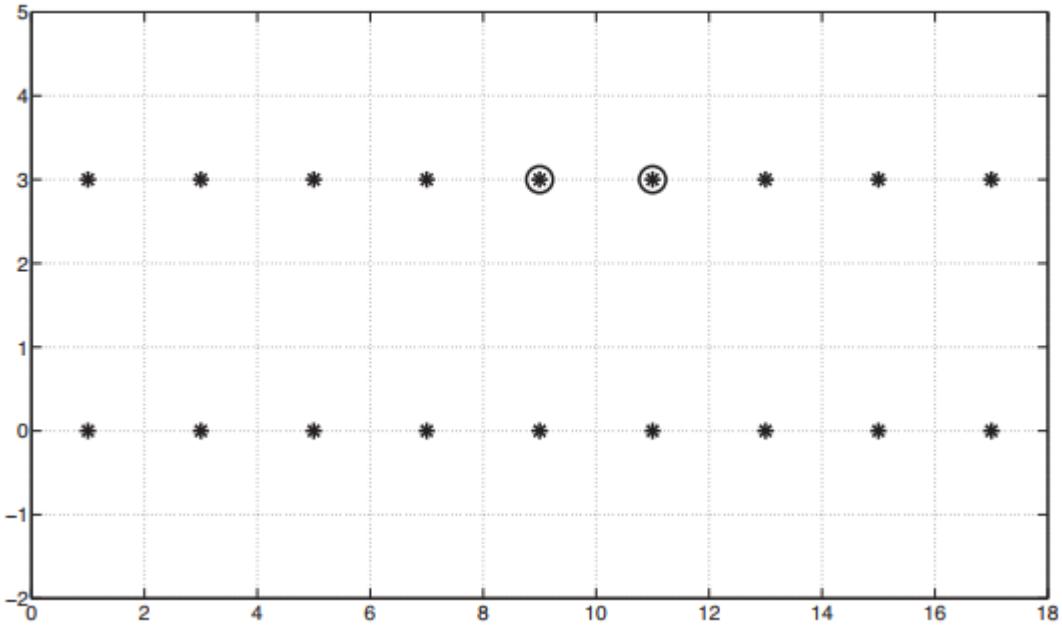
بخش اول: پرسش‌های تشریحی

سوال ۱) در شکل ۱، تعدادی نقاط داده را نشان می‌دهیم که روی شبکه اعداد صحیح قرار دارند (توجه داشته باشید که محور x فشرده شده است؛ فاصله‌ها باید با استفاده از مختصات شبکه واقعی اندازه‌گیری شوند.) فرض کنید الگوریتم K-Means را روی این دادهها اعمال می‌کنیم. با استفاده از K = 2 و با در نظر گرفتن دو نقطه داده دایره‌ای بعنوان مراکز اولیه، خوشه‌های نهایی را که پس از همگرایی K-means به دست می‌آیند رسم کنید (محل تقریبی مراکز جدید را نشان دهید و تمام نقاط اختصاص داده شده به هر مرکز را با هم گروه‌بندی کنید). راهنمایی: کوتاهترین فاصله اقلیدسی را در نظر بگیرید.



شکل ۱

سوال ۲) توضیح دهید در هر یک از کاربردهای زیر چگونه از خوشه‌بندی استفاده می‌شود.

الف) customer segmentation

ب) dimensionality reduction

ج) anomaly detection (outlier detection)

د) image segmentation

سوال۳) الگوریتم DBSCAN برای اجرای مناسب نیازمند تعیین درست پارامترهای خود می‌باشد به‌طوری که مقدارهای مختلف برای پارامترها می‌تواند باعث تولید خوشه‌های متفاوت و به تعداد مختلفی شود. راهکاری برای تعیین این پارامترها ارائه دهید.

