تمرین هفتم درس یادگیری ماشین
دکتر باباعلی

سیدعلیرضا مولوی

فهرست مطالب

فهرست تصاویر

فهرست جداول

۱ داده ها

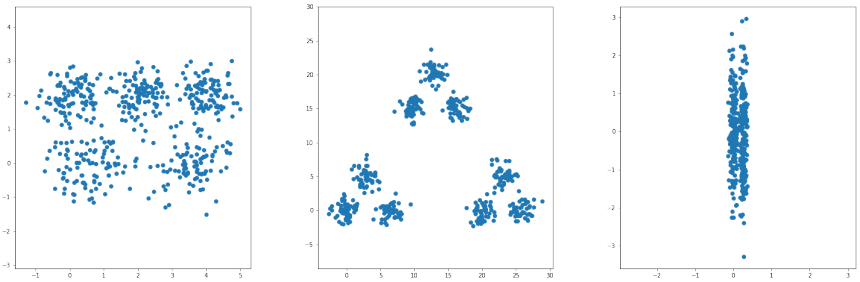
در تمرین ما ۳ مجموعه داده داریم. در جدول ۱ تعداد و بعد داده ها را مشخص کرده ایم. در شکل ۱ شکل داده ها را نمایش داده ایم.

جدول ۱: تعداد و بعد داده ها

نام	تعداد	بعد
مجموعه داده اول	۴۰۰	۲
مجموعه داده دوم	۴۵۰	۲
مجموعه داده سوم	۵۰۰	۲

شکل ۱: مجموعه داده ها

(آ) مجموعه داده اول (ب) مجموعه داده دوم (ج) مجموعه داده سوم



۱.۱ پیش پردازش

به دلیل اینکه در فرآیند خوشه بندی KMeans از معیار فاصله استفاده می شود، داده ها باید استاندارد سازی [۱] و یا نرمال سازی [۲] شوند.

$$X_{standard} = \frac{X - \mu_X}{\sigma_X} \quad (۱)$$

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (۲)$$

البته به دلیل اینکه در دستور عمل های تمارین گفته شده است که بر روی داده خام کار کنیم، بنابراین پیش پردازش را انجام نمی دهیم. (البته در تمرین F صریحا از ما خواسته شده است که داده ها را استاندارد سازی کنیم.)

[۱] Standarize
[۲] Normalize

۲ پیاده سازی

به دلیل اینکه توضیحات کامل پیاده سازی در صورت سوالات در تمارین فایل ژوپیتر قرار دارد، برای جلوگیری از تکرار مطالب، الگوریتم از اول توضیح داده نمی شود و فقط جزییات خاص مربوط به پیاده سازی ذکر می شود.

۱.۲ الگوریتم استاندارد سازی

برای پایداری عددی، ما تغییر زیر را در الگوریتم به وجود می آوریم:

$$X_{standard} = \frac{X - \mu_X}{\sqrt{\sigma^2 + \epsilon}}$$

که ϵ مقدار کوچکی است، که مانع از صفر شدن کسر می شود.
الگوریتم در HW7.scaler.StandardScaler پیاده سازی شده است.

۲.۲ مقدار دهی اولیه centroid ها

در تمرین خواسته شده است که پیاده سازی تصادفی و KMeans++ را انجام دهیم. پیاده سازی در فایل HW7.initializer.py قرار دارد.
در مدل های LLoyd Algorithm و Kmeans، مقدار دهی اولیه centriod ها از طریق پارامتر init مشخص می شود.

۳.۲ الگوریتم LLoyd Algorithm

الگوریتم LLoyd Algorithm مشابه الگوریتم خواسته شده در تمرین فایل ژوپیتر پیاده سازی شده است.

