به نام خدا

جبر خطی کاربردی دکتر امیرمزلقانی

تمرین سری پنجم (از فصل ششم و هفتم)

نیم‌سال دوم ۰۲-۰۱

بخش تئوری

سوال اول

بیشترین مقدار عبارت زیر را بدست آورید.

عبارت:

$$Q(x) = 7x_1^2 + 3x_2^2 - 2x_1x_2$$

محدودیت:

$$x_1^2 + x_2^2 = 1$$

سوال دوم

تجزیه مقدار منفرد ماتریس زیر را بدست آورید.

$$\begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix}$$

سوال سوم

مقادیر منفرد ماترس زیر را بدست آورید.

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

سوال چهارم

ثابت کنید مجموع درایه‌های روی قطر اصلی هر ماتریس قطری شدنی برابر است با مجموع مقادیر ویژه ماتریس.

سوال پنجم

معادله خط زیر را با کمک روش کمترین مربعات خطا بر روی داده‌های زیر پیدا کنید:

داده‌ها: (۴, ۳), (۲, ۱), (۱,۱) , (۰,۱-)

$$y = B_0 + B_1 x$$

سوال ششم

فرض کنید a, b بردارهایی به طول ۱ باشند و همچنین حاصل ضرب داخلی این دو بردار برابر ۰.۵- باشد اندازه بردار تفاضل را بدست آورید.

سوال هفتم

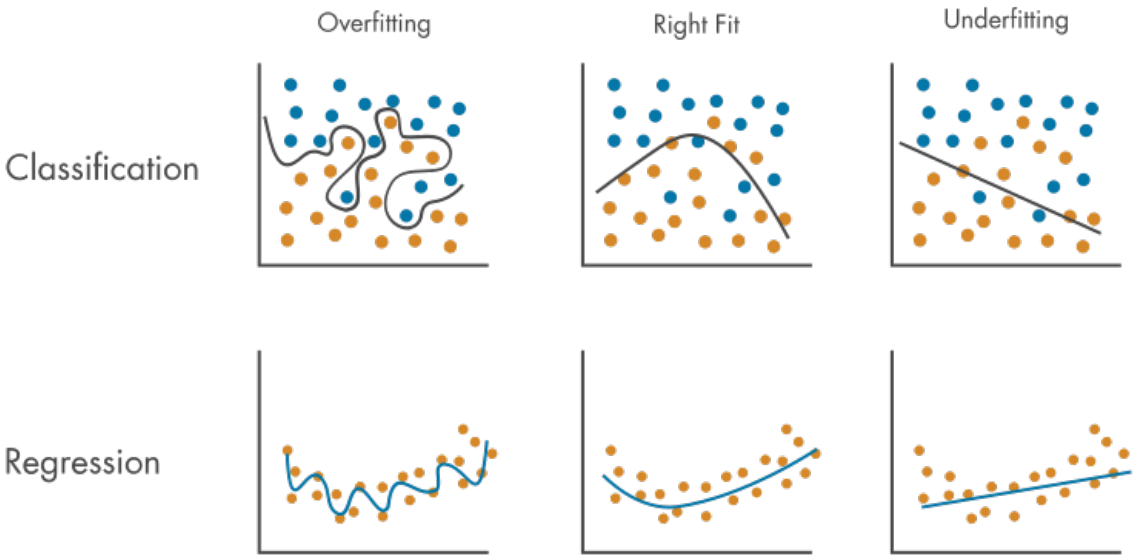
اثبات کنید که عبارت زیر همواره مثبت معین است.

$$Q(x_1, x_2) = 3x_1 - 4x_1x_2 + 6x_2^2$$

بخش پیاده‌سازی

تعریف مسئله

در بسیاری از مسئله ها در حوزه علم داده، با شرایطی رو به رو می شویم که داده های به نسبت کمی از ابعاد بالا (High Dimensional Data) داریم و نیاز داریم که مدلی مبتنی بر یادگیری ماشین طراحی کنیم که بتواند کارایی لازم را در کاربرد مورد نظر ما داشته باشد. اما یکی از چالش هایی که در این موارد به وجود می آید این است که در صورتی که مدل، بر روی تعداد داده های محدود، از ابعاد بالا آموزش ببیند، ممکن است دچار بیش برازش (Overfitting) شود که باعث می شود مدل ما در داده های آموزشی عملکرد خوبی داشته باشد ولی در مواجهه با داده های جدید عملکرد ضعیفی داشته باشد. تصویر زیر نمایشی از Overfitting در برابر Right Fit می باشد:



یکی از راه حل هایی که می توان از آن برای حل این مشکل استفاده کرد، کاهش ابعاد داده (*Dimensionality Reduction*) می باشد. در این تمرین قصد داریم با استفاده از مفاهیمی که در درس جبر خطی خوانده اید (*SVD and PCA*) به کاهش ابعاد داده ها بپردازیم. قابل ذکر است که کاهش ابعاد داده ها، علاوه بر حل مشکل بیش برازش، باعث کاهش هزینه محاسباتی، تسهیل مصورسازی داده‌ها (*Visualization*) و ... می شود.

