بسمه تعالی

الگوشناسی آماری ۷۲۵-۴۰ (گروه ۲) | **نیمسال دوم ۹۲-۹۱**

تمرین سری پنجم: خوشه‌بندی | **مدرس: سلیمانی**

موعد تحویل: ۵ خرداد ۹۲ | **نمره از ۷۰+۵**

سوال ۱ (۱۵ نمره): خوشه‌بندی k-means

۱.۱. تابع هدف k-means برای خوشه‌بندی داده‌های $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$ به صورت $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_k\}$ را در نظر بگیرید:

$$J(\mathcal{C}) = \sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k \right\|^2$$

a. (۴ نمره) نشان دهید تابع هدف بالا، برابر با تابع هدف $J'(\mathcal{C})$ است:

$$J'(\mathcal{C}) = \sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)} \right\|^2$$

b. (۳ نمره) با توجه به تابع هدف $J'(\mathcal{C})$، معیار خوشه‌بندی k-means را توصیف نمایید. آیا طبق این معیار k-means به خوشه‌هایی که تعداد داده‌هایشان تقریباً برابر باشد، تمایل بیشتری دارد یا فقط میانگین فواصل دورن-خوشه‌ای را حداقل می‌کند؟

۲.۱. (۵ نمره) نشان دهید در روش k-means هر دو گام E و M (قبل از همگرایی) باعث کاهش مقدار تابع هزینه‌ی k-means می‌شوند.

۳.۱. (۳ نمره) دقیقاً چه حالتی از اجرای الگوریتم EM برای پیدا کردن GMM معادل k-means می‌شود؟ (پارامترهای GMM چگونه باشد و به جای محاسبه‌ی احتمال posterior برای شماره خوشه، در گام E چه اقدامی انجام می‌شود)

سوال ۲ (۱۰ نمره): الگوریتم EM

مجموعه داده‌ی $\{\binom{-1}{1}, \binom{2}{-1}, \binom{-3}{*}, \binom{2.5}{1.5}, \binom{*}{-0.5}, \binom{0.5}{0.5}\}$ شامل ۵ نمونه با دو ویژگی (که برخی از نمونه‌ها مقدار یکی از ویژگی‌هایشان نامعلوم است) را در نظر بگیرید. فرض کنید داده‌ها از توزیع یکنواخت (uniform) زیر در فضای دوبعدی حاصل شده‌اند:

$$p(\boldsymbol{x}) = \begin{cases} \dfrac{1}{|a-b| \times |c-d|}, & a \le x_1 \le b, c \le x_2 \le d \\ 0, & othewise \end{cases}$$

چنان‌چه از الگوریتم EM برای پیدا کردن پارامترها $\boldsymbol{\theta} = [a, b, c, d]$ استفاده شود و مقادیر اولیه $\boldsymbol{\theta}^0 = [-5, -5, 5, 5]$ در نظر گرفته شود:

a) با شروع از $\boldsymbol{\theta}^0$ گام‌های E و M را انجام دهید.

b) مقدار تقریبی پارامترهای حاصل را پس از همگرایی مشخص نمایید.

سوال ۳ (۱۵ نمره): خوشه‌بندی سلسله‌مراتبی

۱.۳. (۸ نمره) فرض کنید $\mathcal{C}_i$ مجموعه داده‌های موجود در خوشه‌ی i–ام و $\mathcal{C}_i \cup \mathcal{C}_j$ خوشه‌ی حاصل از ادغام $\mathcal{C}_i$ و $\mathcal{C}_j$ را مشخص کند. نشان دهید چنان‌چه برای پیدا کردن فاصله‌ی خوشه‌ی حاصل از ادغام $\mathcal{C}_i$ و $\mathcal{C}_j$ با خوشه‌ی $\mathcal{C}_k$، از معیار فاصله‌ی زیر استفاده شود:

$$d(\mathcal{C}_i \cup \mathcal{C}_j, \mathcal{C}_k) = \alpha_i d(\mathcal{C}_i, \mathcal{C}_k) + \alpha_j d(\mathcal{C}_j, \mathcal{C}_k) + \beta d(\mathcal{C}_i, \mathcal{C}_j) + \gamma \left| d(\mathcal{C}_i, \mathcal{C}_k) - d(\mathcal{C}_j, \mathcal{C}_k) \right|$$

