به نام خدا

جبر خطی

۰۰ـ۰۱ـ۲

پروژه‌ی نهایی

تاریخ بارگذاری: ۱۴۰۱/۰۳/۰۲، تاریخ تحویل: ۱۴۰۱/۰۴/۰۱

۱. با استفاده از قضیه‌ی درون‌یابی لاگرانژ و با بهره‌جویی از Matlab یا Python، منحنی‌ای بیابید که از نقاط داده شده در جدول ۱ عبور نماید. در جدول مذکور، $f(x) = e^{\sin(3x)}$ می‌باشد.

جدول ۱: نقاطی که منحنی می‌بایست از آن‌ها عبور کند.

x	0.1000	0.8225	1.5450	2.2675	2.9900
$f(x)$	1.3438	1.8667	0.3690	1.6426	1.5516

آ) معادله‌ی منحنی تخمین زده شده را بیابید.

ب) در یک نمودار، $f(x)$ و منحنی یافته شده را در بازه‌ی $[0, 3]$ رسم نمایید.

ج) نمودار اندازه‌ی خطای بین دو منحنی را در بازه‌ی $[0, 3]$ رسم نمایید.

۲. در این تمرین، هدف تخمین پارامترهای θ_i با استفاده از الگوریتم حداقل مربعات[1] برای سیستم

$$y = 2x_1 + 3x_2 - 4x_3$$

می‌باشد. بردارهای x_1 و x_2، بردارهایی با توزیع نرمال $\mathcal{N}(0, 1)$ به طول ۱۰۰۰۰ و $x_3 = x_1 - x_2 + \mathcal{N}(0, 10^{-20})$ می‌باشد. مدل را به صورت

$$\hat{y} = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

فرض کنید.

[1] Least-Square

آ) ابتدا ۷۰۰۰ داده را به عنوان داده‌ی یادگیری[۲] و ۳۰۰۰ داده را به عنوان داده‌ی ارزیابی[۳] جدا کنید و برای دادگان یادگیری، با استفاده از Matlab یا Python و با بهره‌جویی از الگوریتم حداقل مربعات، پارامترهای $\theta_i, i = 1, 2, 3$ را تخمین بزنید و سپس با استفاده از دادگان ارزیابی، درستی تخمین خود را بررسی نمایید. پاسخ سیستم اصلی (y) و مدل تخمین زده شده $(\hat{y})$ را به ازای ۳۰۰۰ داده‌ی ارزیابی در یک نمودار رسم کنید.

راهنمایی: برای تمامی دادگان، خروجی سیستم را بیابید و سپس، ۷۰۰۰ عنصر از بردارهای x و عناصر متناظر در y را به عنوان دادگان آموزشی در نظر بگیرید و ۳۰۰۰ داده‌ی باقی‌مانده را به عنوان دادگان ارزیابی فرض کنید.

ب) آیا تخمین درستی از پارامترهای θ در بند پیشین زده شده‌است؟ اگر پاسخ منفی است، دلیل خود را شرح دهید.

ج) اکنون با استفاده از الگوریتم گرام‌ـاشمیت[۴] ، برای دادگان آموزش، عملیات متعامدسازی را انجام دهید و همانند بند اول پارامترهای θ را تخمین بزنید. پاسخ سیستم اصلی (y) و مدل تخمین زده شده $(\hat{y})$ را به ازای ۳۰۰۰ داده‌ی ارزیابی در یک نمودار رسم کنید.

راهنمایی: متعامدسازی را فقط برای ۷۰۰۰ عنصر x که به عنوان داده‌ی آموزش برگزیده‌اید انجام دهید و سپس، خروجی‌های سیستم را برای این ۷۰۰۰ داده به دست آورید. نیازی به متعامدسازی دادگان ارزیابی نمی‌باشد.

د) آیا تخمین بند سوم با تخمین بند اول تفاوت دارد؟ در صورت مثبت بودن پاسخ، دلیل خود را اقامه نمایید.

ه) نمودارهای خطا را برای بندهای اول و سوم به صورت مجزا ترسیم نمایید.

