




تحليل كلان داده‌ها

تمرین سوم

{Data Stream, SVD, Recommender System}

مهلت تحویل

۱۴۰۲/۰۴/۰۷



برای ارسال تمرین به نکات زیر توجه کنید.

۱. ملاک اصلی انجام تمرین گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این تمرین یک فایل گزارش در قالب pdf تهیه کنید و در آن برای هر سوال، تصاویر ورودی، تصاویر خروجی و توضیحات مربوط به آن را ذکر کنید. سعی کنید توضیحات کامل و جامعی تهیه کنید.
۲. زبان برنامه‌نویسی برای انجام تمرین‌ها، پایتون(Python) در نظر گرفته شده است.
۳. برای سهولت در انجام تمرینات، توصیه می‌شود که پلتفرم کولب گوگل استفاده نمایید.
۴. در نظر داشته باشید کدهای شما باید قابلیت اجرا در هنگام ارائه را داشته باشند. همچنین بر روی کدهای خود مسلط باشید
۵. کدهای ارسالی خود را برای افزایش خوانایی و درک بهتر به صورت مناسب کامنت‌گذاری کنید.
۶. پاسخ سوالات تشریحی، فایل‌های کد و گزارش خود را در یک فایل فشرده قرار داده و با نام با فرمت HW3_StudentID ارسال نمایید.
۷. در صورت وجود ابهام یا اشکال می‌توانید از طریق [کانال](#) با ما در ارتباط باشید(داخل کانال لینک گروه نیز موجود می‌باشد)

هدف از این سوال آشنایی شما با کتابخانه ^۱scikit-multiflow و استفاده از آن برای طبقه‌بندی جریان داده است. در کلاس با طبقه‌بند VFDT آشنا شدید. در اینجا می‌خواهیم با طبقه‌بند جدیدتری به نام ^۲Adaptive Random Forest classifier آشنا بشیم و برای طبقه‌بندی کردن دیتاست mnist به صورت استریم از آن استفاده کنیم. مراحل زیر را دنبال کنید.

۱- نصف تا یک صفحه نحوه عملکرد Adaptive Random Forest classifier را توضیح دهید.

۲- دیتاست mnist را در محیط برنامه‌نویسی خود لود کنید.

۳- با استفاده از DataStream از کتابخانه scikit-multiflow از دیتاست خود یک استریم بسازید.

۴- با استفاده از کتابخانه scikit-multiflow از Adaptive Random Forest classifier برای طبقه‌بندی دیتاست mnist استفاده کنید. دقت کنید باید بصورت incremental مدل را برای هر داده آموزش دهید. می‌توانید با آزمون و خطا پارامترهای مدل را برای نتیجه بهتر تغییر دهید.

۵- نمودار صحت طبقه‌بندی مدل در طول کل آموزش و همچنین نمودار صحت طبقه‌بندی مدل برای هر ۱۰۰ دیتا اخیر را رسم کنید.

^۱<https://scikit-multiflow.readthedocs.io/en/stable/index.html>

^۲ Heitor Murilo Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabricio Enembreck, Bernhard Pfahringer, Geoff Holmes, Talel Abdessalem. Adaptive random forests for evolving data stream classification. In Machine Learning, DOI: 10.1007/s10994-017-5642-8, Springer, 2017.

۱- توزیع گوسی با مشخصات زیر را در نظر بگیرید که میانگین صفر و ماتریس کواریانس آن واحد است:

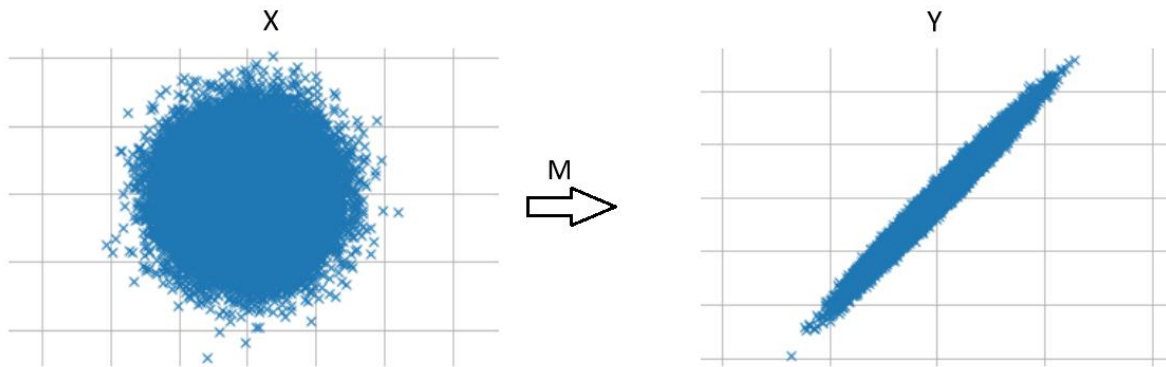
$$N \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

ما از توزیع بالا نمونه دو بعدی تولید می کنیم $\begin{bmatrix} x \\ y \end{bmatrix}$ و اگر به تعداد n بار این کار را انجام دهیم در نهایت ماتریس X که داده های آن از توزیع گوسی نمونه برداری شده اند به شکل زیر می باشد:

$$X = \begin{bmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{bmatrix}_{2 \times n}$$