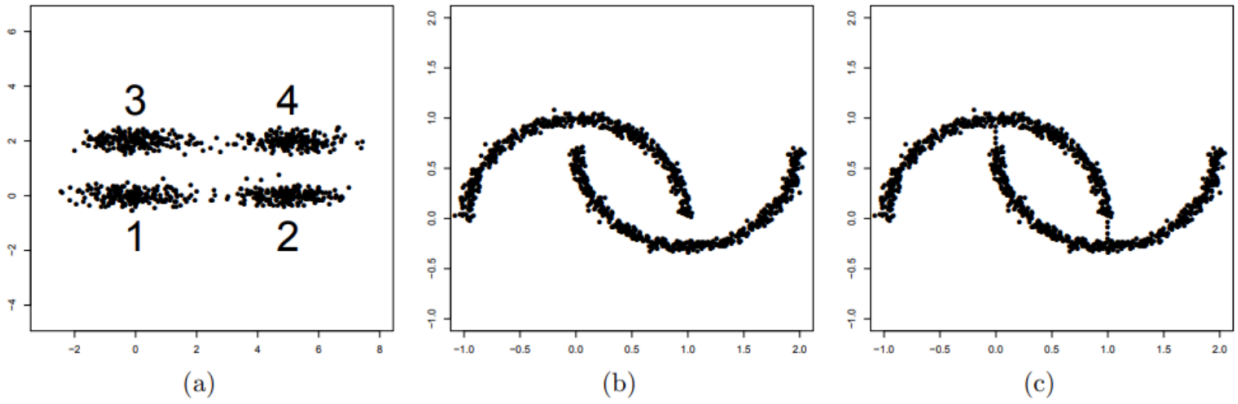
سوال۴) تفاوت دو الگوریتم policy iteration و value iteration در یادگیری تقویتی چیست؟

سوال۵) خوشه‌بندی سلسله مراتبی

الف) در خوشه بندی سلسله مراتبی، برای محاسبه فاصله بین دو خوشه، از معیارهای CompleteLink، singleLinkو AverageLink استفاده میشود. این سه معیار را توضیح دهید و از لحاظ پیچیدگی زمانی و حساسیت به داده پرت مقایسه کنید.

ب) با توجه به شکل ۲، در مجموعه داده a اگر با استفاده از خوشه‌بندی سلسله‌مراتبی با k=2 خوشه بندی صورت گیرد با استفاده از معیارهای شباهت خوشه‌ای CompleteLink، singleLink و AverageLink چهار دسته‌ی مشخص شده به چه خوشه‌ای تعلق می‌گیرند؟

ج) کدام یک از سه معیار فاصله (در صورت وجود) می‌تواند دو شکل ماه‌مانند در مجموعه داده‌b وc را با موفقیت جدا کند؟ پاسخ خود را به طور خلاصه توضیح دهید.



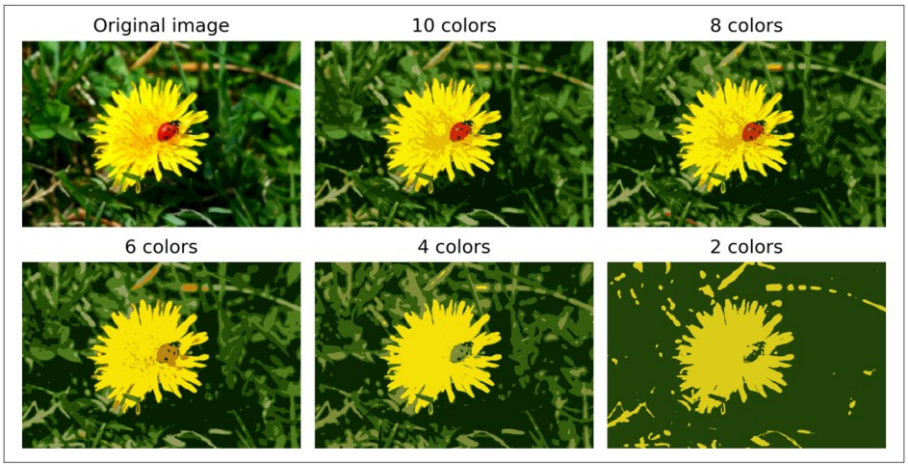
شکل ۲

بخش دوم: پیاده‌سازی

سوال اول:

همانطور که گفته شد یکی از کاربردهای الگوریتم k-means در تقسیم‌بندی(قطعه‌بندی) تصاویر می‌باشد. در این سوال قصد داریم یک نسخه بسیار ساده از قطعه‌بندی تصویر را پیاده‌سازی کنیم که به تقسیم‌بندی رنگ معروف است. اگر پیکسل ها رنگ مشابهی داشته باشند، به سادگی پیکسل ها را به همان بخش(خوشه) اختصاص می‌دهیم. در برخی از کاربردها، همین تسک ساده ممکن است کافی باشد. برای مثال، در تحلیل تصاویر ماهواره‌ای برای اندازه‌گیری میزان کل مساحت جنگل در یک منطقه می‌تواند مفید باشد.

