بنام خدا

الگوشناسی آماری (CE-725)

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

تمرینات سری ششم (خوشه‌بندی و EM)

بهار ۱۳۹۱

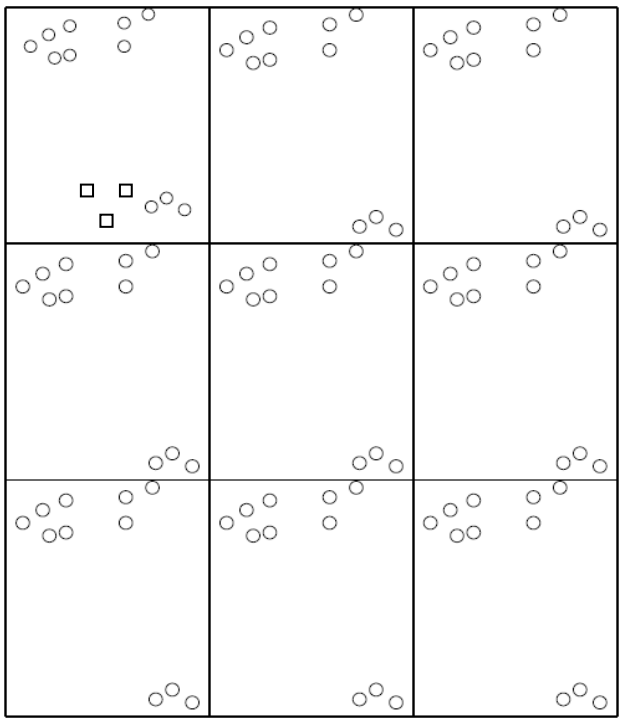
به نکات زیر توجه فرمائید:

۱. زمان تحویل تمرینات در سایت درس مشخص شده است. دقت نمائید که زمانبندی‌های تعیین شده قابل تغییر نیستند.

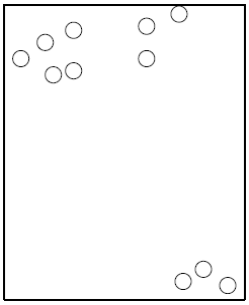
۲. تمرینات را با عنوان SPR-HWx-8xxxxxxx (مثلا SPR-HW1-88300785) و در یک فایل فشرده با همین نام به آدرس Muhammadi@dml.ir ایمیل زده و در اولین جلسه بعد از زمان تحویل، بصورت پرینت شده تحویل استاد درس دهید.

۳. گزارش شما باید مختصر و مفید باشد. برای تمرینات پیاده‌سازی که با لوگوی مشخص شده‌اند باید کد متلب نوشته شده ضمیمه گزارش شده و تمامی خروجی‌های برنامه‌ها در گزارش شما ذکر شوند.

سوال ۱) الف) الگوریتم K-means را بصورت دستی بر روی داده‌های زیر اجرا نمائید. دایره‌ها نشان دهنده داده‌های موجود بوده و مربع‌ها، مراکز خوشه‌ها را نشان می‌دهند. مراکز ابتدایی خوشه‌ها در ابتدا بصورت زیر داده شده‌اند. مرزهایی را که توسط این مراکز پدید می‌آیند را نشان داهید (فضا به چه نحوی افراز می‌شود؟). بر روی تصاویر داده شده، مراحل الگوریتم را تا حالت همگرایی، مرحله به مرحله، نشان دهید.



ب) فرض کنید ما توسعه‌ای از الگوریتم K-means را داریم که به جای استفاده از فاصله‌ی اقلیدسی از فاصله‌ی ماهالانوبیس استفاده می‌کند و خوشه‌های تولیدی آن به جای دایره، دارای شکل گاوسی می‌باشند (در هر مرحله از الگوریتم، پس از نسبت دادن هر کدام از داده‌ها به یکی از مراکز و تشکیل دادن خوشه‌های موقتی، از الگوریتم ML برای تخمین پارامترهای گاوسی جدید هر خوشه استفاده نموده و گاوسی‌های هر خوشه را بروز می‌کنیم). اگر این الگوریتم از همان مراکز خوشه‌ی ابتدایی بخش قبلی استفاده کند، در نهایت حدس می زنید گاوسی‌های نهایی چگونه باشند؟ بیضی‌های معرف هر مدل گاوسی را در شکل روبرو نمایش دهید (از محاسبات دقیق خودداری نموده و شکل‌های تقریبی را حدس بزنید – کوواریانس‌های ابتدایی را که باید بصورت تصادفی در نظر گرفته شوند، شما برای راحتی کار، مطابق نیاز خود در نظر بگیرید).