نمونه های بالا از توزیع گوسی تولید شده اند، پس ماتریس X دارای میانگین $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ و کواریانس $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ خواهد بود. حال یک ماتریس دو در دو، $M_{2 \times 2}$ را تعریف می کنیم. با استفاده از این ماتریس داده های X را به فضای جدید منتقل می کنیم. این داده ها در فضای جدید را با ماتریس Y نشان می دهیم به عبارتی:

$$Y_{2 \times n} = M_{2 \times 2} \times X_{2 \times n}$$



با توجه به موارد گفته شده بخش های زیر را انجام دهید:

الف) تجزیه svd ماتریس M و همچنین تجزیه svd ماتریس کواریانس Y را به صورت زیر در نظر بگیرید:

$$M = U_M \Sigma_M V_M^T \quad \text{covariance}(Y) = C = U_C \Sigma_C V_C^T$$

چه ارتباطی بین Σ_M و Σ_C وجود دارد؟ چه ارتباطی بین U_M و U_C وجود دارد؟ صرفاً نشان دادن این ارتباط کافی نیست و باید نحوه رسیدن به این ارتباط را ثابت کنید.

ب) تجزیه svd ماتریس M و همچنین تجزیه svd ماتریس Y را به صورت زیر در نظر بگیرید:

$$M = U_M \Sigma_M V_M^T \quad Y = U_Y \Sigma_Y V_Y^T$$

چه ارتباطی بین Σ_Y و Σ_M وجود دارد؟ این ارتباط را یا باید ثابت کنید یا می توانید از شبیه سازی کامپیوتری برای نشان دادن این ارتباط کمک بگیرید و در این مورد توضیح دهید. در نهایت باید این ارتباط را نشان دهید.

در این بخش، دیتاست Y_1 و Y_2 در اختیار شما قرار گرفته است. Y_1 و Y_2 به صورت زیر بدست آمده است. (M یک ماتریس دو در دو می باشد):

$$Y_1 = M \times X_{2 \times n} \quad Y_2 = M^T \times X_{2 \times n}$$

(توجه کنید هنگام لود کردن دیتاست Y_1 و Y_2 ، قبل استفاده از آن ترانپاده بگیرید.)

ماتریس X هم همانطور که قبلا اشاره شد دارای نمونه هایی از توزیع گوسی با میانگین صفر و ماتریس کواریانس واحد است.

ج) با توجه به مواردی که در بخش های قبلی فرا گرفتید. ماتریس M را پیدا کنید.

د) در این بخش از توزیع گوسی با میانگین صفر و ماتریس کواریانس واحد ، ۵۰ هزار نمونه تولید کنید و با استفاده از ماتریس M و M^T که در بخش قبل بدست آوردید به فضای جدید منتقل کنید. داده ها در فضای جدید را \hat{Y}_1 و \hat{Y}_2 بنامید. حال Y_1 ، \hat{Y}_1 و Y_2 و \hat{Y}_2 را در کنار هم رسم کنید و با هم مقایسه کنید و ببینید ماتریس M را چگونه بدست آورده اید.

برای تجزیه svd و تولید نمونه گوسی می توانید از توابع آماده زیر استفاده کنید:

`numpy.linalg.svd()`

`numpy.random.multivariate_normal()`

هدف از این بخش پیاده‌سازی الگوریتم Stochastic Gradient Descent برای ساختن یک سیستم توصیه‌گر شخصی سازی شده (Personalized) مبتنی بر Matrix Factorization است. در این رویکرد ما به دنبال پیدا کردن دو ماتریس P و Q هستیم به نحوی که $R \cong QP^T$ ، در این رابطه ماتریس R بیان‌گر ماتریس تعامل کاربر-آیتم می باشد. اندازه ماتریس R برابر با $m \times n$ است که m تعداد کاربرها و n تعداد آیتم‌ها می باشد.

مجموعه داده‌ای که در اختیار شما قرار داده شده است، فاقد امتیاز کاربر به آیتم‌ها بوده و صرفاً نشان‌دهنده تعامل کاربر با آن آیتم است (implicit feedback).

تابع خطا به صورت زیر تعریف می شود. (برای درک بهتر این قسمت توصیه می شود اسلایدهای سری دوم بخش سیستم توصیه‌گر را مشاهده فرمائید).

$$E = \left(\sum_{(u,i) \in \text{training}} \log(\text{sigmoid}(r_{ui} - q_i \cdot p_u^T)) \right) + \lambda \left(\sum_i \|q_i\|_2^2 \sum_u \|p_u\|_2^2 \right)$$

الف) \mathcal{E}_{ui} برابر مشتق تابع خطا نسبت به r_{ui} است. این عبارت را با استفاده از مشتق گیری به دست آورید (با عبارات ریاضی)

ب) الگوریتم Stochastic Gradient Descent را پیاده سازی کنید. ارزیابی خود را بر روی ۱۰ آیتم برتر انجام دهید مقدار λ را برابر ۰.۱ و تعداد تکرار نهایتاً تا ۱۰۰ در نظر بگیرید. مقدار بهینه برای نرخ یادگیری را بدست آورید.

برای حل این سوال از قالب آماده موجود در زمینه استفاده کنید.