تمرین دوم

هدف: آشنایی با شبکه‌های خودسازمانده کوهونن

کد: کد این فعالیت را به زبان پایتون و با استفاده از پلتفرم گوگل کولب بنویسید.

گزارش: ملاک اصلی انجام فعالیت گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این فعالیت یک فایل گزارش در قالب pdf تهیه کنید و در آن برای هر سوال، تصاویر ورودی، تصاویر خروجی و توضیحات مربوط به آن را ذکر کنید. سعی کنید توضیحات کامل و جامعی تهیه کنید.

تذکر ۱: مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیدا برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری بلامانع است، اما کپی کردن غیرمجاز است.

تذکر ۲: مجموعه‌های داده مورد استفاده را به جز در مواردی که صریحا در صورت سوال ذکر شده باشد، حتما قبل از استفاده بصورت تصادفی به سه بخش آموزش (۷۰ درصد داده‌ها)، آزمون (۲۰ درصد داده‌ها) و اعتبارسنجی (۱۰ درصد داده‌ها) تقسیم نمایید.

تذکر ۳: مدل‌های تخمین‌گر را بر اساس معیار میانگین مربعات خطا ارزیابی نمایید.

راهنمایی: در صورت نیاز میتوانید سوالات خود را در خصوص پروژه از تدریس‌یار درس، از طریق ایمیل زیر بپرسید.

E-mail: ann.ceit.aut@gmail.com

ارسال: فایل های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID_HW۰۲.zip تا تاریخ ۱۴۰۰/۰۸/۲۸ ارسال نمایید. شایان ذکراست هر روز تاخیر باعث کسر ۱۰٪ نمره خواهد شد.

در این تمرین قصد داریم با استفاده از شبکه خودسازمانده کوهونن، یک مجموعه‌داده با ابعاد بالا را بصری‌سازی کرده و یک نگاشت مناسب برای کاهش بعد آن مجموعه‌داده ارائه دهیم. مجموعه‌داده مورد استفاده در این تمرین، مجموعه‌داده RCV است که از طریق پکیج sklearn در دسترس می‌باشد. برای بارگذاری این مجموعه‌داده می‌توانید از دستورات زیر در پایتون استفاده نمایید.

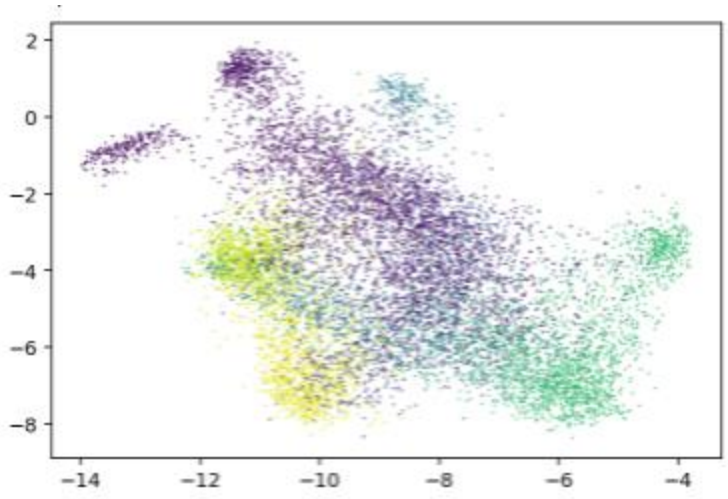
```
from sklearn.datasets import fetch_rcv1
rcv1 = fetch_rcv1()
```

برای انجام این تمرین به سوالات زیر پاسخ دهید.

۱- در فایل گزارش خود به طور کامل فرایندهای کاهش بعد و بصری‌سازی مجموعه‌داده‌های با ابعاد بالا توسط شبکه خودسازمانده کوهونن را توضیح دهید.

۲- مجموعه‌داده مورد استفاده را بارگذاری نمایید. با توجه به اینکه هدف این تمرین، دسته‌بندی داده‌ها نیست، نیازی به تقسیم مجموعه‌داده به سه دسته آموزشی، آزمون و اعتبارسنجی وجود ندارد.

۳- یک مدل خودسازمانده کوهونن ایجاد کنید. این مدل را با استفاده از داده‌های موجود آموزش دهید. یک نقشه ویژگی از داده‌های موجود بدست بیاورید و در فایل گزارش رسم کنید. شکل ۱، نمونه‌ای از نقشه‌های موجود روی همین مجموعه‌داده را نمایش می‌دهد. توجه نمایید هر نقشه‌ای که بتواند به شما تحلیل مناسبی از داده‌ها بدهد قابل قبول است.



شکل ۱ نمونه‌ای از نقشه‌های ویژگی روی مجموعه‌داده مورد استفاده

٤- با توجه به نقشه ویژگی استخراج شده در سوال قبل، توزیع داده‌های موجود در مجموعه‌داده را بطور کامل از لحاظ جداپذیری، میزان پیچیدگی مجموعه‌داده، میزان نویزی بودن و موارد مشابه دیگر، تحلیل کنید.

٥- با استفاده از شبکه آموزش دیده، داده‌های موجود را خوشه‌بندی کرده و عملکرد شبکه خودسازمانده کوهونن را در خوشه‌بندی این داده‌ها ارزیابی نمایید. برای ارزیابی خوشه‌بندی می‌توانید از معیار خلوص[۱] استفاده نمایید.

توجه نمایید، برای محاسبه میزان خلوص خوشه‌بندی می‌توانید طبق رابطه زیر عمل کنید. در این رابطه،N تعداد کل داده‌های موجود، M تعداد خوشه‌های تولید شده و D تعداد کل کلاس‌های موجود در مجموعه‌داده را نمایش می‌دهد. این رابطه، در هر خوشه کلاس غالب را می‌یابد. کلاس غالب، کلاسی است که تعداد رکوردهای موجود از آن کلاس در خوشه، بیشتر از بقیه کلاس‌هاست. سپس تعداد رکوردهای کلاس‌های غالب هر خوشه را با هم جمع کرده و بر تعداد کل رکوردها تقسیم می‌نماید. در صورتی که خوشه‌بندی به نحوی انجام شود که تمام رکوردهای موجود در یک خوشه، از یک کلاس باشند، مقدار این معیار به ۱ می‌رسد.

$$\text{purity} = \frac{1}{N}\Sigma_{\text{m}\in\text{M}} \max_{\text{d}\in\text{D}} |m \cap d|$$

موفق باشید

[۱] Purity