دانشگاه صنعتی شریف

دانشکدهٔ فیزیک

تمرین سری (۲) درس موضوع‌های منتخب در فیزیک آماری

نام و نام خانوادگی: اشکان دماوندی

شمارهٔ دانشجویی: ۴۰۰۱۰۲۵۷۷

استاد درس: دکتر محمدرضا رحیمی‌تبار

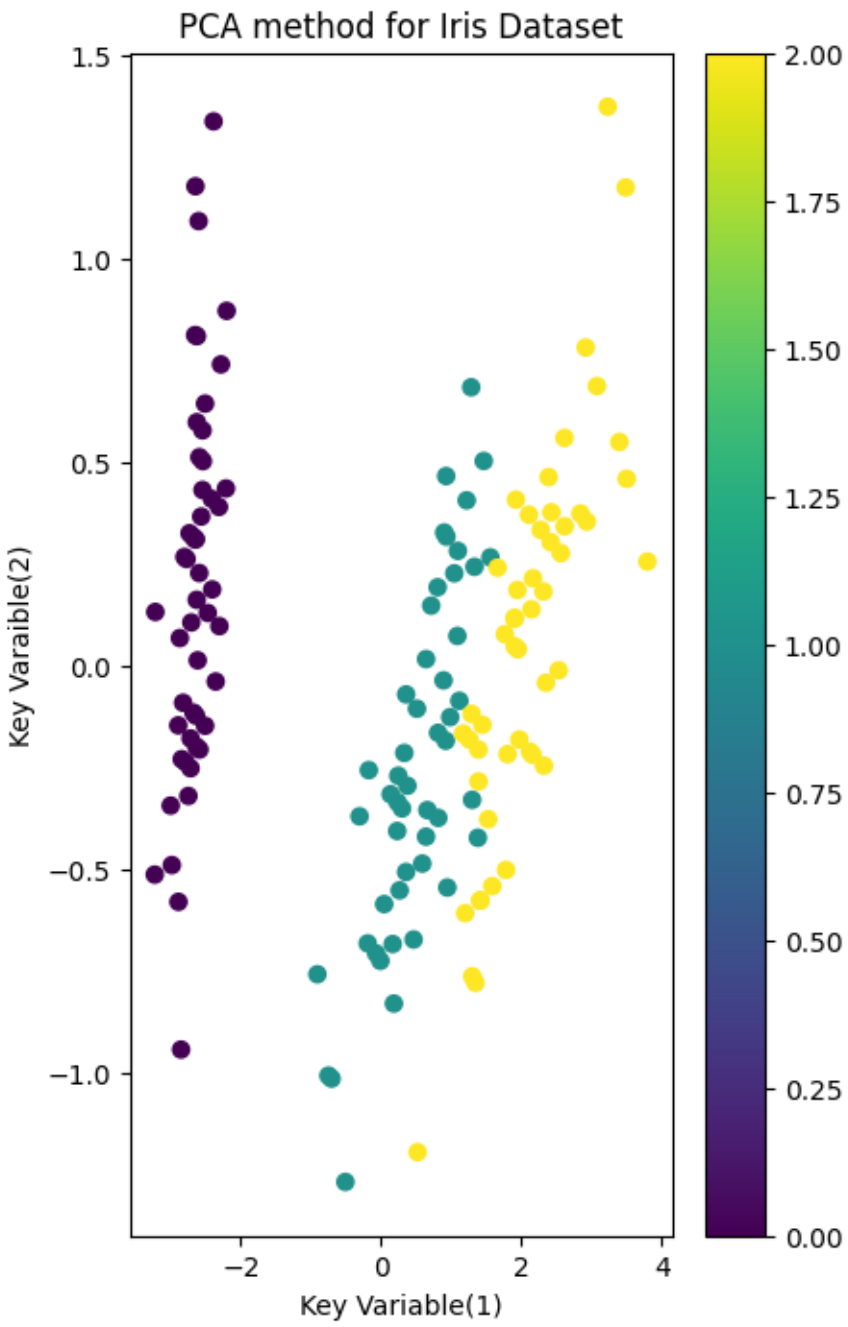
نیم‌سال ۱۴۰۳-۲

1 Latent Variables

متغیرها (Variables) به‌نوعی مکانی برای ذخیرهٔ ویژگی‌های داده هستند. متغیرهای Latent متغیرهایی هستند که بر اساس متغیرهای قابل‌مشاهده (Observable Variables) تعریف و استنباط می‌شوند. این نوع متغیرها می‌توانند متغیرهای تصادفی در یک مدلسازی آماری باشند که بر اساس متغیرهای مشاهده یا اندازه‌گیری‌شده تعریف می‌شوند و می‌توانند ارتباط بین آن‌ها را توضیح دهند. متغیرهای Latent معمولاً ویژگی‌هایی را توصیف می‌کنند که به‌صورت مستقیم قابل‌مشاهده یا اندازه‌گیری نیستند. به‌عنوان مثال اگر هدف یک بررسی آماری اندازه‌گیری میزان تأثیرگذاری یک خبر منتشرشده در شبکه‌های اجتماعی باشد، میزان تأثیرگذاری را نمی‌توان به‌صورت مستقیم اندازه‌گیری کرد و برای اندازه‌گیری آن باید تغییراتی که آن خبر در یک‌سری متغیرهای قابل‌اندازه‌گیری یا مشاهده ایجاد کرده را با استفاده از روش‌های آماری مدلسازی کرد؛ بنابراین میزان تأثیرگذاری به‌صورت متغیر یا مدلی بر اساس متغیرهای قابل‌اندازه‌گیری یا مشاهده گزارش داده می‌شود.

2 Methods for Dimensional Reduction of Multidimensional Data

