بهینه سازی در علوم داده‌ها

نیم‌سال دوم ۱۴۰۰-۱۴۰۱

مدرس: دکتر مجتبی تفاق

دانشکده علوم ریاضی

آزمون میان‌ترم

تاریخ بارگذاری: ۶ خرداد ۱۴۰۱ ساعت ۹

موعد تحویل: ۶ خرداد ساعت ۲۱

- مهلت تحویل پاسخ‌ها تا ساعت ۲۱ روز ۶ خرداد ۱۴۰۱ می‌باشد.
- پاسخ‌های خود را در سامانه‌ی cw و در قسمت مربوط به میان‌ترم بارگذاری کنید.
- اشکالات خود درباره‌ی سوالات را می‌توانید با تیم درس مطرح کنید.
- سوال‌هایی که نیاز به برنامه‌نویسی داشته باشند، باید حتما با زبان Julia نوشته شوند. همچنین باید کد خود (به صورت فایل jl.) و نتایج اجرای آن را همراه با پاسخ ارسال کنید. در صورتی که کد ارسالی شما قابل اجرا نباشد، ممکن است بخشی از نمره‌ی تمرین را از دست بدهید.
- استفاده از اینترنت و همفکری با دیگران در طول امتحان مجاز نیست اما می‌توانید از کتاب و جزوه کمک بگیرید.

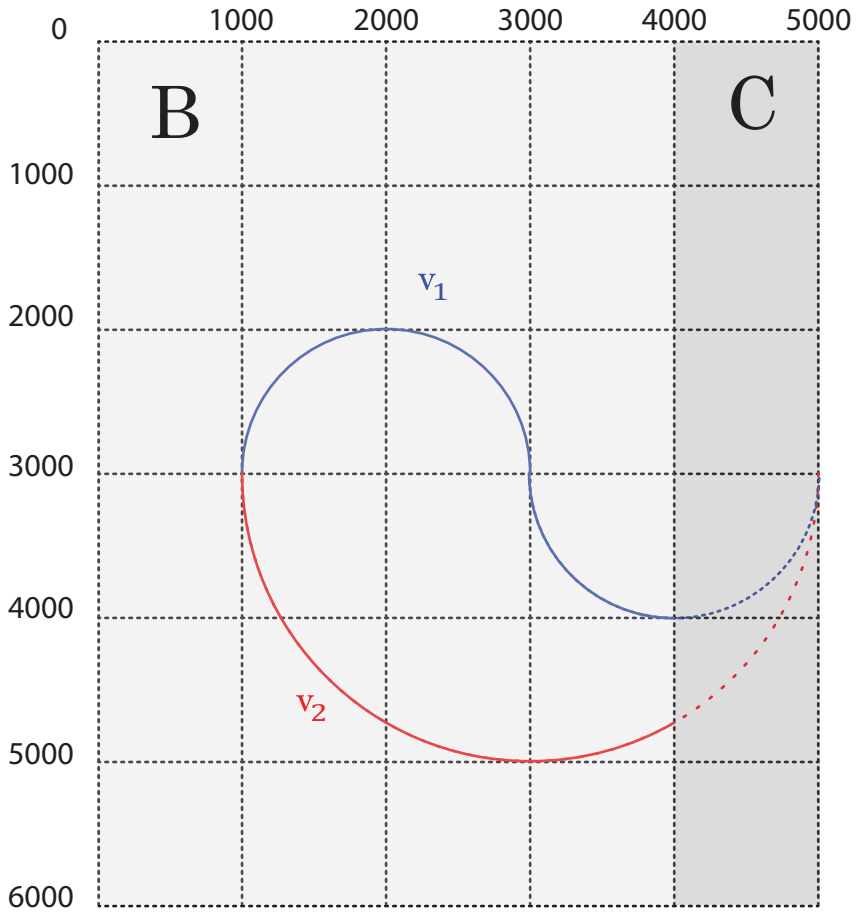
پرسش ۱ ماتریس دسته‌بندی کننده

ماتریس $A_{n\times n}$ و بردار یکه‌ی $x \in \mathbb{R}^n$ داده شده‌اند. ماتریس A بردارهای فضای $\mathbb{R}^n$ را نسبت به x به دو دسته تقسیم می‌کند؛ بدین صورت که برای هر بردار $z \in \mathbb{R}^n$ اگر عدد حقیقی t وجود داشته باشد که زاویه‌ی بین z و tx کوچکتر و یا مساوی با $\alpha > \frac{1}{2}$ باشد، آنگاه $z^T Az \geq 0$ و بردار z متعلق به کلاس ۱+ است و در غیر این صورت z متعلق به کلاس ۱ـ خواهد بود یعنی $z^T Az < 0$ (در واقع برای هر x و α یک A با خاصیت فوق وجود دارد).