شرط همگرایی: اگر در تکرار الگوریتم، تابع خطا به میزان بسیار کمی (توسط پارامتر tol این حد مشخص می شود) تغییر کند، ما فرض می گیریم که الگوریتم همگرا شده است و فرآیند را متوقف می کنیم؛ اگر به همگرایی نرسیم الگوریتم به اندازه تعداد تکرار های مشخص شده، اجرا می شود. (حداکثر تعداد تکرار ها از طریق پارامتر n_iter مشخص می شود)
الگوریتم LLoyd Algorithm در HW7.kmeans.LLoydAlgorithm پیاده سازی شده است.

۴.۲ مدل KMeans

مدل KMeans بر پایه الگوریتم و کد LLoydAlgorithm پیاده شده است، در واقع تعدادی (تعداد توسط پارامتر n_init مشخص می شود) الگوریتم LLoydAlgorithm را اجرا می کند و بهترین آن ها را به عنوان مدل نهایی انتخاب می کند.
مدل از کتابخانه joblib برای موازی سازی اجرای الگوریتم های LLoyd استفاده می کند.
الگوریتم در HW7.kmeans.KMeans پیاده سازی شده است.

۳ تمارین

۱.۳ تمرین A و B

در تمرین A خواسته شده است که ۲۰۰ بار الگوریتم LLoyd با مقدار دهی اولیه تصادفی، تا همگرا شدن اجرا بر روی مجموعه داده اول (شکل ۱آ) کنیم و نتایج را بررسی کنیم؛ تمرین B مشابه تمرین A ولی با مقدار دهی KMeans++ فرآیند را انجام باید دهیم. در شکل ۲ بهترین مدل برای هر دو حالت را مشاهده می کنیم؛ در واقع بهترین مدل برای هر دو حالت مشابه است.

بررسی تعداد تکرار های لازم تا همگرایی: به جدول ۲ توجه کنید. میانگین تعداد تکرار های در حالتی که مقدار دهی اولیه KMeans++ باشد، تقریبا ۱۰٪ کمتر حالت مقدار دهی اولیه تصادفی است. بنابراین روش KMeans++ به صورت قابل توجهی کارایی بهتری دارد. (شرط همگرایی الگوریتم در بخش ۳.۲ گفته شده است.)

centroid های تولید شده در مقدار دهی اولیه: در شکل ۳ centroid تولید شده برای هر دو نوع مقدار دهی اولیه نمایش داده شده است. به دلیل کوچک بودن مجموعه داده، چندان تفاوت محسوسی وجود ندارد. (البته در بررسی تعداد تکرار های لازم تا همگرایی، برتری مقدار دهی KMeans++ بر روش تصادفی محسوس تر است)

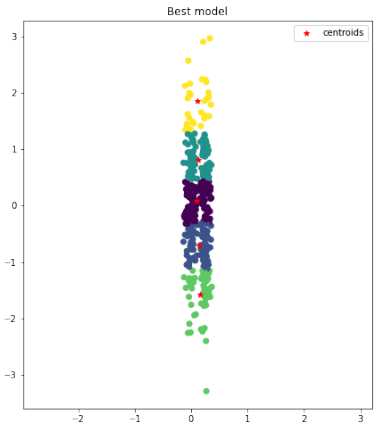
بررسی فاصله نقاط از مرکز خوشه: در شکل ۴ مشخصات مربوط فاصله نقاط از مرکز خوشه ها برای هر اجرای مدل آورده شده است. در حالتی که از KMeans++ برای مقدار دهی اولیه استفاده کردیم میاگین فاصله کمتر و پایدار تر (خط مربوط به میانگین ملایم تر است) است.

جدول ۲: تعداد تکرار های لازم تا همگرایی با توجه به نوع مقدار دهی اولیه

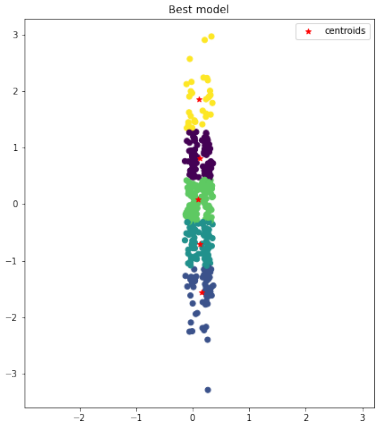
نوع مقدار دهی اولیه	میانگین تعداد تکرار تا همگرایی	انحراف از معیار تعداد تکرار تا همگرایی	بیشینه تعداد تکرار تا همگرایی	کمینه تعداد تکرار تا همگرایی
تصادفی	۱۱٫۰۳۵۰	۴٫۴۰۵۰	۳۰	۳
KMeans++	۹٫۸۹۵۰	۴٫۱۴۶۶	۲۲	۳