در بحث آنالیز داده‌ها و یادگیری ماشین (Machine Learning)، کاهش ابعاد داده‌ها (Dimensional Reduction) از روش‌هایی است که تحلیل و visualize کردن داده‌ها را بهینه‌تر می‌کند و در مواقعی که با داده‌های پیچیده و یا مجموعه‌ای داده‌ها که ابعاد بالایی دارند کار می‌کنیم، در کاهش نویز آن‌ها مؤثر واقع می‌شود. در این روش ویژگی و اطلاعات اساسی و معنادار داده‌ها با استفاده از کاهش تعداد متغیرهای پیش‌بینی‌کننده (Predictor Variables)، جهت افزایش تعمیم‌پذیری آن‌ها حفظ می‌شود. در بحث Machine Learning، ابعاد یا ویژگی‌ها همان متغیرهای پیش‌بینی‌کننده‌ای هستند که خروجی مدل مورد نظر را تعیین می‌کنند و Dimensional Reduction باعث افزایش کارایی این مدل‌ها می‌شود. روش‌های مختلفی برای Dimensional Reduction وجود دارد که ۴ تا از آن‌ها را معرفی کرده و مثالی از کد پایتون مورد نیاز برای استفاده از آن‌ها قرار می‌دهم. در کدها، از ماژول Scikit-Learn استفاده کردم که ماژولی برای استفاده از متدهای یادگیری ماشین در پایتون است و برای داده‌ها، از Iris flower data set استفاده کردم که معمولاً در Machine Learning برای بحث تشخیص الگو استفاده می‌شود. این data set دارای ۱۵۰ نمونه از سه گونه گل Iris است؛ Iris Setosa، Iris Virginica و Iris Versicolor. هر گونه دارای ۵۰ نمونه است که برایشان چهار ویژگی طول و عرض کاسبرگ‌ها و طول و عرض گلبرگ‌ها اندازه‌گیری شده است.

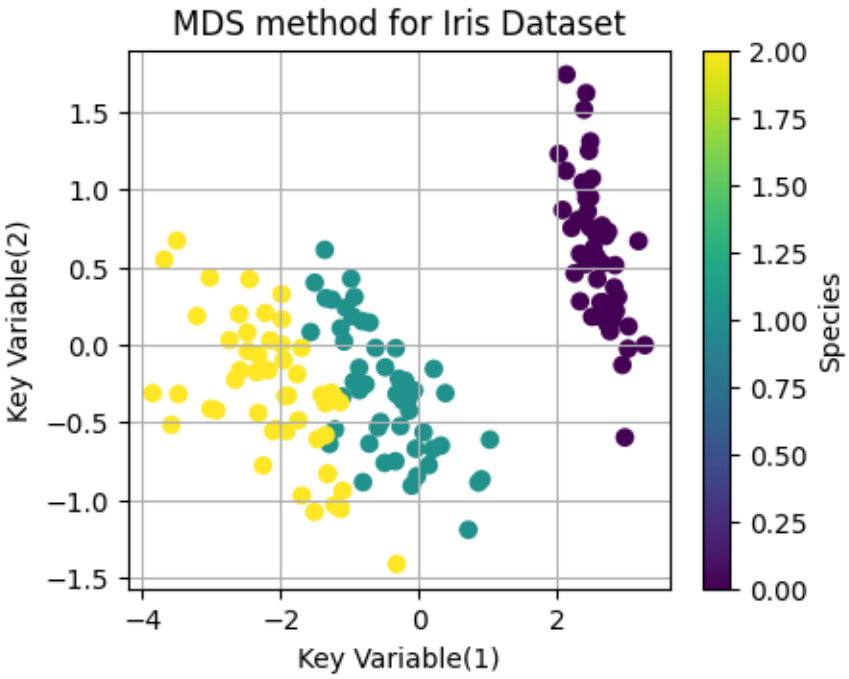


2.1 Principal Component Analysis (PCA)

در این روش داده‌ها به مجموعهٔ جدیدی از متغیرها یا مؤلفه‌های اصلی که ترکیبی خطی از متغیرهای اصلی هستند تبدیل می‌شوند. این روش با انتخاب چند متغیر اساسی، بیشترین واریانس را برای داده‌ها حفظ می‌کند.