آ) از آنجاییکه برای nهای بزرگ محاسبه‌ی $z^T Az$ برای تعداد زیادی بردار z زمان‌بر است، برای یک ماتریس A و بردار x داده شده، قصد داریم از یک classifier برای دسته‌بندی بردارها استفاده کنیم که برای محاسبه‌ی کلاس هر بردار z هزینه‌ی زمانی کمتری نسبت به زمان محاسبه $z^T Az$ بپردازد. روش پیشنهادی خود را شرح دهید و برای ماتریس A و بردار x داده شده در فایل inputs.txt روش خود را در زبان برنامه نویسی julia پیاده سازی کنید.

ب) ماتریس A دارای این ویژگی است که برای بردارهای z که زاویه‌ی بین tx و z برابر α است، $z^T Az = 0$ خواهد بود. برای بردار x داده شده در فایل inputs.txt و $\alpha = 1.57$ ماتریس A را طوری بیابید که بردارهای فضا را به دو دسته تقسیم کند؛ دسته‌ی اول بردارهایی که زاویه‌ی آن‌ها با tx با دقت دو رقم اعشار برابر با α باشد و دسته‌ی دوم سایر بردارها.

پرسش ۲ بازسازی تصویر آسیب دیده



دوست ایمان، عرفان، فایل یک تصویر سیاه و سفید قدیمی را که به صورت یک ماتریس ۵۰۰۰×۶۰۰۰ پیکسل ذخیره شده، برای او ارسال کرده است، اما به دلیل اختلال در شبکه اینترنت تصویر به صورت ناقص به دست ایمان رسیده است. برای بازسازی کامل تصویر ایده‌ی جالبی به ذهن ایمان رسیده است! او قصد دارد با پیدا کردن خم‌های درون تصویر و تقریب زدن آن‌ها با توابع ریاضی، و سپس با امتداد دادن آن توابع قسمت ناقص تصویر را بازیابی کند. فرض کنید تصویر به صورت یک ماتریس $A = [B|C]$ باشد که سایز ماتریس‌های B و C به ترتیب ۶۰۰۰×۴۰۰۰ و ۶۰۰۰ ×۱۰۰۰ است که B به طور کامل به دست ایمان رسیده است اما هیچ اطلاعاتی از C بجز ابعاد آن در دست نیست. ایمان موفق شده دو خم v_1 و v_2 را در C پیدا کرده و در فایل‌های v1.txt و v2.txt ذخیره کند، به این صورت که برای tـاُمین عدد فایل داریم $B[v_1[t],\ 1000+t] = 1$. او همچنین متوجه شده که از قضا این خم‌ها تابع هستند؛ یعنی، به عنوان مثال برای خم v_1 در هر ستون $1000 \le j \le 4000$ از ماتریس B دقیقا یک سطر $1 \le i \le 6000$ وجود دارد که $B[i,j] = 1$. بنابراین ایمان خم v_1 (v_2) را به صورت یک بردار به طول ۳۰۰۰ ذخیره کرده که در آن $v_1[k]$ به این معنی است که $B[v_1[k],\ 1000+k] = 1$ است $(1 \le k \le 3000)$.