شکل ۲: بهترین مدل نهایی بر روی مجموعه داده یک

(آ) بهترین مدل نهایی با مقداردهی تصادفی

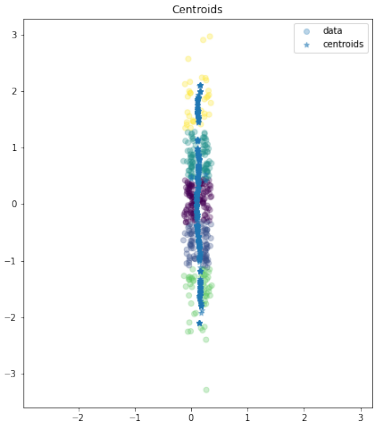


(ب) بهترین مدل نهایی با مقداردهی KMeans++

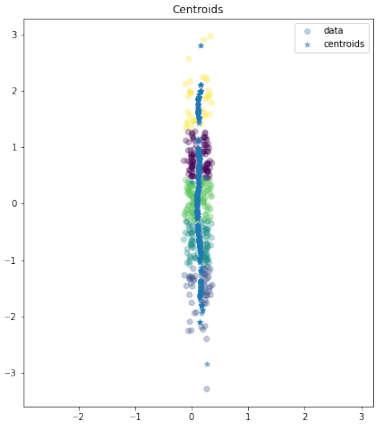


شکل ۳: نتیجه مقدار دهی اولیه centroid ها بر روی مجموعه داده یک

(آ) centroid ها با مقداردهی تصادفی

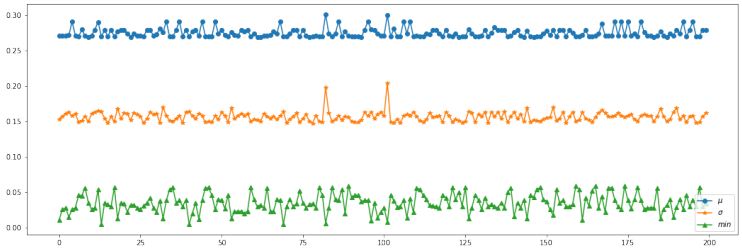


(ب) centroid ها با مقداردهی KMeans++

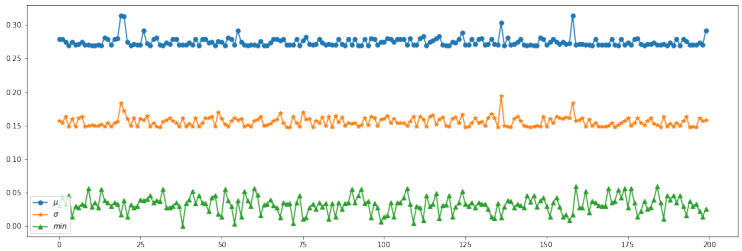


شکل ۴: میانگین، انحراف از معیار و کمینه فاصله نقاط از مرکز خوشه ها در هر تکرار، بر روی مجموعه داده یک

