda_۲(A) \leq \cdots \leq \lambda_n(A) \leq \lambda_n(B) \leq \lambda_{n+۱}(A)$$

😇 اگر خیلی در رابطه با کاربرد این نکته کنجکاو هستید، از این قضیه استفاده میشود تا اثبات کنیم افزایش پیچیدگی یک شبکه عصبی یا مدل یادگیری عمیق در صورت وجود رگولاریزر مناسب در تابع هزینه شبکه، میتواند باعث از بین رفتن پدیده Double-Descent بشود. (با مفاهیم گفته شده در طول درس آشنا خواهید شد!)

۷. (۱۰ نمره) فرض کنید از توزیع با چگالی زیر، نمونه‌های $X_1, X_2, \cdots, X_n$ را دریافت کرده‌ایم، تخمینگر بیشینه likelihood را برای θ بدست بیاورید.

$$f(x) = \frac{1}{\theta^2} x e^{-\frac{x}{\theta}}, \quad 0 < \theta < \infty$$

۸. (۲۰ نمره) توزیع نرمالی با واریانس معلوم σ^2 و میانگین مجهول μ داریم. با فرض اینکه n داده‌ی $X_1, X_2, \cdots, X_n$ از این توزیع بدست آورده‌ایم، ابتدا تخمینگر MLE و MAP را برای μ بدست بیاورید، در نهایت بررسی کنید در صورت میل دادن تعداد سمپل‌ها به بی نهایت چه اتفاقی برای این تخمینگرها می‌افتد. (برای تخمینگر MAP تصور کنید که توزیع پیشینه خود یک توزیع نرمال با میانگین γ و واریانس β^2 است.)

۹. (۲۰ نمره) فرض کنید یک توزیع گوسی چند متغیره بصورت $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ داریم که در آن متغیر تصادفی را میتوان بصورت زیر نوشت :

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$$

و به طبع آن خواهیم داشت :

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

و برای ماتریس کوواریانس :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$$

- حال اثبات کنید توزیع شرطی $p(\mathbf{x}_a|\mathbf{x}_b)$ یک توزیع گوسی است که مشخصه‌هایی به شرح زیر دارد :

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}{\boldsymbol{\Sigma}_{bb}}^{-1}(\mathbf{x}_\mathrm{b} - \boldsymbol{\mu}_b)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}{\boldsymbol{\Sigma}_{bb}}^{-1}\boldsymbol{\Sigma}_{ba}$$

- حال اثبات کنید توزیع حاشیه‌ای $p(\mathbf{x}_\mathrm{a})$ توزیع گوسی‌ای است که مشخصه‌هایی به شرح زیر دارد:

$$E[\mathbf{x}_a] = \boldsymbol{\mu}_a$$

$$Cov[\mathbf{x}_\mathrm{a}] = \boldsymbol{\Sigma}_{aa}$$

۱۰. (۶۰ نمره) در این سوال میخواهیم با الگوریتم بهینه‌سازی گرادیان کاهشی و ارتباط آن با عملگر $Pseudo-inverse$ آشنا شویم. فرض کنید مسئله بهینه سازی بصورت $\min_x \quad ||Ax-b||^2$ داریم،

- اثبات کنید اگر این مسئله چند جواب داشته باشد، جواب به فرم $x^* = A^\dagger b$ کمترین نرم۲ را در میان آنها دارد.
- فرض کنید میخواهیم این مسئله را با الگوریتم گرادیان کاهشی حل کنیم. بدین صورت که در مرحله اول مقدار متغیر بهینه سازی را با صفر مقداردهی میکنیم ($x^{(0)} = 0$) و در هر مرحله متغیر را بصورت $x^{(t+1)} = x^{(t)} - \nu A^T(Ax^{(t)} - b)$ آپدیت میکنیم که در آن ν نرخ یادگیری است. اثبات کنید در این صورت جواب مسئله در صورت انتخاب پارامتر یادگیری بصورت مناسب پس از تعداد ایتریشن کافی به جواب $x^* = A^\dagger b$ میل می‌کند.
- اثبات کنید یک کران مناسب برای نرخ یادگیری میتواند $\frac{2}{\sigma^2_{max}(A)}$ باشد، بدین معنا که اگر نرخ یادگیری بصورت $\nu \leq \frac{2}{\sigma^2_{max}(A)}$ انتخاب شود، در هر مرحله از پیاده سازی الگوریتم بهینه‌سازی، تابع هزینه نسبت به مرحله قبل کاهش خواهد داشت: $||Ax^{(t+1)} - b|| \leq ||Ax^{(t)} - b||$
(تمامی نرم‌ها ، نرم۲ هستند.)

سوالات عملی (۳۰۰ نمره)

۱. (۱۰۰ نمره) فایل نوتبوکی در اختیار شما قرار داده شده است. راهنمایی های لازم برای نحوه انجام تمرین در فایل نوتبوک انجام شده است. در این تمرین با کار با کتابخانه pandas ، نحوه تبدیل داده خام به داده مناسب برای ورودی مدل یادگیری و در نهایت پیاده سازی یک مدل یادگیری ماشین مبتنی بر یک روش بهینه‌سازی متناوب آشنا میشوید. برای بخش ترین کردن مدل به مقاله Paper مراجعه شده و تا بخش stochastic gradient descent خوانده شود.

۲. (۱۰۰ نمره) فایل نوتبوکی که در اختیارتان قرار داده شده است راکامل کنید. در این تمرین به کمک SVM به دسته بندی داده‌هایی که در فایل Heart Disease Dataset.csv قرار دارند میپردازیم. توصیفی از این داده ها در فایل Dataset Description.pdf آمده است. برای دریافت نمره کامل بخش MySVM باید دقت بهترین هسته پیاده‌سازی شده روی داده‌های ازمون حداقل ۹۰ درصد باشد. برای دقت های بالاتر محسوس، در صورت توضیح مناسب دلیل،نمره امتیازی در نظر گرفته خواهد شد.

۳. (۱۰۰ نمره) فایل نوتبوکی که در اختیارتان قرار داده شده است راکامل کنید. در این تمرین با رگرسیون خطی، رگرسیون ناپارامتری و نزدیک‌ترین همسایه و پیاده‌سازی آنها آشنا میشوید.