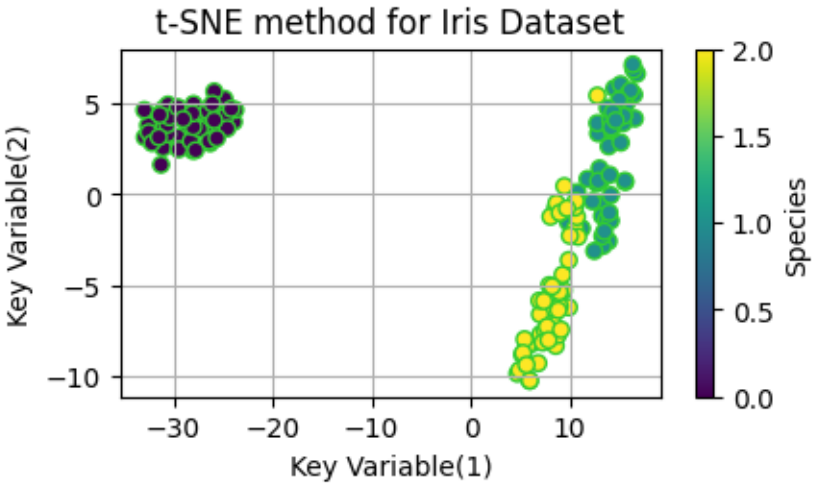
2.2 Multidimensional Scaling (MDS)

این روش برای تشخیص شباهت‌ها و یا تفاوت‌های مجموعه‌ای از داده‌ها استفاده می‌شود. در این روش این شباهت و یا تفاوت با نمایش دادن data set مورد نظر در ابعاد کمتر صورت می‌گیرد و فاصلهٔ بین داده‌ها را بر اساس همان شباهت‌ها و تفاوت‌های جفتی‌شان نمایش می‌دهد.



2.3 T-Distributed Stochastic Neighbor Embedding (t-SNE)

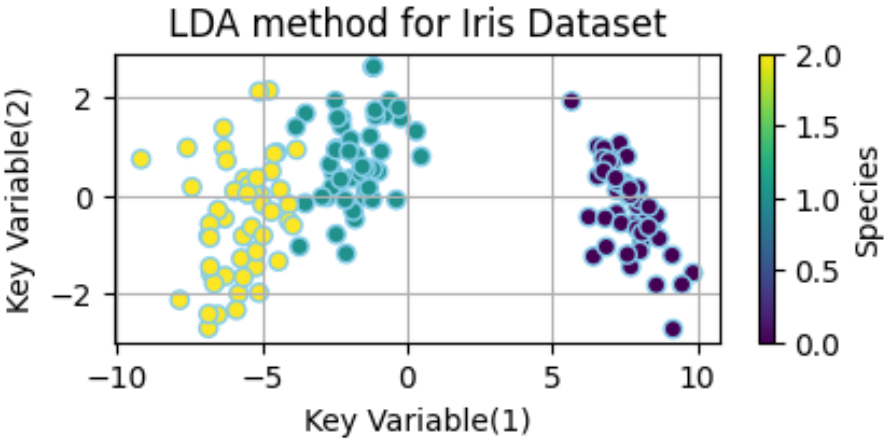
این روش روشی غیر خطی است که برای visualize کردن داده‌های پیچیده و با ابعاد بالا استفاده می‌شود. در این روش داده‌ها در عین کاهش ابعادشان، روابط بینشان حفظ می‌شود. البته از این روش معمولاً برای کاهش ابعاد داده‌های ۲ یا ۳ بعدی استفاده می‌شود.



2.4 Linear Discriminant Analysis (LDA)

این روش که در یادگیری ماشین نظارت‌شده (Supervised Machine Learning) برای حل و مدلسازی آماری مسائل Classification برای مجموعه داده‌هایی که دارای چند Class هستند استفاده می‌شود. در این

روش چندین Class که دارای ویژگی‌های مختلفی هستند، از طریق کاهش ابعاد طبقه‌بندی و جدا می‌شوند. این روش در بهینه‌سازی مدل‌های Machine Learning مؤثر است.



References

[1] https://www.sciencedirect.com/

[2] https://en.wikipedia.org/

[3] https://www.displayr.com/

[4] https://www.ncrm.ac.uk/

[5] https://www.ibm.com/

[6] https://www.geeksforgeeks.org/

[7] https://encord.com/

[8] https://online.stat.psu.edu/