پ) آیا نتایج خوشه‌بندی مراحل (الف) و (ب) فوق یکسان هستند؟ چرا؟

سوال ۲) فرض کنید مجموعه داده‌ی ۶ عضوی زیر را داشته باشیم:

$$S=\{a=(0,0),\ b=(8,0),\ c=(16,0),\ d=(0,6),\ e=(8,6),\ f=(16,6)\}$$

بر روی این داده‌ها الگوریتم K-means با k=3 را می‌خواهیم اعمال کنیم. معیار ارزیابی فاصله در الگوریتم را همان فاصله اقلیدسی در نظر می‌گیریم. قبل از مطرح نمودن سوال ابتدا به دو تعریف زیر توجه نمائید:

- K-Starting Configuration (K-SC): یک زیر مجموعه‌ی kتایی از S، که مراکز اولیه ما را نشان می‌دهند. مثلا {a,b,c}.
- K-partition (K-P): یک افراز از S به k زیر مجموعه‌ی غیر تهی را گوئیم. مثلا {{a,b,e}, {c,d}, {f}}.
- یک K-P پایدار خوانده می‌شود، اگر تکرار اجرای مراحل الگوریتم K-means بر روی آن باعث ایجاد تغییرات در خوشه‌های نتیجه شده نشود.

الف) تعداد 3-SCهای موجود بر روی S داده شده چند تاست؟

ب) جدول زیر را پر کنید:

3-P	پایدار است؟	یک 3-SC اولیه که با اجرای K-means بتواند به این 3-P منجر شود.	تعداد 3-SCهای منجر شونده به این 3-P؟
{a,b,e}, {c,d}, {f}			
{a,b}, {d,e}, {c,f}			
{a,d}, {b,e}, {c,f}			
{a}, {d}, {b,c,e,f}			
{a,b}, {d}, {c,e,f}			
{a,b,d}, {c}, {e,f}			

سوال ۳) در کلاس خوشه‌بندی سلسله مراتبی توضیح داده شد، که الگوریتم پایه آن به صورت زیر است:

- از یک نقطه در داخل یک خوشه شروع کن.
- تا زمانی که تنها یک خوشه باقیمانده باشد تکرار کن:
  - نزدیکترین جفت خوشه را پیدا کن.
  - آنها را ترکیب کن.
- درخت خوشه‌های ترکیب شده را برگردان.

برای اینکه الگوریتم بالا را به یک رویه مشخص تبدیل کنیم نیاز به تعیین معیار نزدیکی دو خوشه داریم (چند روش برای تعریف فاصله بین دو خوشه عبارتند از: single-link, complete-link, average-link). در این مساله می خواهیم از معیار دیگری برای فاصله بین دو خوشه غیر متصل استفاده کنیم، بدین صورت که فاصله بین دو خوشه غیر متصل X و Y برابر است با اینکه چه مقدار مجموع مربعات در زمانی که آن دو خوشه را ترکیب می‌کنیم افزایش می‌یابد:

$$\Delta(X,Y)=\sum_{i\in X\cup Y}||\vec{x}_i-\mu_{\overline{X\cup Y}}||^2-\sum_{i\in X}||\vec{x}_i-\mu_{\bar{X}}||^2-\sum_{i\in Y}||\vec{x}_i-\mu_{\bar{Y}}||^2$$

که در آن μ_i مرکز ثقل خوشه i و x_i یک داده از خوشه مربوطه می باشد. $\Delta(X,Y)$ مقدار هزینه ترکیب دو خوشه X و Y به یک خوشه را مشخص می‌کند. پس در الگوریتم بالا دو خوشه‌ای به یکدیگر نزدیکتر هستند که هزینه ترکیب آنها کمینه باشد.

