به نام خدا

دانشگاه صنعتی امیرکبیر

دانشکده‌ی مهندسی کامپیوتر

پروژه نهایی درس یادگیری ماشین

دکتر احسان ناظرفرد

طراح سوال:

سید اردلان قریشی

محمدرضا امامی ناصری

بهمن ۱۳۹۹

توضیحات مهم:

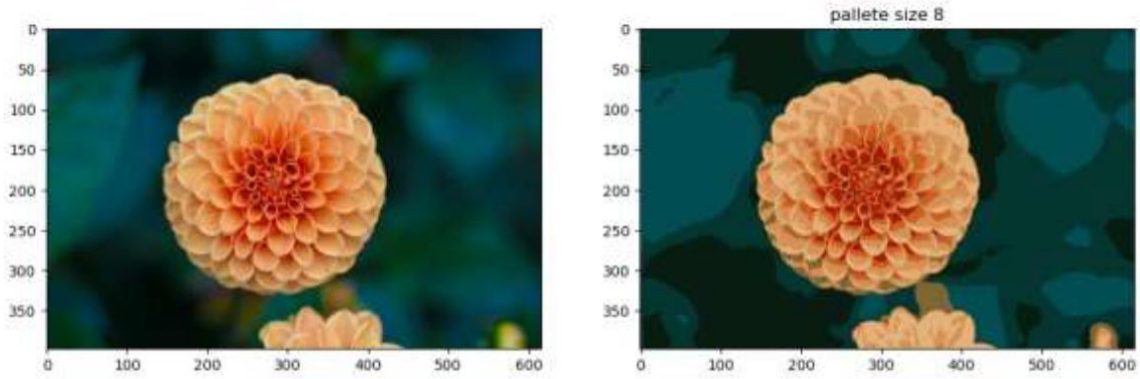
- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان #StudentId_Project.zip بارگذاری نمایید (به عنوان مثال 99131000_Project.zip).
- مهلت انجام تمرین تا **ساعت ۸:۰۰** روز **۱۵ اسفند** می‌باشد و به هیچ وجه تمدید نمی‌شود. **پس از این ساعت امکان بارگذاری و ویرایش تکالیف در سامانه وجود نخواهد داشت.**
- پروژه بدون گزارش فاقد ارزش می‌باشد و **نمره‌ای به آن تعلق نمی‌یابد**.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید (به بهترین گزارش نمره تشویقی تعلق می‌گیرد).
- مطابق قوانین دانشگاه هرگونه کپی برداری **ممنوع** می‌باشد و در صورت مشاهده نمره‌ی **هر دو طرف** صفر در نظر گرفته می‌شود.
- در صورت داشتن هرگونه ابهام می‌توانید از طریق ایمیل زیر سوال خود را مطرح نمایید:

MLAUTFALL99@gmail.com

۱- با استفاده از یک کتابخانه‌ی آماده که در آن الگوریتم خوشه‌بندی[1] K-means وجود دارد، موارد زیر را پیاده‌سازی نمایید.

۱,۱) تصاویر bee.jpg و parrots.jpg را خوانده و نمایش دهید. هر تصویر از تعدادی پیکسل[2] ساخته شده است و رنگ هر پیکسل با استفاده از ترکیب سه رنگ قرمز، سبز و آبی (RGB) ساخته می‌شود؛ به همین دلیل بعد از خواندن تصویر، مشاهده می‌کنید که تصویر خوانده شده یک ماتریس با مشخصات W*H*3 است به طوری که W و H اشاره به عرض[3] و طول[4] تصویر دارد و ۳ نشان دهنده‌ی هر کدام از سه رنگ RGB است. بنابراین، پیکسل‌های تصویر، داده‌های مورد نیاز مسئله می‌باشند که هر کدام دارای سه ویژگی هستند. پیکسل‌ها را با تعداد خوشه‌های ۲،۳،۴،۵،۶،۱۰،۱۵،۲۰ خوشه‌بندی کنید.