تصاویر موجود در فولدر مربوط به این سوال را بخوانید و نمایش دهید. هر تصویر از تعدادی پیکسل تشکیل شده است. رنگ هر پیکسل با استفاده از ترکیب سه کانال قرمز، سبز و آبی(RGB) ساخته می‌شود. به همین دلیل بعد از خواندن تصویر، تصویر خوانده شده یک ماتریس با مشخصات W*H*3 است که W و H اشاره به عرض و طول تصویر دارد و ۳ نشان دهند رنگ‌های موجود در تصویر می‌باشد. پیکسل‌های تصویر، داده‌های مورد نیاز مسئله می‌باشد که هر کدام دارای سه ویژگی است. شکل ۳ نمونه‌ای از تقسیم‌بندی رنگ را با تعداد خوشه‌های مختلف نشان می‌دهد.



شکل ۳

الف) پیکسل‌ها را با تعداد خوشه‌های ۲،۳،۴،۵،۶،۱۰،۱۵،۲۰ با استفاده از الگوریتم k-means خوشه‌بندی کنید.

ب) بعد از هر بار خوشه‌بندی تصاویر، رنگ پیکسل‌های درون هر خوشه را، رنگ پیکسل مرکز خوشه قرار دهید و تصویر حاصل را نمایش دهید.

سوال دوم:

الگوریتم خوشه‌بندی DBSCAN را پیاده‌سازی نموده و برای ۵ مجموعه داده موجود در فولدر مربوط به این سوال، نمونه‌ها را همراه با خوشه‌ای که به آن تعلق دارند، رسم کنید. داده‌ها می‌توانند متعلق به هیچ خوشه‌ای نباشند و می‌توانند هنگام نمایش به عنوان نویز 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.

*سوال سوم (امتیازی):

مجموعه داده Olivetti faces شامل ۴۰۰ تصویر با ابعاد ۶۴ × ۶۴ پیکسل در مقیاس خاکستری از چهره ها است. هر تصویر به یک وکتور ۱ بعدی با اندازه ۴۰۹۶ مسطح شده است. از ۴۰ نفر مختلف (هر کدام ۱۰ بار) عکس گرفته شد و هدف آموزش مدلی است که بتواند

پیش بینی کند که کدام شخص در هر تصویر نشان داده شده است. توجه داشته باشید که مجموعه داده از قبل بین ۰ و ۱ مقیاس شده است.

مجموعه داده را با استفاده از تابع sklearn.datasets.fetch_olivetti_faces() بارگذاری کنید. سپس آن را به یک مجموعه آموزشی، یک مجموعه اعتبارسنجی و یک مجموعه آزمایشی تقسیم کنید. از آنجایی که مجموعه داده کوچک است، احتمالاً نیاز دارید که از نمونه‌گیری طبقه‌ای استفاده کنید تا اطمینان حاصل کنید که تعداد یکسانی از تصاویر برای هر نفر در هر مجموعه وجود دارد.

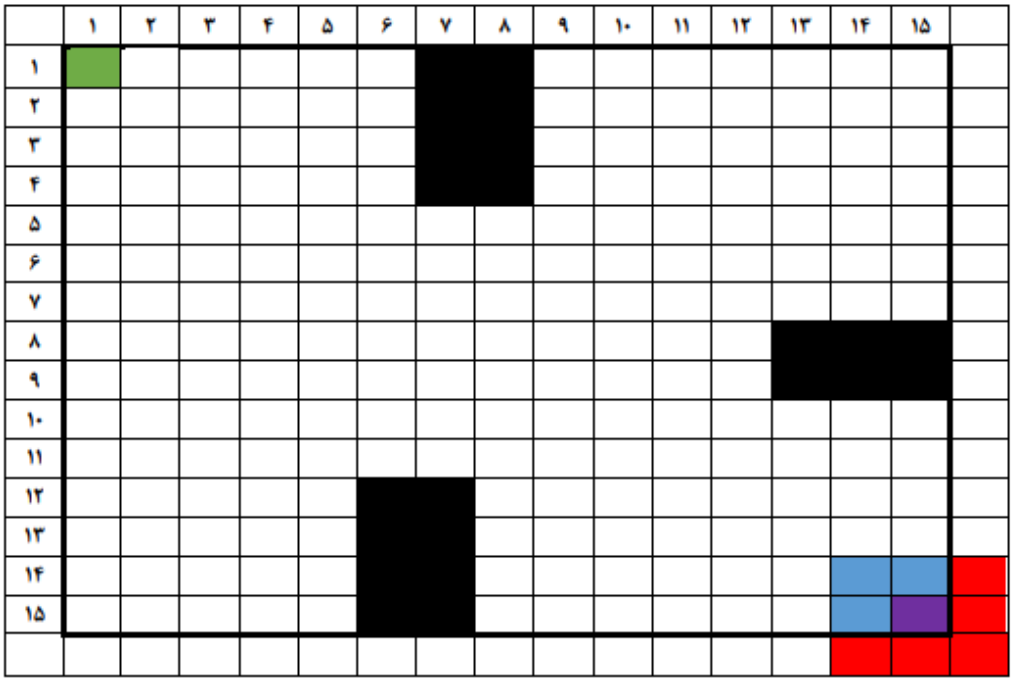
الف) توضیح دهید که نمونه‌گیری طبقه‌ای (stratified sampling) به چه معناست.

