جبر خطی پاییز ۱۴۰۳

پروژه کامپیوتری

استاد: دکتر امیری

طراحان: ماردین نیچی و مانی میرشعبانی

دانشگاه تهران

دانشکده علوم مهندسی

در این پروژه میخواهیم با یک ابزار جبر خطی برای کاهش بعد آشنا بشویم. Prin- Component Analysis ciple یا همان PCA پایه هایی در فضا را پیدا خواهد کرد که داده ها بیشترین پراکندگی را در آن راستا دارند. در این بخش پروژه مرحله به مرحله جلو میرویم تا در نهایت PCA را اثبات وسپس پیاده کنیم.

در ابتدا باید با دیتاست D آشنا شویم. هر دیتاست یک ماتریس $n * d$ است که در هر سطر یک نمونه داده و در هر ستون یک ویژگی یا attribute قرار دارد. بطور مثال در شکل۱ نمونه های ما بترتیب x_i تا x_n و ویژگی های هر داده بترتیب X_i تا X_d است. x_{ij} نیز مقدار ویژگی j نمونه i را نشان میدهد.

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

شکل ۱

در درس آمار با واریانس σ آشنا شدید واریانس یک معیار پراکندگی برای داده ها است که با $E[(x-\mu_{X_i})^2]$ محاسبه میشود که μ_{X_i} میانگین مقدار ویژگی X_i برای تمام داده هاست. وقتی تعداد ویژگی ها از یک بیشتر میشود آماره کوواریانس معرفی میشود که وابستگی خطی بین دو ویژگی را نشان میدهد و بصورت $\sigma_{Y,T} = E[(x_Y - \mu_Y)(x_T - \mu_T)]$ محاسبه میشود.

۱ـ روشی ارائه دهید و با استفاده از دیتاست D، correlation matrix را محاسبه کند.

correlation matrix ماتریسی $d*d$ است که درایه i,j آن مقدار کوواریانس میان ویژگی i و j را نشان میدهد.

۲ـ فرض کنید بردار u_1 که $u_1^T.u_1 = 1$، اولین برداری باشد که تصویر داده ها روی آن بیشترین واریانس را دارد. ثابت کنید که $\sigma_{u_1} = u_1^T \Sigma u_1$ خواهد بود. (راهنمایی: تعریف واریانس را بنویسید و برای تصویر x_i روی بردار

u_1 از $u_1^T x_i$ استفاده کنید. سپس توان دو را باز کرده و ساده کنید.)

حال برای بیشینه کردن واریانس، کافی است که مساله بهینه سازی $\max_u \ u^T \Sigma u$ را حل کنیم. مقدار واریانس را با بزرگ کردن نرم u_1 میتوان بزرگ کرد ($u^T.u =\rightarrow \inf$ باعث میشود که $u^T \Sigma u \rightarrow \inf$) که این کار نه تنها جواب خواسته شده را به ما نمیدهد بلکه حل مسئله را دچار مشکل میکند. پس باید قید $u^T.u = 1$ را به مسئله ی بهینه سازی اضافه کرد. برای حل مسئله بهینه سازی با قید از روشی به اسم ضرایب لاگرانژ استفاده میکنیم. این روش قیود مسئله را همراه یک ضریب به عنوان پنالتی به خود تابع هدف اضافه میکند که در نهایت مسئله به شکل زیر درمیاید

$$\max_u \ u^T \Sigma u - \lambda \left(1 - u^T u\right)$$

توضیح بیشتر ضرایب لاگرانژ خارج از اهداف این پروژه است. برای اطلاعات بیشتر در این زمینه میتوانید به منابع بهینه سازی و تحقیق در عملیات مراجعه کنید.

۳ـمیدانیم که نقاط بهینه در نقاطی از تابع قرار میگیرند که مشتق آن برابر صفر باشد. پس از تابع $-u^T \Sigma u - \lambda \left(1 - u^T u\right)$ نسبت به u مشتق بگیرید و آن را برابر صفر قرار دهید و اثبات کنید که برداری که داده ها بیشترین واریانس را روی آن دارند در واقع بردار ویژه Σ و واریانس داده ها روی این بردار مقدار ویژه متناظر با آن است.

حالا میتوانیم حدس بزنیم که احتمالا بردارهایی که تصویر داده ها روی آن ها بیشترین واریانس را دارد، بردار ویژه های Σ است. برای این که ادعای خود را ثابت کنیم از استقرا استفاده میکنیم.
پایه استقرا برای $n = 1$ را در بالا اثبات کردید.
فرض استقرا: $\{u_1, u_2, \dots u_{j-1}\}$، $j-1$ بردار اولی باشند که داده ها در آن بیشترین واریانس را دارند. این بردار ها متعامد هستند و بردار ویژه های ماتریس Σ خواهند بود.
گام استقرا: jامین بردار که داده ها در آن بیشترین واریانس را دارند، بردار ویژه Σ متناظر با j امین بزرگترین مقدار ویژه است و با تمامی بردار های قبلی متعامد است.
در این قسمت مسئله بهینه سازی ما ۲ نوع قید دارد. یکی متعامد بودن و دومی نرمال بودن بردار است. با استفاده از ضرایب لاگرانژ قید ها را به تابع هدف اضافه میکنیم و به رابطه ی زیر میرسیم

$$\max_{\mathbf{v}} J(\mathbf{v}) = \mathbf{v}^T \Sigma \mathbf{v} - \alpha(\mathbf{v}^T \mathbf{v} - 1) - \sum_{i=1}^{j-1} \beta_i (\mathbf{u}_i^T \mathbf{v} - 0)$$

۴ـاز این تابع نسبت به v مشتق بگیرید و مانند قسمت قبل آن را برابر با صفر قرار بدهید. در نهایت اثبات کنید که v نیز بردار ویژه ی Σ است.

۵ـ هر داده در پایه های جدید چگونه نشان داده میشود؟

۶ـ ثابت کنید جمع واریانس ها در پایه های جدید برابر است با جمع روی مقدار ویژه های متناظر با پایه ها.

تا اینجا فهمیدیم که برای بدست آوردن PCA باید correlation matrix را بدست بیاوریم و تصویر داده ها را بر روی بردار ویژه های آن محاسبه کنیم. اکنون از شما میخواهیم که کد PCA را در زبان برنامه نویسی پایتون پیاده سازی کنید.
توجه کنید که مجاز به استفاده از توابع آماده نیستید و کد هر قسمت را باید از پایه (from scratch) پیاده سازی کنید. بطور مثال برای بدست آوردن بردار ویژه ها میتوانید از QR-Algorithm استفاده کنید.
پاسخ سوالات را یا در یک فایل pdf یا بصورت mark down در فایل نوت بوک بنویسد.

در فاز بعدی پروژه باید به کمک کدی که در فاز قبلی پیاده سازی کرده اید، فایل نوتبوکی که در اختیارتان قرار گرفته شده است را کامل کنید.
اطلاعات مورد نیاز برای این قسمت در فایل نوت بوک آمده است.