- بعد از هر بار خوشه‌بندی تصاویر، رنگ پیکسل‌ها را با رنگ مرکز خوشه‌ای که در آن قرار می‌گیرند جایگزین کنید و تصویر حاصل را نمایش دهید.



شکل ۱ تصویر سمت چپ تصویر واقعی و تصویر سمت راست با استفاده از الگوریتم k-means با تعداد خوشه ۸ ایجاد شده است

۱,۲) در این بخش مجموعه داده‌ی Shill Bidding Dataset.csv را بارگذاری کنید.

الف) یکی از روش‌های تعیین تعداد خوشه‌های بهینه در الگوریتم k-means استفاده از روش elbow است؛ این روش را توضیح دهید.

ب) تعداد خوشه‌ها را از ۱ تا ۱۰ تغییر دهید و الگوریتم را اجرا نمایید. با توجه به روش elbow بهترین تعداد خوشه، برای خوشه‌بندی مجموعه داده را مشخص نمایید و دلیل انتخاب خود را

[1] clustering
[2] pixel
[3] width
[4] height

توضیح دهید. نمودار هزینه بر حسب تعداد خوشه[5] را رسم کنید. (برای تابع هزینه می‌توانید از distortion یا inertia استفاده نمایید).

ج) معیار purity را به ازای تعداد خوشه برابر با ۲ (k=۲) محاسبه نمایید.

۲- با استفاده از الگوریتم خوشه‌بندی DBSCAN برای هر یک از مجموعه داده‌های موجود در پوشه‌ی مربوط به این سوال، نمونه‌ها را همراه با خوشه‌ی نسبت داده شده[6] رسم کنید. به این نکته توجه کنید که داده‌ها می‌توانند متعلق به هیچ خوشه‌ای نباشند و می‌توانند هنگام نمایش به عنوان نویز[7] تلقی شده و نمایش داده شوند. پس از اجرای الگوریتم خوشه‌بندی برای هر یک از مجموعه داده‌ها معیار purity را به دست آورده و به طور کیفی تاثیر نوع مجموعه داده بر کیفیت خوشه‌بندی را مقایسه و تحلیل کنید (در این سوال استفاده از کتابخانه آزاد است).

مجموعه داده‌ها: Compound – pathbased – rings – spiral – D31

۳- در این بخش می‌خواهیم دو الگوریتم value iteration و policy iteration را برای محیط[8] Frozen Lake مانند شکل زیر پیاده‌سازی نماییم. (برای آشنایی بیشتر می‌توانید به مستندات ارائه شده در پانویس[9] مراجعه کنید.)

S	F	F	F	H
F	F	H	H	F
F	F	F	F	F
H	H	F	H	F
F	F	F	F	G

[5] graph of cost sequence
[6] assigned cluster
[7] noise
[8] environment
[9] Link1, Link2, Link3

در این محیط عامل[10] با شروع حرکت از خانه‌ی شروع[11] (S) می‌خواهد به خانه‌ی هدف[12] (G) برسد. در این بین خانه‌های یخ زده[13]ای (F) هم وجود دارد. همچنین گودال[14]هایی (H) نیز در نقشه (محیط) موجود است. عامل باید از طریق خانه‌های یخ‌زده حرکت کرده و به خانه‌ی هدف برسد. توجه کنید که عامل به هدف محیط شامل احتمال گذار وضعیت‌ها[15] و میزان پاداش[16] دسترسی دارد. در این بخش قصد داریم سیاست بهینه[17] را برای محیط ۵ در ۵ شکل بالا به دست آوریم. هنگامی که عامل به خانه‌ی G یا H برسد یک اپیزود[18] تمام می‌شود. در صورتی که عامل در خانه G قرار بگیرد، پاداش ۱+ می‌گیرد. در مابقی موارد عامل پاداشی دریافت نمی‌کند. در پیاده‌سازی الگوریتم‌ها ۰٫۸۵=γ در نظر بگیرید (انتخاب شرط خاتمه‌ی مناسب به عهده‌ی شما می‌باشد).