ب) الگوریتم k-means را پیاده‌سازی کرده و تصاویر را با استفاده از آن خوشه بندی کنید. برای تعیین تعداد مناسب خوشه‌ها از روش elbow استفاده کرده و نمودار آن را ترسیم کنید.

ج) خوشه ها را ترسیم کنید. آیا تصاویر چهره‌هایی که در هر خوشه هستند مشابه یکدیگرند؟

سوال چهارم:

مسئله مسیریابی برای ربات‌ها از مسائل مهم و اساسی برای ربات‌های امروزی می‌باشد. مواردی مانند عدم برخورد با موانع موجود و پیدا کردن بهترین مسیر در محیط داده شده از این قبیل مسائل می‌باشند. این مسائل را می‌توان در ترکیب با مسائل MDPحل کرد. محیط زیر را طبق توضیحات داده شده در نظر گرفته، و پس از پیاده سازی موارد خواسته شده، به سوالات هر بخش پاسخ داده و تحلیل خود را ارائه کنید. یک محیط grid طبق شکل زیر با ابعاد ۱۵ در ۱۵ را در نظر بگیرید. ربات ما در ابتدا در نقطه (۱۵و۱۵) قرار گرفته است. هدف ربات این است که به خانه (۱و۱) برود. ربات مورد نظر ما در هر استیت قادر به انجام ۹ عمل مختلف میباشد. ۸ عمل برای جابجایی در جهت های ۸ گانه (حرکت های مورب مجاز است) و یک عمل برای باقی ماندن در نقطه فعلی. برای ربات دو سری مجموعه حالت داریم. مجموعه حالت‌های قابل دسترسی و مجموعه حالت‌های ممکن. همسایه‌ی ممکن همسایهای است که خارج از محدوده محیط نباشد و مانع نباشد. همسایه ی در دسترس همسایه ای است که با یکی از اکشن های ممکنه بتوان به آن رسید. برای مثال در موقعیت ابتدایی ربات نقاط قرمز نشان داده شده نقاطی غیر قابل دسترسی هستند و نقاط آبی نشان داده شده نقاط قابل دسترس می‌باشند. همچنین تمامی حالت‌های موجود در محیط که مانع نیستند نیز نقاط ممکن می‌باشند.



شکل ۱ – ربات در ابتدا در خانه بنفش – هدف رسیدن به خانه سبز – نقاط مشکلی مانع

در انجام هر اکشن ربات با احتمال p به جهت انتخابی می‌رود و در غیر این صورت به یکی از همسایگان "ممکن و در دسترس" لیز میخورد. دقت شود که احتمال لیز خوردن به همه خانه‌ها به صورت یکنواخت و یکسان است. برای انجام هر حرکت بخاطر وجود انرژی مصرف شده و زمان تلف شده پاداش منفی‌ای درنظر گرفته شده است. همچنین هنگام برخورد با مانع هزینه‌ی برخورد با مانع نیز برای ربات در نظر گرفته شده است. همچنین در هنگام رسیدن به هدف ربات پاداش دریافت می‌کند.

