تمرین سری سوم درس داده کاوی

۱. قسمت اول – MLP

- **دادگان قسمت اول :**

  - تصاویری از افراد مختلف در این مجموعه داده موجود است . از هر فرد حداقل دو تصویر وجود دارد .
  - در این تمرین فقط افرادی در نظر گرفته شده اند که حداقل ۱۵۰ تصویر از آن ها در مجموعه داده موجود باشد که تعداد آن ها برابر ۲ نفر می باشد.
  - با استفاده از دستور زیر میتوانید داده ها را به صورت مستقیم دانلود و در کد خود استفاده کنید.(میتوانید از colaboratory google و یا Jupyter notbook استفاده کنید)

```
from sklearn.datasets import fetch_lfw_people
lfw_people = fetch_lfw_people(min_faces_per_person=150, resize=0.4)
X = lfw_people.data
y = lfw_people.target
```

- **شرح مسئله :**
  - پس از load کردن دادگان ۲۵ درصدآن را به عنوان test setو باقی را به عنوان train set در نظر بگیرید
  - با استفاده از PCA تعداد Feature ها را کاهش دهید
  - بوسیله ی شبکه عصبی Multi-layer Perceptron داده ها را دسته بندی کرده و مدل خود را ارزیابی کنید
  - Confusion matrix را ارائه داده و نتایج را با توجه به آن تحلیل کنید

- **لینک های کمکی :**

what is Google colab?

PCA

Multi-layer Perceptron

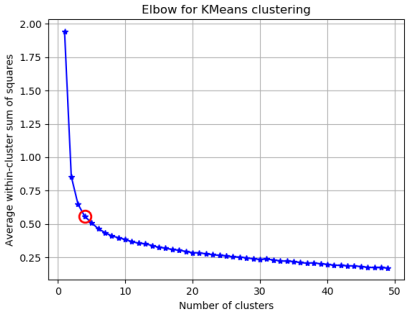
Recommended video

۲. قسمت دوم – clustering

- **دادگان قسمت دوم :**
  - در این قسمت از تمرین مجموعه دادگانی برای انجام عمل خوشه بندی به شما داده شده است. این مجموعه شامل ۳ دادگان است که هر کدام شامل ۲ ویژگی هستند.
  - از این لینک میتوانید مجموعه دادگان (dataset) را دانلود کنید.
- **شرح قسمت دوم :**
  - ابتدا هر ۳ دادگان موجود را لود کنید. سپس هر یک از نقاط این دادگان ها را در نمودار دوبعدی نمایش دهید (۳ نمودار ۲ بعدی باید رسم کنید) و تحلیل کنید که طبق تعاریف مختلف clustering نظیر تعریف مبتنی بر density ، انتظار داریم در هر مجموعه داده چطور خوشه بندی انجام شود و مشخص کنید که به ازای چند کلاستر بهترین خوشه بندی را برای هر کدام از دادگان ها خواهیم داشت.

  - با استفاده از الگوریتم Lloyd (k-means) عمل خوشه بندی را انجام دهید. تمامی K مرکز را به صورت رندوم از $x_1 . \dots . x_n$ (مجموعه داده ها) انتخاب کنید. تمامی K مرکز بدست آمده را در کنار داده های اصلی نمایش دهید. در این قسمت باید برای هر کدام از ۳ دادگان، مقدار K مناسب پیشنهاد دهید و با Kی پیشنهادی عمل خوشه بندی توسط kmeans را انجام دهید و خطای SSE را در هر مورد گزارش کنید.

  - یکی از راه های پیدا کردن Kی مناسب برای عمل خوشه بندی رسم نمودار خطی خطا-K (خطا در محور عمودی و K در محور افقی) است. برای این منظور ابتدا برای هر دادگان به ازای $K = 1.2. \dots . 15$ ، حداقل ۲۰۰ بار اعضای centroid ها را به صورت رندوم انتخاب کنید و سپس مقدار خطا متناظر با بهترین مقدار اولیه را به ازای تمامی K های ذکر شده محاسبه کنید و نمودار خطا-K را رسم کنید. با استفاده از تحلیل این نمودار، مشخص کنید که در هر دادگان در چه Kای نقطه ی زانویی "knee" رخ میدهد. نقطه زانویی نقطه ای است که در آن به یکباره شیب کاهش خطا کم شود و شکستگی ایجاد شود. مثالی از knee point یا elbow point در تصویر زیر آمده است:



توجه : لزومی ندارد برای هر دادگان فقط یک نقطه زانویی داشته باشیم و ممکن است تعداد نقاط زانویی بیشتر باشد. در صورت وجود چند نقطه زانویی مقدار K مربوطه را با رسم نمودار ذکر کنید.

  - با استفاده از الگوریتم fuzzy c-means ، همانند مرحله قبل داده ها را به K خوشه تقسیم کنید. مقادیر K را مانند قسمت قبل در نظر بگیرید. سپس داده هایی که در بیش از دو خوشه قرار میگیرند را در نمودار با رنگ متفاوت مشخص کنید.

- با استفاده از الگوریتم DBSCAN، عمل خوشه بندی را برای هر ۳ دادگان انجام دهید. نتیجه clustering را با نمودار نمایش دهید (داده های موجود در خوشه های یکسان را با رنگ مشابه نمایش دهید).
- نتایج ۳ الگوریتم را از نظر اندازه خطا باهم مقایسه کنید. کدام الگوریتم برای خوشه بندی این ۳ دادگان بهتر کار میکند؟ برای عادلانه تر بودن شرایط مقایسه برای هر دادگان هر الگوریتم را ۲۰۰ بار اجرا کنید و بهترین نتیجه را در جدول برای هر دادگان و هر الگوریتم ذکر کنید.

نحوه ارسال فایل های تمرین :

- کد های خود را به همراه فایل مربوط به توضیحات به صورت یک فایل فشرده در سامانه ایلرن ثبت کنید.
- ترجیحا فایل کد ارسالی به فرمت ipynb. باشد.
- نام فایلی که آپلود می کنید شامل نام نام خانوادگی و شماره دانشجویی باشد . مثال :

DM – HW3 - your name - your student id

پیشنهادات و نکات مربوط به انجام تمرین :

- برای انجام قسمت مربوط به نوشتن کد سعی کنید از jupyter notebook استفاده کنید.
- استفاده از پکیج ها و کتابخانه های آماده بلامانع است.
- برای خواندن داده ها میتوانید از دستور pandas.read_csv() استفاده نمایید.

موفق باشید

فروردین ۱۴۰۱