برای پیاده سازی ایده‌ی ایمان

آ) روش مناسبی پیشنهاد دهید؛

ب) با استفاده از روش ارائه شده در قسمت الف خم‌های v_1 و v_2 را به درون C امتداد داده و تصویر C را کامل کنید. برای این منظور فرض کنید f_1 تابعی است که v_1 را تقریب می‌زند. کافیست به ازای $4000 \leq k \leq 5000$ مقدار جزء صحیح $f_1(k)$ را به عنوان تقریبی از $v_1(k)$ در نظر بگیرید. در انتها بردارهای u_1 و u_2 به طول ۱۰۰۰ را در فایل خروجی ذخیره کنید که $u_i[k] = \lfloor f_i(k) \rfloor$ $(i = 1, 2)$. در انتها برای $4000 \leq k \leq 5000$، ماتریس C را به صورت $C[u_i[k],\ k] = 1$ تقریب بزنید و سپس نمودار خم‌های v_1 و v_2 را به طور کامل رسم و به صورت فایل png. ارسال کنید.

پ) برای محاسبه‌ی خطای این روش چه راهی پیشنهاد می‌دهید؟

ت) اگر v_2 نیم دایره‌ای به مرکز (۳۰۰۰،۳۰۰۰) و شعاع ۲۰۰۰ باشد، خطای روش توصیف شده در قسمت ب را پیدا کنید.

پرسش ۳ جریان بیشینه

فرض کنید G یک گراف جهت‌دار n رأسه و m یاله است و s و t دو رأس متمایز از G هستند. طوری که s یال ورودی ندارد و t یال خروجی. یک جریان از G را مقداردهی به یال‌های G با اعداد حقیقی از ۰ تا ۱ می‌نامیم طوری که برای هر رأس $v \in V(G)\backslash\{st\}$، مجموع مقداردهی یال‌های ورودی به v برابر با مجموع مقدار دهی یال‌های خروجی از v باشد. مقدار یک جریان f را مجموع مقداردهی یال‌های خروجی از s در نظر می‌گیریم.
در مسئله‌ی جریان بیشینه، به دنبال یافتن جریانی از G هستیم که دارای مقدار بیشینه باشد.
در این پرسش، می‌خواهیم با استفاده از روش‌های کمینه سازی با کمک رگرسیون که در درس خواندیم، به این مسئله حمله کنیم. اگر رأس‌ها با شماره‌ی ۱ تا n و یال‌ها را با شماره‌ی ۱ تا m شماره‌گذاری کنیم، یک جریان را می‌توان با برداری از $\mathbb{R}^m$ چون $f \in \mathbb{R}^m$ نمایش داد که برای یال $e \in \{1, ..., m\}$، f_e برابر با مقداردهی یال e باشد.

آ) قیدهای مسئله‌ی جریان بیشینه را به صورت قیدهایی خطی، به شکل برابری یا نابرابری بر روی بردار f بنویسید. سپس یک تابع هدف به صورت $||Af||^2$ برای ماتریس A ارائه کنید که مسئله‌ی جریان بیشینه، معادل با بیشینه‌سازی این تابع هدف بر روی قیود f باشد. ثابت کنید مسئله‌ی خطیِ ساخته شده، معادل با جریان بیشینه است.

ب) مسئله‌ی بخش قبل، یک مسئله‌ی بیشینه‌سازی است. در حالی که در مسئله‌ی رگرسیون، تلاش برای کمینه‌سازی تابع هدف داریم، یکی از راه‌های ساده برای حل این مشکل در مسائل، کم کردن اختلاف هدف با عددی بزرگ است. یک کران بالا برای پاسخ مسئله ارائه کنید و مسئله را به یک مسئله‌ی کمینه‌سازی تبدیل کنید (استفاده از مسئله‌ی دوگان مجاز نیست، ولی فکر کردن به آن بعد از امتحان توصیه می‌شود).

پ) در مسئله‌ی رگرسیون خطی، در صورتی که تابع هدف $||Af - b||^2$ باشد. استقلال خطی روی ستون‌های ماتریس A را فرض می‌کردیم. آیا این مسئله این شرط را دارد؟ ثابت کنید.

ت) یک راه حل برای حل کامل مسئله، استفاده رگرسیون قیددار غیر خطی است. در این صورت دو تابع $h(x)$ و $g(x)$ داریم که هدف کمینه سازی $h(x)$ با شرط $g(x) = 0$ است. طوری که g, h هر دو مشتق پذیر باشند. مسئله‌ی ساخته شده را به این فرمت در آورید.
دقت کنید احتمالا در بخش اول، تعدادی قید نابرابری داریم، این قیود را چگونه در $g(x)$ یا $h(x)$ قرار دهیم؟
نکته: برای حل عملی مسئله، می‌توانید یک relaxation بر حسب عدد ϵ در نظر بگیرید. سپس خطا بر روی قیدها یا تا حد ϵ مجاز بگیریم و خانواده از توابع $h_\epsilon(x)$ و $g_\epsilon(x)$ ارائه کنیم.