حالت‌های زیر حالت‌های ممکن در محیط هستند:

- حالت پایه:
  - o احتمال انجام اکشن و رفتن به استیت بعدی برابر ۰,۸
  - o احتمال لیز خوردن ربات و رفتن به یک خانه "ممکن و در دسترس" یا ماندن در استیت فعلی برابر ۰,۲
  - o هزینه برخورد با مانع ۱-
  - o برای انجام هر حرکت به خاطر از دست وقت و انرژی پاداش منفی برابر با ۰,۰۱-
  - o پاداش رسیدن به خانه هدف برابر با ۱۰۰۰
- حالت حرکت بدون هزینه: همانند حالت پایه می باشد با این فرق که هزینه حرکت برابر با ۰ در نظر گرفته شود. و در صورت برخورد با مانع پاداش منفی ۰,۰۱برای ربات در نظر گرفته شود.
- حالت حرکت با هزینه زیاد: همانند حالت پایه میباشد با این تفاوت که هزینه هر حرکت برابر با ۱- (برای همه اکشنها به جز اکشن ماندن در خانه) ، هزینه برخورد با مانع برابر با ۱۰- و پاداش رسیدن به خانه هدف را برابر با ۱۰۰در نظر بگیرید.

با توجه بهMDP تعریف شده، توابع مشخص شده در فایل نوتبوک پیوست شده به همراه تمرین را کامل نموده تا مراحل زیر را پیاده سازی کرده و به سوالات مربوطه پاسخ دهید. لازم به ذکر است که اگر فرمت ارائه شده در توابع را رعایت ننمایید نصف نمره از شما کاسته خواهد شد.

۱. سیاست بهینه را برای حالت پایه با استفاده از روش policy iteration به دست آورید. مقدار discount factor برابر با ۰,۹ در نظر گرفته شود.
۲. سیاست بهینه را با روش policy iteration را برای حالت بدون اصطکاک به دست آورده و با نتایج مرحله قبلی مقایسه کنید. در مقایسه طول مسیر طی شده توسط ربات را در نظر داشته باشید. مقدار discount factor برابر با ۰,۹ در نظر گرفته شود.
۳. حال حالت با اصطکاک زیاد را در نظر گرفته و سیاست بهینه را با استفاده از روش policy iteration به دست آورده و با دو حالت قبل مقایسه کنید. مقدار discount factor برابر با ۰,۹ در نظر گرفته شود.
۴. با توجه به مراحل 1 و 2 بهترین حالت برای ریوارد محیط را در نظر گرفته و نقش تفاوت مقدارهای مختلف برای discount factor را برای 4 مقدار مختلف در مسئله بررسی کنید. تحلیل خود از نتایج به دست امده و همچنین آینده نگری ربات با توجه به discount factor تعیین شده را بررسی کنید.
۵. الگوریتم value iteration را برای محیط داده شده اجرا کرده و نتایج به دست آمده را با بهترین نتیجه قسمت قبل مقایسه کنید.

نکات مربوط به تحویل تمرین

- کدهای خود را ترجیحا به زبان پایتون و در محیط jupyter پیاده‌سازی کنید.
- نظم و خوانایی در نوشتن گزارش و کدها از اهمیت بالایی برخوردار است. کدهای خود را تا حد امکان کامنت‌گذاری کنید.
- در پیاده‌سازی بخش‌های مختلف امکان استفاده از کتابخانه‌های آماده مربوط به الگوریتم‌های یادگیری ماشین رابه طور کلی ندارید. مگر در مواردی که در صورت سوال ذکر شده باشد.
- برای خواندن داده‌ها می‌توانید از کتابخانه pandas و برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. برای محاسبه معیارهای ارزیابی مانند دقت، ماتریس درهم‌ریختگی و تقسیم داده‌ها به مجموعه‌های آموزش و آزمون نیز استفاده از کتابخانه آماده مجاز است.
- در صورتی که داده‌ها را به دسته‌های آموزشی، تست (ویا validation) تقسیم می‌کنید، درصد هر کدام را در گزارش ذکر کنید.
- فایل‌های کد و گزارش خود را در قالب یک فایل فشرده با فرمت HW04_StdNumber.zip که StdNumber شماره دانشجویی شماست، در سامانه بارگذاری کنید.
- سوالات ستاره دار(*) دارای نمره اضافی بر تمرین است.
- مطابق قوانین دانشگاه هر گونه کپی‌برداری ممنوع است و در صورت مشاهده، نمره **هر دو طرف صفر** داده می‌شود.
- در صورت وجود هر گونه سوال یا ابهامی با ایمیل درس در تماس باشید:

ml.ce.aut@gmail.com