تمرین سری چهارم مباحثی در ریاضیات

نگار یگانه

40113437

بخش اول) در این بخش هدف استفاده از الگوریتم isomap برای تجسم فواصل بین شهر های ایران بوده است که به دوصورت دستی و استفاه از تابع اماده پیاده سازی شد. همچنین برای هر کدام از پیاده سازی ها با استفاده از دوران وانعکاس سعی کردیم به نقشه واقعی خروجی را نزدیک کنیم.

برای پیاده سازی دستی این الگوریتم به این صورت عمل کردیم که ابتدا گراف مجاورت را با پیدا کردن کا نزدیک ترین همسایه های هر شهر ایجاد کردیم سپس با استفاده از الگوریتم دایکسترا فاصله ژئودزیک بین همه زوج شهرها محاسبه می‌شود و ماتریس فواصل ژئودزیک را بدست می سازیم. سپس الگوریتم MDS بر روی ماتریس فواصل ژئودزیک به کار گرفته می شود. بدین صورت که ماتریس Hرا که کاربردش حذف اثر میانگین نقاط و آوردن داده‌ها به حول مبدأ می سازیم که این مارتیس برابر است با تفاضل ماتریس همانی با ماتریسی که تمام درایه هایش $1/n$ است. سپس ماتریس $B = -\frac{1}{2}H(\bar{A})H$ که $\bar{A} = \left(a_{ij}^2\right)$ ساخته می شود . مقدایر و بردار ویژه های این ماتریس را حساب کرده و ضرب حاصل از جذر مقادیر ویژه و بردار ویژ] را بدست می

آوریم. فقط مقادیر ویژه های بزرگتر و بردار ویژه های متناظر با آن ها مورد استفاده قرار می گیرند.

از ماتریس های چرخش برای قابل فهم تر شدن خروجی های هر دو isomap استفاده کردیم البته زوایه هایی که استفاده شد متفاوت است.

با توجه به خروجی های تصویری می توان دید که خروجی ها بهم نزدیک بوده اند و نزدیک نقشه واقعی ایران هستند. از سه معیار هم برای نشان دادن نزدیک بودن خروجی ها استفاده شد : معیار پروکروستس که یک روش آماری برای مقایسه‌ی دو مجموعه از نقاط است که ممکن است تحت تبدیل‌های هندسی مانند انتقال، چرخش یا مقیاس‌دهی قرار گرفته باشند و این تحلیل بهترین هم‌ترازی بین دو مجموعه نقاط را پیدا می‌کند و میزان تفاوت آن‌ها را به صورت عددی نشان می دهد با توجه به مقدار پایین بدست امده برای این معیار می توان متوجه شد اختلاف کمی در ساختار دو خروجی وجود دارد. همچنین مقادیر بالا بدست امده برای همبستگی مولفه ها نیز شباهت دو خروجی را نشان می دهد. معیار دیگه ای که استفاده کردیم حفظ همسایه های نزدیک بود که جهت بررسی این است که آیا نقاطی که در فضای اصلی به‌هم نزدیک بوده‌اند، در فضای جدید نیز نزدیک باقی مانده‌اند یا نه. برای هر نقطه، k همسایه‌ی نزدیک آن در هر دوپیاده سازی استخراج می‌شود. سپس بررسی می‌کنیم چه تعداد از این همسایه‌ها در هر دو تعبیه مشترک هستند. نهایتاً، درصد همسایگی حفظ‌شده به صورت میانگین محاسبه می‌شود که 89 درصد بود.

بخش دوم) در این بخش، دو روش متداول کاهش بُعد یعنی PCAو t-SNEبر روی مجموعه داده‌ی MNISTاعمال شده‌اند و مورد ارزیابی قرار گرفتند.

مجموعه داده‌یMNIST تک بعدی شامل ۷۸۴ ویژگی برای تصاویر ارقام 0 تا 9 است. برای کاهش زمان اجرا، تنها از ۴۰،۰۰۰ نمونه تصادفی استفاده شده است که به صورت متوازن از بین کلاس‌ها انتخاب شده‌اند.

