## تکلیف شماره ۵ یادگیری ماشین: خوشه‌یابی

پیاده سازی الگوریتم خوشه‌یابی k-means :

۱- دادگان ۲ بعدی موجود در data1.mat را درنظر بگیرید.
الف) دادگان را رسم کنید تا از نظر نمایشی تعداد کلاسترهای بهینه مشخص شود.
ب) الگوریتم k-means را با درنظر گرفتن فاصله اقلیدسی برای دادگان موجود در data1.mat با مقدار k=3 با روش مقداردهی اولیه تصادفی پیاده کنید. **(از کد k-means متلب یا کتابخانه مرتبط در پایتون نباید به هیچ وجه استفاده شود، و خودتان باید کد بنویسید.)**
برای نشان دادن اشکال روش مقداردهی اولیه تصادفی، خودتان یک بار بصورت دستی (با توجه به نمایش دادگان) نقاط اشتباهی را بصورت نقاط اولیه الگوریتم وارد کنید، و پس از اجرای الگوریتم، اشتباه خوشه‌یابی را نمایش دهید.
پ) سپس برای اصلاح روش بخش قبلی، روش مقداردهی اولیه با دورترین نقطه را پیاده کنید. (توجه: بدون استفاده از کتابخانه خاص روش خوشه‌یابی)

۲- این بار عمل خوشه‌یابی باید روی دادگان data2.mat که دادگان تصویر از ۵۰۰۰ رقم دست نوشته هستند انجام شود. هر رقم یک تصویر خاکستری با بعد 10x10 پیکسل است، و بصورت بردارهایی با طول ۱۰۰ ذخیره شده است. متغیر X شامل تمام تصاویر در یک ماتریس 5000x100 می‌باشد، و بردار Y برچسب واقعی هر تصویر را دارد.
الف) الگوریتم k-means را با فاصله اقلیدسی روی دادگان X با مقدار دلخواه k پیاده کنید. برای مقادیر اولیه کلاسترها، نقاط تصادفی انتخاب کنید. برای هر اجرای الگوریتم کلاسترینگ، مقداردهی تصادفی را ۱۰ بار تکرار کنید، و بهترین نقطه شروع را بدست آورید. از همین روش پیدا کردن بهترین نقطه شروع در مراحل بعدی هم باید استفاده کنید.
ب) تابع هدف موردنظر مطابق آنچه در کلاس معرفی شد، مجموع فواصل within cluster می‌باشد. این بار برنامه را با k=10 اجرا کنید و منحنی مقادیر تابع هدف را نسبت به تعداد تکرارهای الگوریتم رسم کنید.
پ) مرحله قبل را با k=16 تکرار کنید. منحنی چه تفاوتی نسبت به قسمت قبل دارد؟
ت) الگوریتم را به ازا مقادیر k=1, 5, 10, 15, 16, 20 اجرا کنید، و مقادیر تابع هدف را برای هر مقدار k گزارش کنید. همچنین می‌خواهیم مقدار دقت الگوریتم خوشه‌یابی یعنی نسبت تعداد نمونه‌های درست برچسب خورده به تعداد کل نمونه‌ها را برای تمام مقادیر k مشخص شده پیدا کنیم. همانطور که می‌دانید پیدا کردن چنین چیزی در الگوریتم کلاسترینگ ساده نیست، زیرا الگوریتم فقط خوشه‌یابی انجام می‌دهد و معلوم نیست که نمونه‌های موجود در هر خوشه مربوط به کدام کلاس باشند (یا کدام برچسب را داشته باشند.) با توجه به اینکه مقادیر برچسب واقعی در بردار Y داده شده است، چه روشی برای پیدا کردن میزان دقت الگوریتم خوشه‌یابی پیشنهاد می‌کنید؟ مقدار دقت روش پیاده شده را محاسبه کنید.

**راهنمایی: به برچسب بیشترین تعداد نقاط توجه کنید.**
ث) در مرحله قبل با توجه به توضیحات درس، مقدار اپتیمم k چیست؟ از کجا مشخص کردید؟

۳- یک مجموعه دادگان تصادفی در داخل ۲ بیضی دلخواه بسازید (مثلا دو بیضی افقی که کمی با هم همپوشانی داشته باشند). سپس دادگان را نمایش دهید.
الف) حال دادگان موجود را با روش fuzzy c-means با k=2 خوشه‌یابی کنید و حاصل را نمایش دهید. (توجه: برای الگوریتم fuzzy c-means می توانید از کتابخانه موجود استفاده کنید.)
ب) مشکل روش قسمت قبل در چیست؟ آیا خوشه‌یابی را درست انجام داده است؟ اگر مشکلی دارد، چگونه می‌توان آن را برطرف کرد؟ کد آن را بنویسید و نتیجه را نمایش دهید **(قسمت تشویقی)**

**گزارش: برنامه های نوشته شده به همراه نتایج و شکل های خواسته شده در هر مرحله را ضمیمه کنید. گزارش شامل تحلیل و مقایسات و توضیحات کافی در مورد نحوه پیاده سازی ها، لیست برنامه ها و پارامترهای بکار رفته می‌باشد. لطفا از نوشتن گزارشهای طولانی اجتناب کنید.**
**توجه: تمرین و گزارش باید بصورت انفرادی نوشته شود، و به برنامه‌ها و گزارش های کپی شده نمره‌ای تعلق نمی‌گیرد.**

**نمره تشویقی: نوشتن کد درست و نتیجه درست از قسمت ۳-ب**

**روش تحویل: آپلود در courses**
**موعد تحویل: ۲۸ مرداد ۱۳۹۹**

موفق باشید
سیدین