ث) آیا در مسئله‌ای که ساختید، نقاط بهینه‌ی موضعی، نقاط بهینه‌ی سراسری هستند؟ چرا؟ اگر نه، چگونه می‌توان این مشکل را حل یا کم‌رنگ کرد؟ این کار را انجام دهید.

ج) با توجه مسئله‌ای که ساخته‌اید. الگوریتمی برای حل مسئله‌ی جریان بیشینه ارائه کرده و آن را توضیح دهید. سپس با زبان جولیا، آن را پیاده‌سازی کنید. طوری که در خط اول ورودی دو عدد n و m را با فاصله از هم بخواند، سپس در m خط بعد، در هر خط ۲ عدد که نشانگر یالی از رأس با عدد اول به دوم است را بخواند. سپس در خروجی ابتدا اندازه‌ی جریان بیشینه را چاپ کند و در خط بعد، بردار f را چاپ کند، طوری که درایه‌های بردار اعداد اعشاری باشند و با فاصله از هم جدا شده‌اند.

پرسش ۴ SVM

تعدادی نقطه در صفحه به شما داده شده است. هدف پیدا کردن خطی است که این نقاط را به دو دسته افراز کند و فاصله‌ی مناسبی با نقاط هر دسته داشته باشد.

آ) ابتدا می‌خواهیم نقاط را به دو دسته افراز کنیم. مختصات نقاط را از فایل points.txt بخوانید. مولفه‌ی اول برابر با x نقاط و مولفه‌ی دوم برابر با y نقاط است.
با استفاده از الگوریتم‌های تدریس شده در کلاس و یا روش‌های ابداعی نقاط را به دو دسته افراز کنید. نقاط باید به نحوی دسته بندی شوند که نقاطی که به یکدیگر نزدیک‌تر هستند در یک دسته قرار بگیرند.

پس از دسته بندی، نقاط یک دسته را آبی و نقاط دسته‌ی دیگر را قرمز کنید و نقاط رنگی را در صفحه رسم کنید و خروجی را با نام res1.png ذخیره کنید.

در گزارشتان، روشی که برای دسته‌بندی نقاط استفاده می‌کنید را ذکر کنید و آن را به صورت خلاصه توضیح دهید. در صورتی که این روش در کلاس تدریس نشده است، باید به صورت کامل و دقیق آن را توضیح دهید. روش شما باید به صورت هوشمند باشد و در نمونه‌های مشابه نیز نتیجه‌ی قابل قبولی بدهد. در نتیجه به روش‌هایی که به صورت دستی و یا هاردکد شده نقاط را دو دسته کنند، نمره‌ای تعلق نمی‌گیرد.

توجه کنید که حتما نتیجه‌ی پلات را باید به صورت عکس و با نام گفته شده ذخیره کنید. در غیر اینصورت حتی اگر در هنگام اجرا آن را پلات کنید، نمره‌ای به آن تعلق نمی‌گیرد.

ب) در این بخش می‌خواهیم خطی پیدا کنیم که نقاط آبی در یک طرف آن و نقاط قرمز در طرف دیگر آن باشند و همچنین فاصله‌ی عمود هر نقطه در صفحه تا این خط دست کم ۱ واحد باشد.

ابتدا در گزارشتان به صورت دقیق و کامل، فرمول بندی و مدل سازی‌ای که انجام داده‌اید را ذکر کنید و نحوه‌ی حل آن را توضیح دهید و سپس معادله‌ی خطی که بدست می‌آورید را بنویسید. استفاده از روش‌هایی که به صورت هاردکد شده معادله‌ی خط را بدهند مورد قبول نمی‌باشد اما در صورت نیاز می‌توانید برخی نقاط کمکی را اضافه کنید و از آن‌ها بهره ببرید. خطی که در شرایط بالا صدق می‌کند و جواب نهایی است را l بنامید. خط‌های l_1 و l_2 را خطوطی در نظر بگیرید که موازی با l باشند و فاصله‌شان تا l برابر با ۱ واحد باشد.

