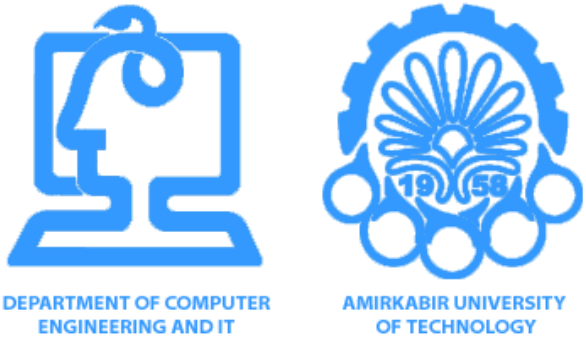

دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر

تمرین چهارم درس یادگیری ماشین

دکتر ناظرفرد

پاییز ۹۸

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت زیپ ذخیره کرده و با عنوان #studentId_HW.zip بارگزاری نمایید.
- مهلت انجام این تمرین تا ساعت **۲۳:۵۵ روز ۱۳ بهمن** می‌باشد **و به هیچ وجه تمدید نمی‌شود**.

تمرین تشریحی: (۱۰٪)

۱- صحت هر یک از موارد زیر را بررسی کرده و دلایل خود را توضیح دهید.

- ماشین‌های بردار پشتیبان[1] پارامتریک‌اند.
- مقدار حاشیه بدست آمده برای دو ماشین بردار پشتیبان با کرنل‌های متفاوت که برای داده‌های یکسان آموزش دیده‌اند، می‌تواند معیاری برای میزان کارایی مدل باشد.
- ماشین‌های بردار پشتیبان همواره در برابر بیش برازش مقاوم می‌باشند.
- وجود داده‌های پرت و نویز بر روی ماشین‌های بردار پشتیبان بی تاثیر است.

[1] Support Vector Machine (SVM)

در انجام تمرینات زیر به نکات زیر توجه نمایید:

- تمرین بدون گزارش فاقد ارزش می‌باشد و نمره‌ای به آن تعلق **نمی‌یابد**.
- کامنت گذاری کدها در حد **لازم و کافی** الزامی می‌باشد.
- گذاشتن عنوان[۲] برای نمودارها و برچسب‌گذاری[۳] محورهای نمودار **الزامی** می‌باشد.
- در پیاده‌سازی بخش‌های مختلف، امکان استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را به طور کلی ندارید. **موارد مجاز در صورت سوال بخش‌ها ذکر شده است.**
- برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. هم‌چنین برای خواندن داده‌ها به عنوان ورودی می‌توانید از pandas استفاده کنید.
- برای بهبود سرعت برنامه توصیه می‌شود که تا حد ممکن از عملیات ماتریسی استفاده شود.
- برای محاسبه معیارهای ارزیابی مانند دقت و ماتریس پریشانی می‌توانید از کتابخانه استفاده نمایید. **درصورت پیاده‌سازی آن‌ها نمره تشویقی تعلق خواهد گرفت. (۵+)**
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید. **در این راستا برای گزارش‌هایی با ظاهر عالی نمره تشویقی در نظر گرفته شده است. (۵+)**
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره هر **دو طرف صفر** در نظر گرفته می‌شود.

تمرینات برنامه‌نویسی: (۱۰+٪۹۰)

۲- پیاده سازی ماشین بردار پشتیبان

الف) دیتاست lsvt-voice-rehabilitation را از آدرس زیر[۴] دانلود کرده و به سوالات زیر پاسخ دهید. برای ارزیابی از 10-fold cross validation استفاده کنید. در این تمرین مجاز به استفاده از کتابخانه می‌باشید. داده‌ها را به روش‌های زیر دسته‌بندی کنید و به سوالات زیر پاسخ دهید:

- کرنل خطی
- کرنل چند جمله‌ای (پارامترهای d,r)
- RFB (پارمتر گاما)
- سیگموئید (پارمتر r)

ب) معیار دقت[5] و F1 را برای هریک از دسته‌بندهای بالا بدست آورید. (برای هر یک از پارامترهای یاد شده، حداقل سه مقدار متفاوت در نظر بگیرید)
ج) تاثیر پارامترهای هر کرنل بر کارایی مدل‌ها را تحلیل کنید.
د) بهترین مدلی را که یافته‌اید مشخص نمایید.