این معیار با انتخاب‌های زیر برای پارامترهای α_i، α_j، α_i و α_i به معیار فاصله‌ی بین خوشه‌ای در روش‌های مختلف سلسله‌مراتبی تبدیل خواهد شد:

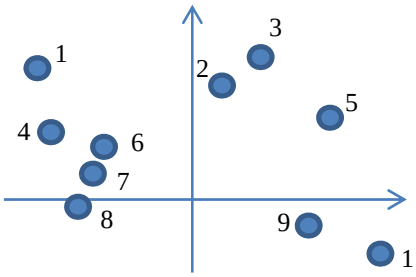
a. Single-link: $\alpha_i = \alpha_j = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$

b. Complete-link: $\alpha_i = \alpha_j = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$

c. Average-link: $\alpha_i = \frac{N_i}{N_i+N_j}, \alpha_j = \frac{N_j}{N_i+N_j}, \beta = \gamma = 0$

N_i تعداد داده‌های موجود در خوشه‌ی $\mathcal{C}_i$ را مشخص می‌کند.

۲.۳. داده‌های موجود در شکل زیر را در نظر بگیرید:



a. (۵ نمره) بدون پیاده‌سازی روش خوشه‌بندی Single-Link، Dendrogram حاصل از اعمال این روش را بر روی مجموعه داده‌ی بالا رسم نمایید.

b. (۲ نمره) تعداد خوشه‌های مناسب طبق این Dendrogram را تعیین نمایید.

سوال ۴ (۳۵ نمره): پیاده‌سازی و مقایسه روش‌های خوشه‌بندی

مجموعه داده‌ی موجود در فایل "two-circles.mat" را در نظر بگیرید که تعداد خوشه‌ها $C = 2$ فرض شده است. در این تمرین قصد داریم عملکرد روش‌های خوشه‌بندی مختلف را روی این مجموعه داده بررسی نماییم.

۱.۴. (۶ نمره) تابعی برای محاسبه‌ی معیار RandIndex جهت ارزیابی خوشه‌بندی داده‌ها پیاده‌سازی نمایید که بردار برچسب به‌دست آمده برای داده‌ها توسط یک الگوریتم خوشه‌بندی و همچنین برچسب مطلوب داده‌ها را می‌گیرد و مقدار معیار RandIndex را برمی‌گرداند.

۲.۴. روش خوشه‌بندی طیفی (spectral) برش نرمال (NCut):

a. (۷ نمره) روش NCut را پیاده‌سازی کنید (برای تشکیل گراف همسایگی از تابع knn که در اختیار شما گذاشته شده، استفاده نمایید). این روش باید با دریافت پارامترهای k و σ و همچنین مجموعه داده‌ی بدون برچسب، نتیجه‌ی خوشه‌بندی داده‌ها را مشخص نماید.

b. (۷ نمره) با تغییر پارامترهای گراف همسایگی به صورت $k = \{3,5,7,10,15,20\}$ و مقدار انحراف استاندارد $\sigma = \{0.1,0.5,1,2,5,10\} * \bar{d}$ برای تابع وزن‌دهی گاوسی که $\bar{d}$ متوسط فواصل داده‌ها با k–نزدیک‌ترین همسایه‌هایشان را نشان می‌دهد، نتایج خوشه‌بندی را به‌دست آورید. مقدار RandIndex به دست آمده را به ازای این تغییر پارامترها در یک شکل کلی (به ازای هر مقدار k یک نمودار) رسم نمایید؟

۳.۴. (۱۵ نمره) نتایج روش‌های خوشه‌بندی Single-Link و Complete-Link، k-means، EM+GMM، NCut ($k = 3, \sigma = \bar{d}$) را روی مجموعه داده‌ی مذکور، با تعیین مقدار RandIndex مشخص نمایید. شکل‌هایی که برای برچسب‌گذاری داده‌ها توسط این روش‌ها به دست می‌آید را در گزارش نشان دهید.

(برای پیاده‌سازی روش‌های خوشه‌بندی سلسله‌مراتبی، k-means و EM+GMM می‌توانید به ترتیب از توابع linkage، kmeans و gmdistribution.fit در Matlab استفاده نمایید)

موفق باشید