۳. داده‌های دنیای واقعی اغلب شامل ویژگی‌های زیادی می‌باشند. از متداول‌ترین روش‌های کاهش ابعاد مجموعه داده‌ها، آنالیز مؤلفه اصلی یا PCA[۵] می‌باشد. PCA مؤلفه‌های اصلی را شناسایی و به ما کمک می‌کند تا به جای این‌که تمامیِ ویژگی‌ها را مورد بررسی قرار دهیم، یک مجموعه از ویژگی‌هایی را که دربردارنده اطلاعات اصلی است، تحلیل کنیم. در واقع، PCA آن ویژگی‌هایی که ارزش بیشتری دارند را برای ما استخراج می‌کند و برای این کار از جبرخطی کمک می‌گیرد.

آ) الگوریتم PCA را به تفصیل شرح دهید.

ب) تعدادی داده در آدرس درج شده وجود دارد. می‌توانید از ۱۰۰۰ داده به جای کل داده‌ها استفاده کنید. داده‌ها را در Matlab یا Python نمایش دهید.

[2] Train
[3] Test
[4] Gram-Schmidt
[5] Principal Component Analysis

ج) بر روی داده‌ها، الگوریتم PCA را بدون استفاده از کتابخانه‌ی آماده پیاده‌سازی کنید و داده‌های جدید را مشاهده کنید. بعد داده‌ها را با قسمت قبل مقایسه کنید. (برای پیدا کردن بهترین تعداد ویژگی، از روش آرنج[۶] استفاده کنید.)

۴. در زندگی روزمره، هزاران عکس و فایل‌های عکسی را به اشتراک می‌گذاریم. اما زمانی که نیاز به اشتراک‌گذاری عکس با حجم زیاد داشته باشیم، باید آن را فشرده کنیم. در این بخش قصد داریم فشرده‌سازی عکس به کمک روش SVD را بررسی کنیم. تجزیه مقادیر تکین یا SVD روشی دیگر برای کاهش بعد می‌باشد. لازم به ذکر است که این روش، بهترین روش برای فشرده‌سازی یک عکس نیست! اما در این سؤال تا حدودی به اهمیت مقادیر تکین در معرفی مطالب مهم یك مجموعه پی خواهیم برد.

آ) الگوریتم SVD را به تفصیل شرح دهید.

ب) عکسی را به دلخواه انتخاب کنید و در Matlab یا Python آن را مشاهده کنید (در صورتی که سایز عکس بزرگتر از ۳۰۰ در ۳۰۰ پیکسل است، سایز آن را به همین مقدار تغییر دهید. هم‌چنین در صورت رنگی بودن عکس، آن را به سیاه و سفید تبدیل کنید).

ج) حال الگوریتم SVD را بر روی عکس انتخابی ــ که اکنون به صورت ماتریس می باشد ــ بدون استفاده از کتابخانه‌ی آماده، پیاده‌سازی کنید.

د) تصویر را با مقادیر تکین مختلف نمایش دهید و با تصویر اصلی مقایسه کنید. هم‌چنین میزان کم‌حجم‌سازی را برای هر یک از حالت‌ها بیان کنید.

ه) اگر تنها ۳۰ درصد از بزرگ‌ترین مقادیر تکین را نگه داریم و بقیه را صفر کنیم، به ازای ماتریس جدید، تصویر جدید چگونه خواهد بود؟

۵. توانایی پیش‌بینی انتخاب‌های کاربران، یک تجارت بزرگ است. بسیاری از خدمات اینترنت در حال مطالعه انتخاب‌ها و ترجیحات مصرف‌کننده هستند تا بتوانند محصولاتی را که ممکن است مصرف‌کننده به آن‌ها علاقه داشته باشد، ارائه دهند. توجه داشته باشید که داده‌های مربوط به فیلم‌ها، آهنگ‌ها، محصولات مصرفی و غیره را اغلب می‌توان به شکل یک بردار مرتب کرد. سپس این نمایش‌های برداری را می توان در الگوریتم‌های مختلف برای مقایسه موارد و پیش‌بینی شباهت‌ها استفاده کرد. در این پروژه از جبرخطی برای ایجاد شباهت در سلیقه بین کاربران مختلف استفاده خواهیم کرد. ایده‌ی اصلی این است که اگر رتبه‌های فعلی را از یک کاربر خاص بدانیم، با مقایسه آنها با رتبه‌بندی سایر کاربران در یک پایگاه داده، می توانیم کاربرانی با سلیقه مشابه را پیدا کنیم. در نهایت، می‌توانیم به مواردی که توسط آن، کاربران رتبه بالایی دارند نگاهی بیاندازیم و این موارد را به کاربر فعلی پیشنهاد کنیم. در این سؤال، ما به مجموعه دادگان MovieLens نگاه خواهیم کرد که شامل حدود یک میلیون رتبه بندی از ۳۹۵۲ فیلم توسط ۶۰۴۰ کاربر است. ما با یافتن شباهت‌های بین سلیقه‌ی کاربران، یک سیستم توصیه‌گر بسیار ساده ایجاد خواهیم کرد.