تمامی نقاط آبی و قرمز را به همراه خط l و دو خط l_1 و l_2 رسم کنید و خروجی را با نام res2.png ذخیره کنید. تمامی کدهایی که برای هر دو بخش سوال می‌نویسید را باید همراه با گزارش آپلود کنید.

پرسش ۵ مجموعه‌های آفین

آ) فرض کنید $\mathcal{A} \subset \mathbb{R}^m$ یک مجموعه افین (یک زیرفضای انتقال‌یافته) باشد. به طور دقیق‌تر میتوان گفت که مجموعه‌ای از فرم‌ها به صورت زیر است که در آن $A \in \mathbb{R}^{n \times m}$ و $a \in \mathbb{R}^n$:

$$\mathcal{A} = \{x | Ax = a\}$$

می دانیم که به این روش نوشتن یک مجموعه افین، فرم ضمنی آن مجموعه می‌گویند. اما یک مجموعه افین را همیشه می‌توان به شکل دیگری به نام شکل صریح نوشت.
به عبارت دیگر می‌توان گفت همیشه یک عدد صحیح p، یک ماتریس $B \in \mathbb{R}^{m \times p}$ و یک بردار $b \in \mathbb{R}^m$ وجود دارد که :

$$\mathcal{A} = \{By + b | y \in \mathbb{R}^p\}$$

حال فرض کنید به شما A و a داده شده باشد، روشی برای یافتن pوBو b ارائه دهید. چگونه می‌توان کوچکترین p قابل استفاده را تعیین کرد؟

ب) ما همواره می توانیم فرم صریح را به فرم ضمنی تبدیل کنیم .الگوریتمی ارائه دهید که با آن بتوان با داشتن B و b به Aو a دست یافت.

پ) روش خود را برای روی داده‌های زیر اجرا کنید.

$$B = \begin{bmatrix} 37 & -100 & -15 \\ -16 & 230 & 20 \\ -66 & 100 & -10 \\ 96 & -34 & 52 \\ 11 & 24 & -20 \end{bmatrix}$$

$$b = \begin{bmatrix} 2 \\ 15 \\ 2 \\ -13 \\ 2 \end{bmatrix}$$

ت) فرض کنید $\mathcal{A}$ و$\mathcal{B}$ دو مجموعه ی آفین در$\mathbb{R}^n$ باشند. فاصله‌ی بین این دو مجموعه را به صورت زیر تعریف می‌کنیم:

$$dist(\mathcal{A}, \mathcal{B}) = min_{x \in \mathcal{A}, y \in \mathcal{B}} |x - y|$$

ما فرم ضمنی دو مجموعه ی آفین $\mathcal{A}$ و$\mathcal{B}$ را به صورت زیر داریم :

$$\mathcal{A} = \{x | Ax = a\}$$

$$\mathcal{B} = \{y | By = b\}$$

ماتریس های $A \in \mathbb{R}^{m \times n}$ و $B \in \mathbb{R}^{m \times n}$ و بردارهای $a \in \mathbb{R}^m$ و $b \in \mathbb{R}^m$ داده شده است و از شما می‌خواهیم فاصله$\mathcal{A}$ و$\mathcal{B}$ را به روش حداقل مربعات محاسبه کنید بطوریکه :

$$dist(\mathcal{A}, \mathcal{B}) = min_z |Fz + g|$$

پرسش ۶ مجموعه‌های محدب و نامحدب

محدب یا نامحدب بودن مجموعه‌های زیر را مشخص کنید.

- $\mathcal{A} = \left\{A \in \mathbf{S}_+^n \mid A_{ij} \geq (\frac{1}{i})^j,\ i,j = 1,\ldots,n\right\}$
- $\mathcal{B} = \{B \in \mathcal{C} \mid \det(B) \geq 0.5\}$ که $\mathcal{C} = \left\{X \in \mathbf{S}_+^n \mid X_{ii} = 1,\ i = 1,\ldots,n\right\}$
- $\mathcal{D} = \left\{D \in \mathbf{S}_+^n \mid \mathbf{rank}(D) \geq k\right\} \cup \{0\}$

پیروز و موفق باشید.