خروجی های بصری و زمانی هر دو روش چاپ شده اند که واضحا روش PCAزمانی بسیار کمتری داشت. اما در t-SNE، هر رقم تقریباً به صورت خوشه‌ای مشخص تفکیک شده است و در نتیجه خوشه بندی واضح تری دارد.

مقدار trustworthiness برای هر دو روش بسیار بالاست اما برای t-SNE به مقدار 0.04 بیشتر است . البته در نمونه گیری هایی با مقدار پایین تر این اختلاف بیشتر است.

توضیح trustworthiness :

این معیارنشان می‌دهد که ساختار محلی داده‌ها تا چه حد حفظ شده است در واقع اندازه‌گیری می‌کند که آیا همسایگی‌های اصلی در فضای پایین‌بعد نیز حفظ شده‌اند یا خیر و اگر نقاطی که در فضای اصلی دور از هم بوده‌اند، بعد از کاهش بُعد به هم نزدیک شوند، معیار Trustworthiness کاهش پیدا می‌کند.

این معیار با فرمول زیر محاسبه می شود :

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i^k} \max(0, (r(i,j) - k))$$

که در آن برای هر نمونهi ، Nنشان‌دهنده k نزدیک‌ترین همسایه‌های نمونه i در فضای خروجی است، و هر نمونهj ،rank(i,j) نشان‌دهنده رتبه همسایگی نمونه j نسبت به

نمونه i در فضای ورودی است. به عبارت دیگر، هر همسایه غیرمنتظره (نزدیک‌ترین همسایه‌ای که در فضای خروجی ظاهر شده ولی در فضای ورودی در رتبه‌های پایین‌تر قرار دارد) متناسب با رتبه‌اش در فضای ورودی جریمه می‌شود. پس اگر Trustworthinessپایین باشد، به این معنی است که روش کاهش بعد نقاطی را که در واقع دور از هم بوده‌اند، به‌صورت اشتباه به هم نزدیک کرده است.

ادامه توضیحات پیاده سازی :

در ادامه t-SNE برای perplexity های مختلف اجرا شد و خروجی ها چاپ شد. با افزایش perplexity مقدار انحراف KL کاهش می یابد به طوری مقدار بهینه perplexity با توجه به انحراف KL ، 100 بدست اومد البته این بدین معنا نیست که با افزایش preplexity، به کاهش بعد بهتری از لحاظ بصری می رسیم چون می توان به صورت بصری دید که خوشه در حال ادغام شدن و بدتر شدن هستند.

Perplexity معیاری است برای ایجاد تعادل میان مقیاس های local و global در نمایش داده های کاهش یافته. مقدار این پارامتر مشخص می کند که t-SNE چه تعداد همسایه های نزدیک را در نظر بگیرد.هر چه مقدار آن کم باشد تمرکز بیشتر بر همسایه های خیلی نزدیک خواهد بود که باعث خوشه بندی محلی تر می شود و خوشه ها نزدیک تر بهم شکل میگیرند و ممکن است درک روابط را کمی دشوار کند و هرچه مقدارش بالا تر باشد الگوریتم تعداد بیشتری از نقاط را در نظر می گیرد مه باعث نمایش کلی تر از ساختار می شود ولی ممکن است جزئیات محلی از بین برود.

KL divergence معیاری است که اندازه‌گیری می‌کند چقدر توزیع احتمال داده‌های پربعد با توزیع داده‌های کم‌بعد تفاوت دارد .در واقع، الگوریتم سعی می‌کند KL-Divergenceرا کمینه کند تا بتواند ساختار داده‌ها را بهتر حفظ کند. اما لزوما کاهش این مقدار به معنای نمایش بصری بهتر نیست.

منابع :

trustworthiness — scikit-learn 1.6.1 documentation

جزوه

t-SNE: The effect of various perplexity values on the shape — scikit-learn 1.6.1 documentation