[2] Title
[3] Labeling
[4] https://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation
[5] Accuracy

۳- پیاده‌سازی k-means

دیتاست پیوست شده با نام "data_kmeans_1" را به کمک الگوریتم K-means و با K=2 خوشه‌بندی کنید. مرکز خوشه‌ها را تصادفی انتخاب کنید و از فاصله اقلیدسی استفاده کنید.

الف) نمودار Sum of Squared Error را در هر تکرار رسم کنید.
ب) نمودار Davies-Bouldin Index را در هر تکرار رسم کنید.
ج) داده‌ها را رسم کرده و داده‌های مربوط به هر خوشه را با رنگ متفاوتی نشان دهید.
د) با توجه به این بخش، الگوریتم K-means چه نقطه ضعفی دارد؟

دیتاست پیوست شده با نام "data_kmeans_2" را به کمک الگوریتم K-means و با K=2 خوشه‌بندی کنید. مرکز خوشه‌ها را تصادفی انتخاب کنید و از فاصله اقلیدسی استفاده کنید.

الف) نمودار Sum of Squared Error را در هر تکرار رسم کنید.
ب) نمودار Davies-Bouldin Index را در هر تکرار رسم کنید.
ج) داده‌ها را رسم کرده و داده‌های مربوط به هر خوشه را با رنگ متفاوتی نشان دهید.
د) با توجه به این بخش، الگوریتم K-means چه نقطه ضعفی دارد؟
ه) داده‌ها را با الگوریتمی مناسب خوشه‌بندی کنید.

۴- پیاده‌سازی خوشه‌بندی سلسله مراتبی[6]

دیتاست پیوست شده با نام "data_h" را به کمک الگوریتم خوشه‌بندی سلسله مراتبی، به ترتیب زیر، خوشه‌بندی کنید.

الف) با استفاده از الگوریتم K-means با K=4 و به روش Top-Down تا تقسیم داده‌ها به ۸ خوشه، خوشه‌بندی را ادامه دهید.
راهنمایی: ابتدا داده‌ها به دو خوشه تقسیم شده و سپس با توجه به یک معیار ارزیابی، در هر مرحله یکی از خوشه‌های حاصل در مرحله قبل، به دو خوشه تقسیم شده و این فرایند تکرار می‌شود. انتخاب معیار ارزیابی به عهده دانشجو است. نتیجه خوشه‌بندی در هر مرحله را نمایش دهید.
ب) برای محاسبه فاصله بین دو خوشه، از سه معیار AverageLink وCompleteLink و SingleLink استفاده می‌شود. سه معیار یاد شده را از لحاظ پیچیدگی زمانی و حساسیت به داده نویز بررسی کنید.
ج) با استفاده از الگوریتم K-means با K=8 و به روش پایین به بالا تا تقسیم داده‌ها به چهار خوشه، خوشه‌بندی را ادامه دهید. نتایج حاصل از سه معیار فاصله یاد شده در بالا را با هم مقایسه کنید. (از فاصله اقلیدسی استفاده کنید)

۵- پیاده‌سازی خوشه‌بندی DBSCAN

دیتاست‌های پیوست شده با نام‌های "pathbased-D۳۱-spiral- Compound " را به کمک الگوریتم DBSCAN خوشه‌بندی کنید. سعی کنید برای پارامترهای الگوریتم، مناسب‌ترین مقادیر را بدست آورید.
الف) معیار purity و تعداد خوشه‌های بدست آمده را نمایش دهید.
ب) داده‌ها را رسم کرده و داده‌های مربوط به هر خوشه را با رنگ متفاوتی نمایش دهید. توجه داشته باشید که داده‌ها می‌توانند متعلق به هیچ خوشه‌ای نبوده و نویز محسوب شوند. داده‌های نویز را با رنگی متفاوت نمایش دهید.
ج) با توجه به این بخش، نوع مجموعه دادگان، چه تاثیری بر عملکرد الگوریتم DBSCAN دارد؟

موفق باشید

[6] Hierarchical