توجه: پیشنهاد می شود برای فهم بهتر مفهوم PCA، به ویدیو موجود در این لینک مراجعه کنید.

پیاده سازی

برای این پروژه، یک فایل دیتاست به نام "nndb_flat.csv" در اختیار شما قرار می گیرد که شامل داده های مواد غذایی و اطلاعاتی درباره آن ها می باشد. این دیتاست دارای بیش از ۸۰۰۰ رکورد می باشد که هر رکورد نماینده یک ماده غذایی می باشد و ستون های آن (ID, FoodGroup, ShortDescription, ...) اطلاعات مربوط به آن رکورد را دارد. این دیتاست، ۴۵ ستون دارد.

در این پروژه، قسمت هایی از کد پیاده سازی شده اند و در اختیار شما قرار می گیرند و آنچه از شما انتظار می رود این است که با استفاده از تجزیه مقادیر تکین (SVD) و تحلیل اجزای اصلی (PCA) ابعاد داده های موجود را کاهش دهید. برای این کار باید ماتریس ورودی را با استفاده از SVD تجزیه کنید و از k مقادیر ویژه بزرگتر آن و بردارهای مربوط به این k مقدار ویژه در دو ماتریس دیگر حاصل از تجزیه، برای بازسازی مجدد ماتریس اصلی استفاده کنید. این عمل را برای مقادیر k = 3, 5, 8, 10, 15 انجام دهید و درصد حفظ داده های ماتریس اصلی در ماتریس بازسازی شده را برای همه مقادیر k انجام دهید.

توجه: برخی از ستون های دیتاست به صورت متنی می باشد و شما نیاز دارید که ستون های متنی را از این دیتاست حذف کنید (راهنمایی: از تابع drop در کلاس DataFrame استفاده کنید.)

توجه: قبل از اجرای SVD روی داده ها، نیاز داریم که مقادیر ستون های این دیتاست را Scale کنیم تا SVD به درستی عمل کند و تفاوت نسبت اندازه های مقادیر در ستون های مختلف، در عملکرد این الگوریتم اختلال ایجاد نکند. این قسمت برای شما پیاده سازی شده است و تنها نیاز است که DataFrame شما در متغیری به اسم df ذخیره شده باشد. برای اجرای صحیح کد، نیاز دارید که کتابخانه sklearn را در محیط پایتون خود نصب کنید. (Scikit-Learn Installation)

نکته: برای محاسبه درصد حفظ داده ها می توانید از فرمول زیر استفاده کنید:

$$Retained\ Percentage = \frac{\sum_{i=1}^{k}(\sigma_i)^2}{\sum_{i=1}^{total\_number\_of\_singular\_values}(\sigma_i)^2} * 100$$

مشاهده خواهیده کرد که با افزایش مقدار k، درصد حفظ اطلاعات بیشتر می شود ولی ابعاد داده ها نیز افزایش می یابد.

توجه: برای ارزیابی صحت کدهای خود می توانید درصدهای به دست آمده از کد خودتان را با مقداری که تابع از پیش نوشته شده get_percentage در فایل main.py به دست می آورد مقایسه کنید.

امتیازی: برای تجزیه مقادیر تکین، می توانید از کتابخانه numpy.linalg استفاده کنید ولی پیاده سازی تابع SVD بدون استفاده از این کتابخانه، نمره امتیازی دارد.

مصورسازی (Visualization)

این بخش صرفا برای علاقه مندان، برای درک بهتر PCA و افزایش دقت با افزایش مقدار k آورده شده است و نمره ای به آن تعلق نمی گیرد. شما می توانید با فراخوانی تابع visualize در این فایل، شباهت داده بازسازی شده با داده اصلی را بررسی کنید. همان طور که در قبل اشاره شد، با افزایش تعداد بردارهای تکین استفاده شده در بازسازی داده اصلی، شباهت بیشتر می شود.

نکات:

- برای خواندن داده های دیتاست، از کتابخانه pandas استفاده کنید.
- در طول این پروژه، فقط استفاده از کتابخانه های numpy, pandas, matplotlib مجاز می باشد.

دانشجویان عزیز توجه کنید که:

❊ فایل پاسخ خود را تنها به شکل <<StuNum_HWNum.pdf>> نام گزاری کنید. (به عنوان مثال HW1.pdf_۴۰۰۱۲۳۴۵۶)

❊ فایل پاسخ علاوه بر پاسخ بخش تئوری باید حاوری گزارش و تحلیل نتایج به دست آمده از بخش پیاده‌سازی‌ها باشد.

❊ در صورت شبیه بودن پاسخ تمارین دانشجویان، نمره تمرین بین دانشجویان با پاسخ تمرین مشابه تقسیم خواهد شد.

❊ اگر هرگونه سوال و ابهامی داشتید با یکی از ایمیل‌ها یا آیدی‌های تلگرامی زیر ارتباط برقرار کنید.

farhaaaaa1@gmail.com

sepehr.nk.81@gmail.com

@sepehr_Noey2081