الف) آیا می توانید فرمول مشخص شده در معادله بالا برای $\Delta(X,Y)$ را به شکل ساده‌تری تبدیل کنید؟ فرمول نهایی باید بر اساس اندازه خوشه ها (η_X,η_Y) و $||\mu_{\bar{X}}-\mu_{\bar{Y}}||^2$ (فاصله بین مرکز ثقل دو خوشه) باشد.

ب) تفسیر معیار فاصله ارائه شده چیست (فرمول ساده شده به دست آمده در قسمت قبل می‌تواند در جواب دادن به این قسمت کمک کند)؟

پ) فرض کنید که دو مجموعه دو خوشه‌ای P_1 و P_2 داده شده است. به صورتی که فاصله بین دو مرکز ثقل در مجموعه P_1 بیشتر از فاصله بین دو مرکز ثقل در مجموعه P_2 است. آیا با استفاده از معیار فاصله بالا همیشه دو خوشه در مجموعه P_2 برای ترکیب شدن انتخاب می‌شوند؟ چرا؟

ت) در روش‌های خوشه‌بندی معمولا مشخص نیست که چه تعداد خوشه برای داده‌ها مناسب است. از معیار فاصله و الگوریتم خوشه‌بندی داده شده استفاده کنید و یک روش اکتشافی ارائه دهید تا تعداد خوشه‌های مناسب را به دست آورد؟

سوال ٤) فرض کنید که مجموعه ای از نقاط (X,Y,Z) براساس مدل زیر تولید شده‌اند (X و Y و Z متغیرهای بولین هستند).

- متغیر X با احتمال α برابر ۱ قرار داده می‌شود.
- متغیر Y با احتمال β برابر ۱ قرار داده می‌شود.
- اگر (X,Y)=(1,1) باشد، متغیر Z با احتمال λ_{11} برابر ۱ قرار داده می‌شود.
- اگر (X,Y)=(0,1) باشد، متغیر Z با احتمال λ_{01} برابر ۱ قرار داده می‌شود.
- اگر (X,Y)=(1,0) باشد، متغیر Z با احتمال λ_{10} برابر ۱ قرار داده می‌شود.
- اگر (X,Y)=(0,0) باشد، متغیر Z با احتمال λ_{00} برابر ۱ قرار داده می‌شود.

در این مساله ما نیاز داریم که پارامترهای مدل را به دست آوریم و می‌دانیم که یکی از متغیرها، Y، مشاهده نشده است و به ما یک مجموعه m تایی از نقاط داده شده است $D = \{(x_i, z_i)\ ;\ 1 \le i \le m\}$.

گام‌های E و M را محاسبه کرده و عبارات صریحی برای بروز رسانی پارامترها در فرآیند EM به دست آورید.

سوال ۵) فرض کنید که متغیر تصادفی X دارای توزیع ترکیبی به صورت زیر باشد:

$$p_\alpha(x) = \alpha * p_1(x) + (1-\alpha) * p_2(x)$$

فرض کنید دو توزیع p_1 و p_2 برای ما شناخته شده باشند و فقط متغیر α برای ما ناشناخته باشد. فرض کنید n نمونه داده i.i.d $\{x_1, x_2, ..., x_n\}$ از توزیع X را داشته باشیم. یک الگوریتم EM برای تخمین α ارائه دهید (قدم‌های E و M را در الگوریتم خود بصورت دقیق مشخص نمائید).

سوال ٦) فرض کنید که $Y_1 \sim \exp(1/\theta_1)$ و $Y_2 \sim \exp(1/\theta_2)$ مستقل از هم باشند و $\theta_1 <> \theta_2$. فرض کنید n نمونه داده i.i.d $\{x_1, x_2, ..., x_n\}$ از توزیع $X=Y_1+Y_2$ را داشته باشیم.

الف) توزیع X را بیابید.

نکته ۱: چگالی Y_1 برابر است با $f_{\theta_1}(y) = \theta_1 \exp(-\theta_1 y)$. برای Y_2 هم چگالی به همین صورت می‌باشد.

نکته ۲: ابتدا CDF را برای X به دست آورید، یعنی:

$$F(x) = P(Y_1 + Y_2 < x) = \int_0^x \int_0^{x-y_1} f_{\theta_1}(y_1) f_{\theta_2}(y_2) dy_2 dy_1$$

ب) برای تخمین MLE پارامترهای θ_1 و θ_2 الگوریتم EM مناسب ارائه کنید. قدم‌های E و M و عبارات بروزرسانی مربوطه را بصورت دقیق مشخص نمائید.