[6] Elbow Method

آ) دادگان را بارگذاری کنید. فیلم‌های کاربران باید یک ماتریس 6040×3952 حاوی مقادیر صحیح بین ۰ و ۵ باشند که ۱ به معنای «اصلا دوست نداشتن» و ۵ به معنای «به شدت دوست داشتن» باشد. ۰ در ماتریس به این معنی است که کاربر به فیلم امتیاز نداده است. مرتب‌سازی فیلم‌های کاربران حاوی ۲۰ فیلم محبوب انتخاب شده است. در نهایت، رتبه‌بندی این فیلم‌های محبوب برای کاربر دیگری (که در پایگاه داده نمی باشد)، توسط کاربر آزمایشی [۷] داده می‌شود.

ب) کاربرانی را که کاربر آزمایشی را با آنها مقایسه می‌کنیم، انتخاب کنید (افرادی را انتخاب می‌کنیم که به تمام ۲۰ فیلم مورد بررسی امتیاز داده‌اند. این بدان معنی است که در ردیف های مربوط به مرتب سازی فیلم های کاربران ماتریسی نباید هیچ صفر وجود داشته باشد).

ج) به دنبال کاربری هستیم که نزدیک ترین امتیاز به کاربر آزمایشی را داشته باشد. اما ممکن است تفاوت کوچکی بین چندین کاربر نزدیک وجود داشته باشد. بنابراین کاربرها را بر اساس نزدیکی به کاربر آزمایشی مرتب کنید. از روش فاصله اقلیدسی استفاده کنید. ایراد استفاده از این معیار تشابه را بیان کنید.

د) حال قسمت قبل را با استفاده از روش ضریب همبستگی پیرسون حل کنید. چه نتیجه‌ای می‌گیرید ؟

$$r(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \qquad (۱)$$

ه) فیلم‌هایی که مورد علاقه‌ی کاربر آزمایشی بوده و فیلم‌های توصیه شده برای او را نمایش دهید.

و) با توجه به لیست فیلم‌ها، بردار رتبه‌بندی فیلم خود را ایجاد کنید و آن را MyMovies بنامید. رتبه بندی را بین ۱ تا ۵ مشابه قسمت اول انجام دهید. اگر فیلم خاصی را ندیده‌اید، رتبه آن را تصادفی انتخاب کنید.

ز) مراحل قبل را برای بردار رتبه‌بندی خود انجام دهید و فیلم‌های توصیه شده به شما را به دست آورید.

[7] trial-user

⚠ توجه: خواهشمند است جهت تحویل پروژه، به نکات زیر توجه نمایید:

آ) گزارش ارسالی باید به صورت تایپی و شامل فهرست مطالب، فهرست اشکال و فهرست جداول باشد. همچنین باید تمامی کدهای مربوط، به صورت قابل اجرا ارسال شوند تا قابلیت ارزیابی مستندات ارسالی وجود داشته باشد. در غیر این صورت نمره‌ای به این مستندات تعلق نخواهد گرفت.

ب) در صورت مشاهده‌ی مواردی از کپی‌برداری، دانشجویان خاطی (چه کپی‌کننده و چه کپی‌شونده) مشمول کسر نمره خواهند شد. در صورت احراز مشابهت در یک سؤال، نمره‌ی سؤال مربوطه به طور کامل از کپی‌کننده و کپی‌شونده کسر می‌گردد. در صورتی که مشابهت در حل بیش از نیمی از سؤالات احراز گردد، نمره‌ی پروژه به طور کامل از کپی‌کننده و کپی‌شونده کسر می‌شود.