(آ) با مقدار دهی اولیه تصادفی



(ب) با مقدار دهی اولیه KMeans++

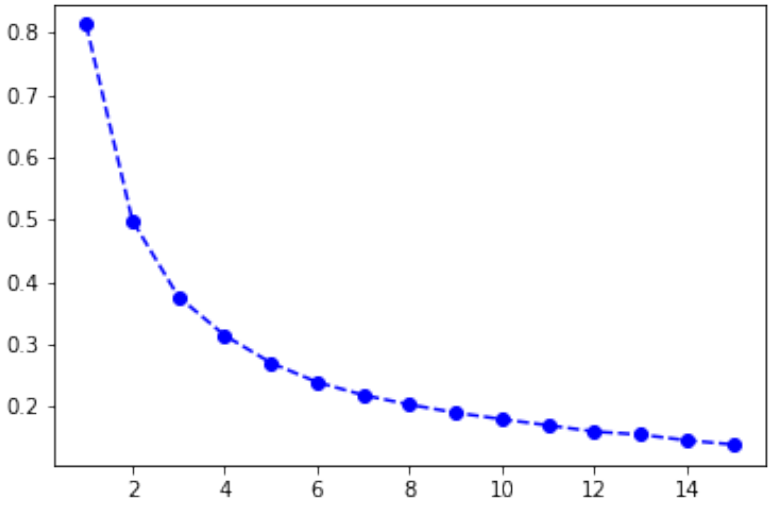


۲.۳ تمرین C

در تمرین C خواسته شده است که مدل KMeans را بر روی مجموعه داده یکم (شکل ۱آ)، برای مقادیر $K = [۱, ۲, ۳, \ldots, ۱۵]$ اجرا کنیم (در تمرین گفته شده برای هر مقدار K مدل را ۲۰۰ بار اجرا کنیم و بهترین را انتخاب کنیم، در حالیکه من از ۱۰ اجرا استفاده کرده ام زیرا در ۱۰ اجرا به همگرایی می رسم) و نمودار خطا را به ازای هر مقدار K رسم کنیم. نتیجه در شکل ۵ نمایش داده شده است.

از نمودار خطا برای یافتن مقدار مناسب برای ابرپارامتر K استفاده می شود؛ و مقادیر مناسب معمولا نقاط شکستگی[۳] نمودار خطا اند. با توجه به شکل ۵ نقطه شکستگی واضحی برای مجموعه داده یکم، وجود ندارد.

شکل ۵: خطای روش KMeans(++) برای مجموعه داده یک به ازای مقادیر مختلف K

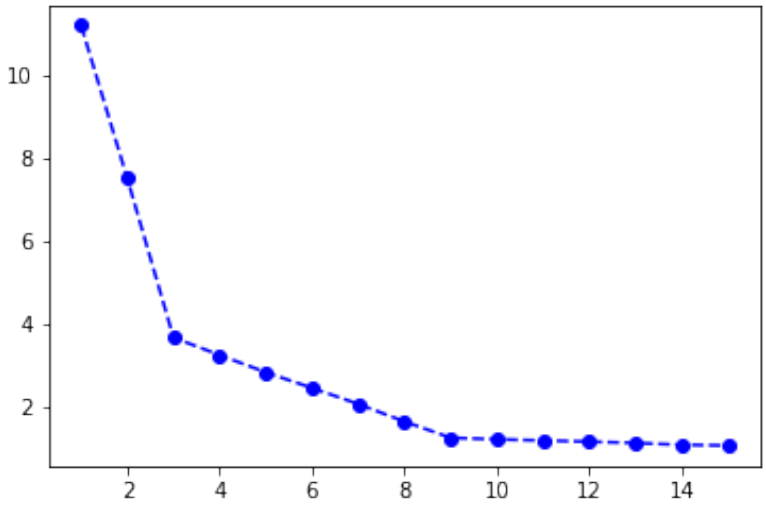


[۳] knee یا elbow نیز گفته می شود.

