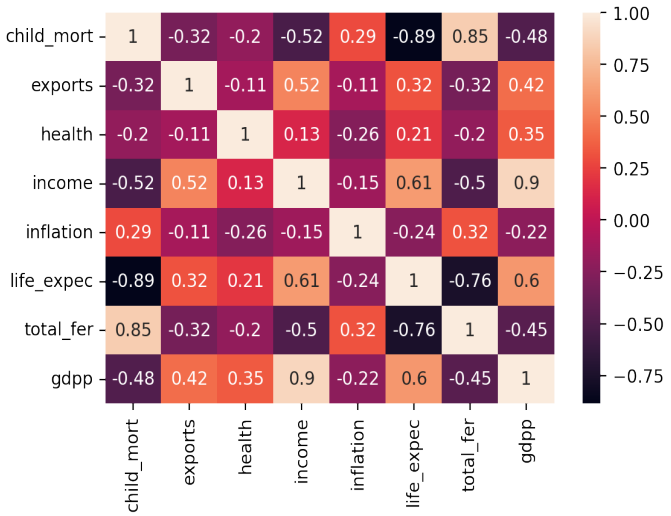
گزارش کار تمرین چهارم درس یادگیری ماشین

محمد لشکری ۴۰۰۱۱۲۰۸۷

۱۳ دی ۱۴۰۰

به دلیل یکتا بودن ویژگی country و عدم تأثیر آن روی خروجی از دادگان حذف شد. همانطور که در شکل ۱ که نمایانگر ماتریس همبستگی است مشاهده می‌شود، ویژگی imports ارتباط کمی با متغیرهای مهمی مثل health دارد. پس این ستون نیز حذف و دادگان باقی مانده برای خوشه بندی استفاده شد که آن را X نامیدیم.



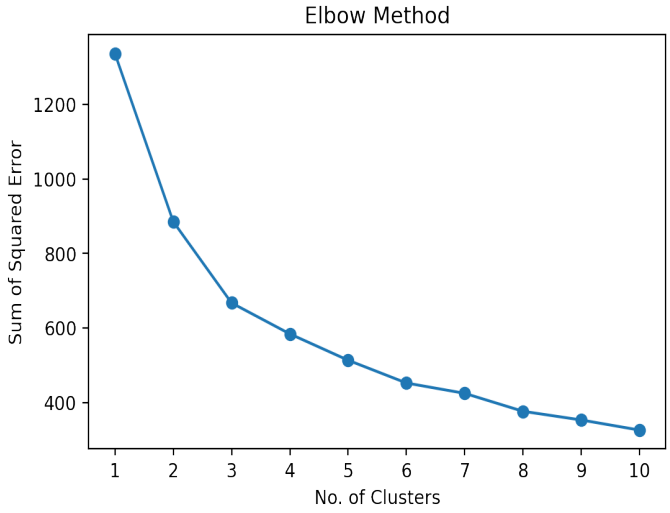
شکل ۱: ماتریس همبستگی

۱ پیش پردازش داده‌ها

برای اینکه تأثیر همه ویژگی‌ها روی معیار اقلیدسی یکسان باشد آن‌ها را به صورت زیر نرمال کرده‌ایم:

$$X := \frac{X - \mu}{\sigma}$$

که در آن μ و σ به ترتیب میانگین و واریانس ویژگی‌های X هستند.



شکل ۲: روش Elbow

۲ خوشه بندی

۱.۲ روش Elbow

روش Elbow یکی از روش‌های پیدا کردن تعداد بهینه خوشه‌هاست. این روش به میزان زیاد و کم شدن تغییرات فاصله اقلیدسی هر نقطه از مرکز خوشه، با افزایش تعداد خوشه‌ها، حساس است. طبق شکل ۲ تعداد ۴ خوشه بهینه است.

۲.۲ تحلیل سایه [1]

این تحلیل بر مبنای فاصله اقلیدسی است و مشخص می‌کند که نمونه داخل خوشه درست افتاده است یا خیر. برای اینکار فاصله هر داده با اعضای خوشه خود را محاسبه می‌کنیم، سپس میانگین می‌گیریم. فاصله نزدیک ترین خوشه به آن نمونه (به جز خوشه ای که در آن قرار دارد) را به همین شکل محاسبه می‌کنیم. فرض کنیم عدد اول a و عدد دوم b باشد. برای نمونه $\mathbf{x}_i \in X$ داریم:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

اگر مقدار فوق به ۱ نزدیک باشد، داده در خوشه درست است و اگر به -۱ نزدیک باشد، نمونه در خوشه اشتباه است. درنهایت با میانگین گیری از s_i ها معیار نهایی برای ارزیابی خوشه بندی، به دست می‌آید.