الف) الگوریتم value iteration را پیاده‌سازی کرده و مقادیر*V را به دست آورید. زمان اجرا و تعداد تکرار[19] مورد نیاز را نمایش دهید. سیاست بهینه را به دست آورید و آن را به صورت یک جدول متشکل از حروف U(بالا)، R(راست)، D(پایین) و L(چپ) نمایش دهید.

ب) الگوریتم policy iteration را نیز مانند حالت قبل پیاده‌سازی کرده و زمان اجرا و تعداد تکرار آن را با مورد قبل مقایسه کنید. سیاست بهینه را مانند قسمت قبل نمایش دهید.

نکته: برای پیاده‌سازی قسمت‌های **الف** و **ب** می‌توانید از ابزار gym استفاده نمایید. برای آشنایی بیشتر به مستندات[20] آن رجوع کنید.

۴- مجموعه داده‌ی SeoulBikeData.csv در فایل مجموعه داده‌ها قرار داده شده است. با استفاده از آن موارد زیر را انجام دهید (استفاده از کتابخانه در تمامی بخش‌های سوال مجاز است).

الف) پیش‌پردازش[21]های لازم را انجام دهید.

ب) همبستگی[22] بین ویژگی‌ها را استخراج کرده و با توجه به آن بهترین ویژگی‌ها را انتخاب کنید. در این مرحله شما باید تعداد ویژگی‌های انتخاب شده را با توجه به یک مدل رگرسیون خطی پایه مورد بررسی قرار داده و بهترین K را پیدا کنید.

[10] agent
[11] start
[12] goal
[13] frozen
[14] hole
[15] state-transition probability
[16] reward
[17] optimal policy
[18] episode
[19] iteration
[20] gym doc (Link)
[21] preprocess
[22] correlation

ج) با استفاده از داده پیش‌پردازش شده مدل رگرسیون لسو[۲۳] را آموزش دهید. نقش پارامتر α در این مدل را بررسی کرده و با جستجو، بهترین مقدار آن را به دست آورید.

د) ویژگی‌های انتخاب شده در بخش **ب** و **ج** را با هم مقایسه کنید. چه نتیجه‌ای می‌گیرید؟

هـ) برای بهبود عملکرد مدل چه پیشنهادی دارید؟

۵- مجموعه داده‌ی heart_failure_clinical_records_dataset.csv را بارگذاری کنید. شما باید با استفاده از ویژگی‌های موجود، هر فرد را بر اساس مقادیر ستون DEATH_EVENT دسته‌بندی کنید (استفاده از کتابخانه در تمامی بخش‌های سوال مجاز است).

الف) پیش پردازش‌های لازم را انجام دهید. این مجموعه داده شامل مقادیر گم شده[۲۴] است. روش‌های مختلفی برای برطرف کردن این مشکل پیشنهاد شده است. درباره‌ی آن‌ها تحقیق کرده و با ذکر دلیل یکی از این روش‌ها را انتخاب کرده و مقادیر گم شده مجموعه داده را برطرف کنید.

ب) بهترین K ویژگی را با توجه به اهمیت آن‌ها انتخاب کنید. همانند بخش **ب** سوال ۴، باید تعداد ویژگی‌های انتخاب شده را با توجه به یک مدل دسته بند پایه[۲۵] مورد بررسی قرار دهید و بهترین K را پیدا کنید[۲۶].

ج) ۳ مدل مختلف رای‌گیری[۲۷] که هر کدام شامل ۳ دسته بند است را برای بهترین K ویژگی آموزش دهید و بهترین مدل را انتخاب کنید.

د) دسته‌بندهای مورد استفاده در بهترین مدل را با استفاده از داده‌های به دست آمده در بخش **ب** (بهترین K ویژگی) به صورت مجزا آموزش دهید. از مقایسه عملکرد دسته بندها به صورت تکی و گروهی چه نتیجه‌ای می‌گیرید؟

با آرزوی موفقیت!

[23] Lasso regression
[24] missing value
[25] base classifier
[26] more info (Link)
[27] voting