سوال ۷) فرض کنید x دارای توزیع Hat با میانگین μ باشد، یعنی:

$$p(x \mid \mu) = \begin{cases} 0 & x \le \mu - 1 \\ 1-(\mu - x) & \mu - 1 \le x \le \mu \\ 1-(x-\mu) & \mu \le x \le \mu + 1 \\ 0 & \mu + 1 \le x \end{cases}$$

فرض کنید داده‌های {1, 3, 6, 7} بوسیله ترکیبی از سه توزیع Hat تولید شده باشند و همچنین داشته باشیم:

$$P(w1)=1/2,\ P(w2)=1/4,\ P(w3)=1/4$$

سه توزیع فوق را طوری بیابید که بیشترین Likelihood را داشته باشیم.

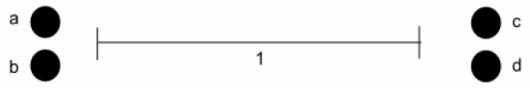
سوال ۸) یک کلاس از الگوریتم‌های خوشه‌بندی که اخیرا بسیار محبوب شده است، الگوریتم‌های خوشه‌بندی طیفی است. بسیاری از این الگوریتم‌ها از نظر پیاده‌سازی بسیار ساده و از نظر عملکرد نسبت به روش‌های سنتی مانند k-means بر روی مسائل مشخص خوشه‌بندی بهتر عمل می‌کنند. در این تمرین می‌خواهیم بررسی کنیم که چرا این روش‌ها جواب مناسبی تولید می‌کنند و در ادامه یکی از این الگوریتم‌ها را پیاده‌سازی می‌کنیم.

اگر یک مجموعه داده متشکل از m نقطه $x_1,\ldots,x_m$ داده شده باشد، ورودی الگوریتم خوشه‌بندی طیفی، یک ماتریس A است که شامل شباهت‌های دو به دوی بین این داده‌هاست. ماتریس A را ماتریس وابستگی می‌نامند. انتخاب اینکه چگونه شباهت بین نقاط را اندازه‌گیری کنیم، موضوعی است که به جنس مساله بستگی دارد. یک ماتریس وابستگی ساده را می‌توان با استفاده از رابطه زیر ایجاد کرد.

$$A(i,j) = A(j,i) = \begin{cases} 1 & d(x_i, x_j) < \theta \\ 0 & otherwise \end{cases}$$

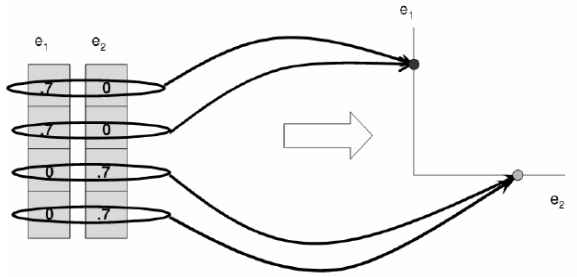
که در آن $d(x_i,x_j)$ نشان دهنده فاصله اقلیدسی بین نقاط x_i و x_j است. ایده عمومی در خوشه‌بندی طیفی این است که نگاشتی از نقاط داده در فضای ویژه ایجاد کند؛ با این امید که نقاط به صورت مناسبی در این فضای ویژه جدا می‌شوند و با اعمال یک روش ساده مانند k-means بر روی این نقاط جدید، می‌توانیم نتایج مناسبی به دست آوریم.

برای مثال، ایجاد ماتریس وابستگی برای مجموعه داده شکل زیر را در نظر بگیرید که از فاصله اقلیدسی با $\theta=1$ استفاده شده است. ماتریس به دست آمده (ماتریس A) در زیر شکل نمایش داده شده است. در این مثال خاص خوشه‌های $\{a,b\},\{c,d\}$ به صورت بلاک‌های غیر صفر در ماتریس وابستگی نشان داده شده‌اند. البته این نتیجه ساختگی است به این دلیل که ما می‌توانیم ماتریس A را با هر ترتیبی از {a,b,c,d} به دست آوریم. برای مثال ماتریس وابستگی ممکن دیگر 'A است.