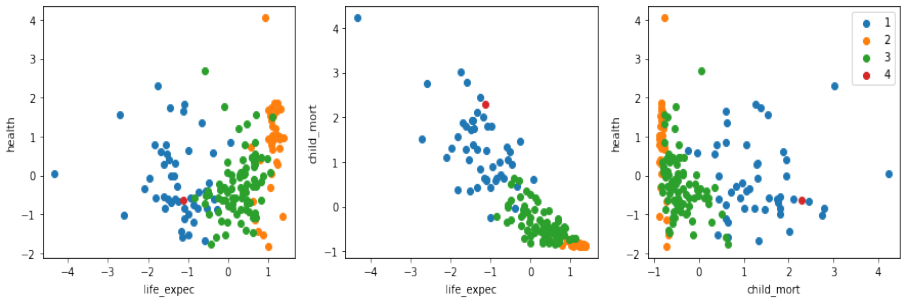
$$SS = \frac{۱}{n} \sum_{i=۰}^{n} s_i$$

[1] Silhouette Analysis

که در آن SS همان امتیاز سایه [۲] است. برای دیتاست مورد نظر $SS = ۰/۳۲$ که از ۱ فاصله دارد اما فاصله آن از -۱ بیشتر است، پس قابل قبول است.

۳.۲ دیداری سازی

ویژگی‌های life_expec, health, child_mort برای دیداری سازی انتخاب شده‌اند که نتایج آن در شکل ۳ قابل مشاهده است.هر نقطه مشخص کننده یک کشور و رنگ آن، نشان‌دهنده خوشه کشور است.لازم به ذکر است ویژگی‌های life_expec, child_mort رابطه عکس دارند. کشوری که در آن وضعیت سلامتی بدتر، نرخ مرگ کودکان زیر ۵ سال بالاتر و نرخ سال‌های زنده ماندن کودکان پایین تر است، وضعیت بدی دارد. پس کشورهای نارنجی رنگ در شکل ۳ نیاز به کمک دارند.



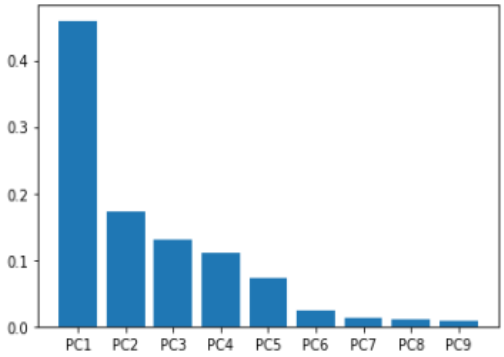
شکل ۳: دیداری سازی

۳ تحلیل مؤلفه‌های اصلی

۱.۳ تحلیل واریانس مؤلفه‌ها براساس نمودار توضیح درصد واریانس

در شکل ۴ نمودار توضیح درصد واریانس برای هر مؤلفه مشخص است. با استفاده از ۵ مؤلفه اول، واریانس داده‌ها به خوبی حفظ می‌شود. پس ۵ مولفه اول استخراج شد و بعد مجموعه داده‌ها به $(۱۶۷, ۵)$ تغییر کرد.

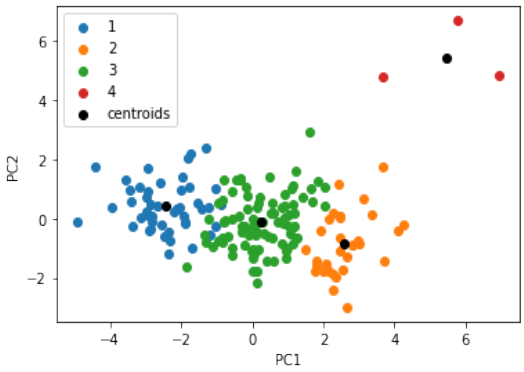
[2] Silhouette Score



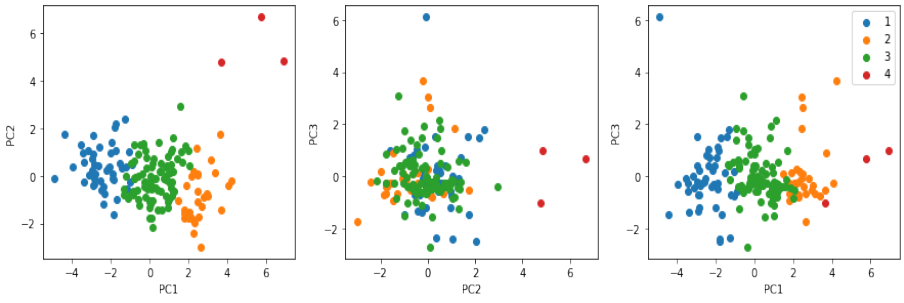
شکل ۴: نمودار توضیح درصد واریانس

۲.۳ خوشه بندی بعد از تحلیل مؤلفه‌ها

در شکل ۵ نتایج حاصل از اجرای الگوریتم KMeans با تعداد ۴ خوشه که بر اساس مؤلفه‌های PC1, PC2 رسم شده‌اند، قابل مشاهده است. همانطورکه مشخص است در مقایسه با نتایج بخش ۳.۲ داده‌ها بهتر خوشه‌بندی شده‌اند و اختلاف هر داده تا مرکز خوشه‌های دیگر بیشتر است.



شکل ۵: خوشه بندی بعد از PCA


PC2
PC1
PC3
PC2
PC3
PC1
1
2
3
4