دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سری پنجم درس یادگیری ماشین

زمستان ۱۳۹۶

سؤال ۱

در روش‌هایی که نیاز به محاسبه مرکز خوشه دارند، معیارهای مختلفی برای محاسبه مرکز وجود دارد. به‌طورمعمول از معیارهایی نظیر میانگین، مد(داده‌ای با بیشترین بسامد تکرار) و میانه استفاده می‌شود. یک مزیت برای هرکدام از موارد یادشده، ذکر کرده و دلیل آن را مختصراً توضیح دهید.

سؤال ۲

در یک مجموعه داده $KMeans$ را به دو روش پیاده‌سازی می‌کنیم و نتایج را نگه می‌داریم.

- در حالت اول بدون هیچ پیش‌پردازشی بر روی دادگان، خوشه‌بندی را انجام می‌دهیم.
- در حالت دوم دادگان را نرمال می‌کنیم؛ بدین معنی که تمامی ویژگی‌ها را در بازه $[0,1]$ نگاشت می‌کنیم و سپس خوشه‌بندی کرده و نتایج را دوباره ذخیره می‌کنیم.

آیا نتایج هر دو روش یکسان است؟ چرا؟

سؤال ۳

در خوشه‌بندی سلسله مراتبی، برای محاسبه فاصله بین دو خوشه، از معیارهای SingleLink, CompleteLink, AverageLink استفاده می‌شود. این سه معیار را ازلحاظ پیچیدگی زمانی و حساسیت به داده پرت مقایسه کنید(برای هر مورد توضیح مختصری نیز بدهید).

سؤال ۴

در روش $DBScan$ تعیین دو فراسنج حداقل نقاط داخل دایره و شعاع دایره، نقش مهمی در خروجی الگوریتم دارد. یک روش برای تعیین این دو فراسنج پیشنهاد دهید.

سؤال ۵

دادگان موجود در فایل اکسل data1 را در نظر بگیرید. ستون A مقادیر x و ستون B و C مقادیر y متناظر را نشان می‌دهند. یعنی هر سطر متناظر با دو نقطه(داده) است. برای مثال سطر دوم را در نظر بگیرید. این سطر مقدار x برابر با 0.9 و مقادیر y 0.06411 و 0.93589 را دارد؛ یعنی دو نقطه زیر را بیان می‌کند:

$$d_1 = (0.9, 0.06411)$$
$$d_2 = (0.9, 0.93589)$$

الف) الگوریتم $KMeans$ را بر روی دادگان داده‌شده به ازای $K = [2,3,4,5]$ پیاده‌سازی کنید. معیار فاصله اقلیدسی و انتخاب نقاط اولیه تصادفی است. برای هر مورد، شاخص دیویس بولدین[1] را محاسبه کرده و بهترین خوشه‌بندی را که بر اساس شاخص ذکرشده به دست آورده‌اید، رسم کنید. شاخص یادشده به شکل زیر محاسبه می‌شود:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\frac{\max_{j}(\mu_i + \mu_j)}{d(c_i, c_j)}$$

n تعداد خوشه‌ها، c_i مرکز خوشه iام، و μ_i میانگین فاصله تمام نقاط خوشه i تا مرکز خوشه(c_i) است.

خروجی موردنظر: گزارش شاخص دیویس بولدین به ازای $K = [2,3,4,5]$ ، گزارش بهترین مقدار K بر اساس شاخص ذکرشده و رسم نمودار بهترین خوشه‌بندی به‌دست‌آمده.

ب) الگوریتم $DBScan$ را بر روی دادگان داده‌شده پیاده‌سازی کنید. از روش پیشنهادی در سؤال ۴ برای تعیین فراسنج‌ها استفاده کنید. خوشه‌بندی به‌دست‌آمده را رسم کنید.

خروجی موردنظر: گزارش مقادیر فراسنج‌های $DBScan$، رسم خوشه‌بندی نهایی.

ج) به‌صورت شهودی، خوشه‌بندی‌هایی که از هر دو قسمت الف و ب رسم کرده‌اید، باهم مقایسه کنید. کدام‌یک بهتر عمل کرده‌است؟ چرا؟

سوال ۶

مجموعه داده‌ی data2 را بارگذاری کنید. این مجموعه داده ۲ بعد داشته که در ۲ ستون قرار داده شده‌اند.

الف) با استفاده از الگوریتم $Kmeans$ با $K = 2$ الگوریتم خوشه‌بندی $\mathrm{Top} - \mathrm{Down}$ را بر روی این مجموعه داده پیاده‌سازی کنید. در این قسمت ابتدا داده‌ها به دو خوشه تقسیم شده و سپس با توجه به یک معیار ارزیابی خوشه‌بندی، در هر مرحله یکی از خوشه‌های حاصل در مرحله‌ی قبل به ۲ خوشه تقسیم شده و این فرآیند تکرار می‌شود. انتخاب معیار ارزیابی مناسب به عهده‌ی دانشجو می‌باشد. نتیجه‌ی خوشه‌بندی را از ابتدای بارگذاری داده‌ها تا تقسیم داده‌ها به ۱۰ خوشه با رنگ یا شکل‌های مختلف برای هر خوشه نمایش دهید.

ب) چگونه می‌توان تعداد خوشه‌ها را به صورت خودکار انتخاب نمود؟ روش ارائه شده را توضیح داده و پیاده سازی کنید. نتیجه‌ی خوشه بندی نهایی را گزارش کنید.

[1] Davies–Bouldin index

ج) الگوریتم خوشه‌بندی $Bottom-Up$ با معیارهای $SingleLink$ ، $CompleteLink$ و $AverageLink$ را پیاده سازی کرده و نتایج ۱۰ تا ۲ خوشه‌ی نهایی را نمایش دهید. نتایج سه روش را با هم مقایسه کنید. (از فاصله‌ی اقلیدسی استفاده کنید)

توضیحات تمرین:

۱- شما باید سورس کد خود به همراه مستندات(پاسخ سؤال‌ها و نتایج پیاده‌سازی که خواسته‌شده است) را در قالب یک فایل zip که نام فایل $xxxxx\_hw5$ که $xxxxx$ شماره دانشجویی شما است، تحویل بدهید.

۲- پیاده‌سازی با متلب یا پایتون باید انجام شود.

۳- در صورت هرگونه سؤال یا ابهام به ceitml17@gmail.com ایمیل بزنید.