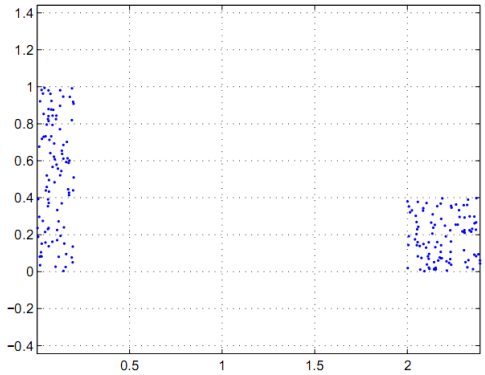


$$A=\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 1 & 1 & 0 & 0 \\ b & 1 & 1 & 0 & 0 \\ c & 0 & 0 & 1 & 1 \\ d & 0 & 0 & 1 & 1 \end{array}, \qquad A'=\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 1 & 0 & 1 & 0 \\ b & 0 & 1 & 0 & 1 \\ c & 1 & 0 & 1 & 0 \\ d & 0 & 1 & 0 & 1 \end{array}$$

نکته مهمی که وجود دارد این است که بردارهای ویژه ماتریس A و 'A مقادیر یکسانی دارند، فقط مکان آنها عوض شده است. بردار ویژه با مقادیر ویژه غیر صفر A، $e_1=(.7,.7,0,0)^T, e_2=(0,0,.7,.7)^T$ بوده و بردارهای ویژه با مقادیر ویژه غیر صفر 'A، $e_1=(.7,0,.7,0)^T, e_2=(0,.7,0,.7)^T$ هستند. خوشه‌بندی طیفی داده‌های اصلی را با استفاده از مختصاتی که بردارهای ویژه مشخص می‌کنند در فضای جدید قرار می‌دهد. یعنی در این روش نقطه x_i را به نقطه $\left(e_1(i),e_2(i),\ldots e_k(i)\right)$ که $e_1,e_2,\ldots e_k$، k بردار ویژه با مقادیر ویژه بزرگتر ماتریس A هستند نگاشت می‌کند.



در این سوال ما عملکرد یک نوع از روش‌های خوشه‌بندی طیفی را بر روی مجموعه داده ساده دیگری که در شکل زیر نشان داده شده است را بررسی و تجزیه و تحلیل می‌کنیم.



الف) فرض کنید در مجموعه داده فوق، خوشه اول m_1 نمونه و خوشه دوم m_2 نمونه داشته باشد. در این صورت چه مقداری را برای θ انتخاب می‌کنید؟ چرا؟

ب) گام بعدی محاسبه k بردار ویژه غالب ماتریس وابستگی است که k مشخص کننده تعداد خوشه‌هاست. برای مجموعه داده فوق و ماتریس وابستگی به دست آمده، آیا مقداری برای θ وجود دارد که بتوان به صورت تحلیلی دو مقدار ویژه و بردار ویژه متناظر آنها را محاسبه کرد؟ اگر وجود ندارد توضیح دهید؟ اگر وجود دارد، این مقادیر ویژه را محاسبه کنید. مقادیر ویژه دیگر دارای چه مقادیری هستند؟

پ) همانطور که گفته شد اکنون ما می‌توانیم نگاشت نقاط داده را با استفاده از k بردار ویژه با مقدار ویژه بیشتر را محاسبه کنیم. برای مجموعه داده فوق بهترین حدس را برای مختصات مراکز خوشه در k=2 را با استفاده از θ هایی که در قسمت (الف) مشخص شده است را به دست آورید.

