دانشگاه اصفهان

تمرین شماره ۴

یادگیری ماشین در تجارت الکترونیک

موعد تحویل: **یکشنبه ۲۲ بهمن ۱۴۰۲**

نام استاد: **دکتر مرجان کائدی**

دستیار حل تمرین: **رضا شیری**

زمستان ۱۴۰۲

نکات مورد نیاز درباره‌ی تمرین

- فایل‌های ارسالی شما باید شما دو بخش کدهای پیاده‌سازی شده و همچنین یک گزارش فنی شامل توضیحات مربوط به تمرین باشد.
- در صورت مشاهده‌ی مشابهت‌های غیر طبیعی، **کسر نمره** به افراد تعلق می‌گیرد.
- در صورت استفاده از یک وب‌سایت یا منبع خاص، در فایل گزارش خود آدرس وب‌سایت یا نام منبع را ذکر کنید. در صورت عدم ذکر این موارد و پیدا شدن مشابهت‌های غیر طبیعی، کسر نمره تعلق می‌گیرد.
- شما می‌توانید حداکثر به‌صورت دو نفره، تمرین را تحویل دهید.
- پاسخ‌های خود را به صورت یک فایل zip به فرمت [name1]_[family1]_[name1]_[family1] [#assignment]_ [std#2]_ [std#1]_ در آورید که در آن [name] نام، [family] نام‌خانوادگی، [assignment#] شماره‌ی تکلیف و [std#] شماره دانشجویی می‌باشد. (به‌طور مثال Reza_Shiri_953611133047_Siavosh_Djazmi_993624008_#1)
- پاسخ‌های خود را تا **ساعت ۲۳:۵۹** روز **یکشنبه ۲۲ بهمن** در قسمت مربوط به تمرین ۳ کلاس تعریف شده در سامانه‌ی کوئرا ارسال کنید.
- مهلت پاسخ‌گویی به این تمرین تا پایان روز یکشنبه ۲۲ بهمن می‌باشد. با توجه به اختلالات مربوط به اینترنت و مشکلات دیگری که برای شما ممکن است پیش بیاید، شما می‌توانید تا **حداکثر ۵ روز** پس از اتمام مهلت ارسال تمرین، پاسخ‌های خود را **بدون کسر نمره**، در سامانه‌ی کوئرا آپلود کنید. توجه داشته باشید که **این زمان به‌هیچ عنوان قابل تمدید نمی‌باشد**.
- همچنین شما می‌توانید سوالات احتمالی خود را از طریق ایمیل rezamdd1998@gmail.com یا آیدی تلگرامی (@creation_bug) بپرسید. (توجه داشته باشید این ایمیل، فقط برای پاسخ‌گویی به سوالات مربوط به تمرین است و ارسال تکالیف به این آدرس نمره‌ای را به همراه نخواهد داشت.)

تمرین شماره ۴
نام درس: **یادگیری ماشین در تجارت الکترونیک**
نام استاد: **دکتر مرجان کائدی**
مهلت تحویل: **یکشنبه ۲۲ بهمن**

این تمرین، سری چهارم و آخرین سری از تمرینات درس یادگیری ماشین در تجارت الکترونیک است.

مجموعه‌داده‌ی مورد استفاده در این تمرین، مجموعه‌داده‌ی **Customer Classification** است که در تمرین قبلی هم مورد استفاده قرار گرفت. این مجموعه‌داده شامل اطلاعات مربوط به مشتریان (مانند منطقه‌ی زندگی، وضعیت بازنشستگی، وضعیت تأهل، میزان درآمد و...) است که با نام Telecust1.csv به همراه تمرین پیوست شده است.

هدف از این تمرین، خوشه‌بندی مشتریان با استفاده از الگوریتم k-میانگین[1] است که جزء الگوریتم‌های بدون‌نظارت به شمار می‌رود.

مجموعه‌داده‌ی ضمیمه‌شده را بخوانید و تنها سه ستون Tenure، Income و employe را نگه دارید. در صورت نیاز، پیش‌پردازش‌های موردنیاز و دلخواه را بر روی آن انجام دهید. با استفاده از تابع train_test_split موجود در کتابخانه‌ی sklearn، ۸۰ درصد داده‌ها را به داده‌ی آموزشی و ۲۰ درصد به داده‌ی آزمایشی اختصاص دهید. (**تنظیمات مقابل را بر روی آن اعمال کنید: shuffle=True, random_state=17**)

خوشه‌بندی[2]

در این تمرین شما باید با استفاده از توضیحاتی که در اسلایدهای مربوط به درس داشته‌اید یک خوشه‌بند k-میانگین پیاده‌سازی کنید. شما باید بر روی داده‌های آموزشی خوشه‌بندی را به‌ازای kهای مختلف (k=2,...,5) انجام دهید. تعداد دورهای انجام الگوریتم را ۱۰۰ و مراکز دسته‌ها را در ابتدا به‌صورت تصادفی در نظر بگیرید.

۱- به‌نظر شما به ازای k=1 الگوریتم فوق به چه صورت در می‌اید؟

در هر آزمایش و در انتهای دور ۱۰۰ام برای داده‌های آموزشی و آزمایشی:

- نمودارهای هر دسته را به‌همراه مرکز دسته‌ی مشخص شده رسم کنید. (مرکز دسته متمایز از داده‌ها باشد.)
- فاصله‌ی درون خوشه‌ای را به‌ازای هر خوشه گزارش دهید.
- مراکز خوشه‌ها و فاصله‌ی بین آن‌ها را گزارش کنید.

برای هر مقدار k شما باید ۴ آزمایش انجام دهید. ۳ آزمایش اول داده‌ها دو بعدی هستند (دو ستون از سه ستون مجموعه‌داده را انتخاب کنید) و آزمایش آخر هم با هر سه ستون انجام می‌شود.

توجه داده باشید که شما مجاز به استفاده از مدل‌های آماده **نیستید** و باید تمام روابط را خودتان پیاده‌سازی کنید.

در صورت آزمایشات بیشتر و بررسی موارد دیگر (مانند تعداد دورهای متفاوت، مقداردهی اولیه‌ی مختلف برای مرکز خوشه‌ها و...) **نمره‌ی اضافه** به شما تعلق می‌گیرد.

[1] K-means
[2] Clustering