۳.۳ تمرین D

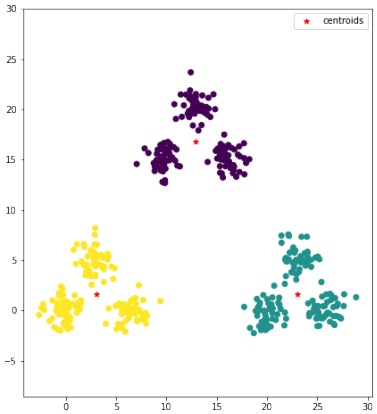
تمرین D مشابه تمرین C است ولی گفته شده است که آزمایشات بر روی مجموعه داده دوم (شکل ۱ب) انجام شود.
با توجه به شکل ۶ ما ۲ شکستگی در نقاط ۳ و ۹ داریم و این مقادیر کاندیدای مناسبی برای مقدار K هستند.
در شکل ۷ ما مدل را با مقدار $K = ۳$ و $K = ۹$ آموزش داده ایم. با اینکه مدل $K = ۹$ به نظر می آید مدل درست تر است ولی مقدار $K = ۳$ نیز قانع کننده است.

شکل ۶: خطای روش KMeans(++) برای مجموعه داده دوم به ازای مقادیر مختلف K

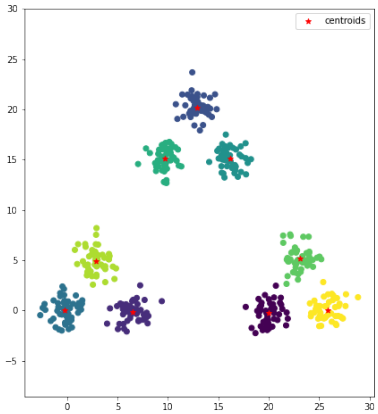


شکل ۷: خوشه بندی بر روی مجموعه داده دوم

(آ) خوشه بندی با مقدار $K = ۳$



(ب) خوشه بندی با مقدار $K = ۹$



۴.۳ تمرین E و F

هدف از این تمرین نشان دادن تاثیر پیش پردازش بر روی نتیجه نهایی مدل KMeans است. در این تمرین از مجموعه داده سوم (شکل ۱ج) استفاده شده است. با اینکه با توجه به شکل مجموعه داده سوم، به نظر میرسد داده از ۵ خوشه تشکیل شده است، ولی در تمرین گفته شده است که $K = ۲$ قرار دهید.
در بخش ۱.۱ در رابطه با پیش پردازش توضیح داده شده است.

بررسی نتایج: در تمرین از ما خواسته شده است داده را استاندارد سازی کنیم و نتایج را بررسی کنیم. در شکل ۸ نتایج را نشان داده ایم.

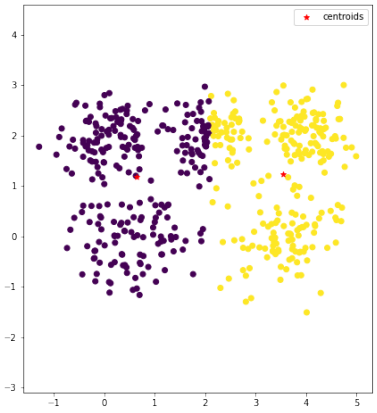
بدون پیش پردازش: شکل ۸آ حالتی است که از داده های خام برای پردازش استفاده می کنیم؛ در این حالت کره‌ی وسط به ۲ قسمت تقسیم شده است و هر قسمت به یک خوشه تخصیص داده شده است؛ که مشخصا این خوشه بندی نامناسب است.

استفاده از پیش پردازش: شکل ۸ب حالتی را نشان می دهد که پیش از آموزش مدل، داده را استاندارد سازی (به بخش ۱.۱ رجوع شود) می کنیم. در این حالت به نظر میرسد که حداقل خوشه ها به طور بهتری تقسیم بندی شده اند.

در واقع نمی توان در این گونه مواقع مدل ها را بررسی کرد مگر اینکه اطلاعات بیشتری در دسترس داشته باشیم، ولی همواره توصیه می شود که پیش از استفاده از مدل هایی که از معیار فاصله استفاده می کنند، داده را استاندارد سازی کنیم.

شکل ۸: خوشه بندی بر روی مجموعه داده سوم

(آ) خوشه بندی بدون پیش پردازش داده



(ب) خوشه بندی بر روی داده استاندارد سازی شده