خوشه‌بندی طیفی در عمل. در پیاده‌سازی خوشه‌بندی طیفی معمولا ماتریس وابستگی با استفاده از کرنل گاوسی ساخته می‌شود:

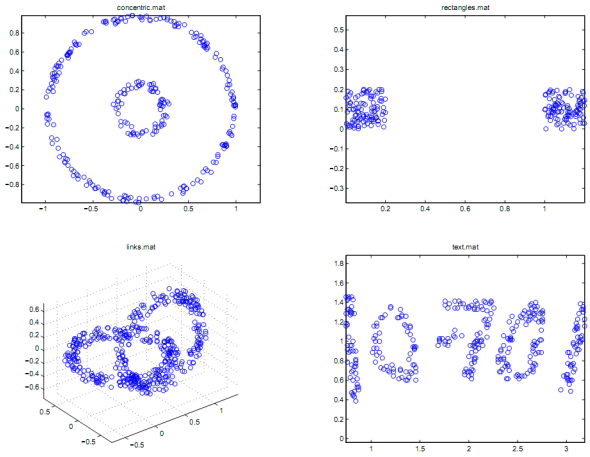
$$A(i,j) = exp\left(-\frac{d(x_i, x_j)^2}{\sigma}\right)$$

که σ پارامتری است که به وسیله کاربر مشخص می‌شود. بهترین ماتریس وابستگی که ما امیدواریم در عمل به آن برسیم یک ماتریس وابستگی نزدیک به ماتریس قطری-بلاکی است. می‌توان نشان داد که در این مورد، بعد از نگاشت به فضای پوشش داده شده به وسیله k بردار ویژه با مقادیر ویژه بزرگتر، نقاطی که متعلق به یک بلاک هستند با معیار فاصله اقلیدسی به یکدیگر نزدیکتر هستند. با توجه به این ایده، شما یکی از روش‌های الگوریتم‌های خوشه‌بندی طیفی را پیاده‌سازی خواهید کرد.

گام های این الگوریتم به صورت زیر هستند:

- ابتدا با استفاده از کرنل گاوسی ماتریس وابستگی A را ایجاد کنید.
- به صورت متقارن سطرها و ستون های ماتریس A را نرمال‌سازی کنید تا ماتریس N را به دست آورید به صورتی که $N(i,j) = \frac{A(i,j)}{\sqrt{d(i)d(j)}}$ و $d(i) = \sum_k A(i,k)$ است (به صورت ماتریسی $N = D^{-(1/2)} A D^{-(1/2)}$).
- ماتریس Y، که ستون‌های آن k بردار ویژه اول ماتریس N است، را ایجاد کنید.
- هر سطر از ماتریس Y را نرمال‌سازی کنید تا دارای طول واحد باشد.
- مجموعه نقاط نگاشت شده را با اجرای الگوریتم k-means خوشه‌بندی کنید (هر سطر Y یک نمونه از داده‌ها را مشخص می‌کند).

الف) الگوریتم k-means را بر روی ۴ مجموعه داده ضمیمه شده (نمایش داده شده در شکل زیر) اجرا کرده و نتایج را نمایش دهید. برای مجموعه داده text.mat، k=6 در نظر گرفته شده و برای بقیه مجموعه داده ها k=2 در نظر گرفته شود.



ب) الگوریتم خوشه‌بندی طیفی ابتدایی را پیاده‌سازی کرده و بر روی مجموعه داده‌های ضمیمه شده و با همان مقادیر k اجرا کنید و نتایج خوشه‌بندی را با $\theta = .025, .05, .2, .5$ نمایش دهید.

پ) با استفاده از نتایج به دست آمده، k-means و خوشه‌بندی طیفی را مقایسه کنید؟

تحلیل نظری: در الگوریتم بالا از ماتریس $N = D^{-(1/2)} A D^{-(1/2)}$ استفاده شد که در آن A ماتریس وابستگی است و $a_{ij} = a_{ji}$ یک مقدار نامنفی است که فاصله بین دو نقطه x_i و x_j را مشخص می کند و D یک ماتریس قطری است که عضو iام قطر اصلی آن d_{ii} برابر با حاصل جمع سطر iام ماتریس A است.

ت) نشان دهید که بردار $v_1 = \left[\sqrt{d_{11}}, \sqrt{d_{22}}, \ldots, \sqrt{d_{nn}}\right]^T$ بردار ویژه ماتریس N با مقدار ویژه $\lambda_1 = 1$ است.

ث) نشان دهید که $P^{\infty} = D^{-(1/2)} v_1 v_1^T D^{(1/2)}$. این ویژگی نشان می‌دهد که اگر نقاط به عنوان گره‌ها در گراف مارکوف با احتمال گذر متناسب با فاصله بین نقاط (عضوهای ماتریس A) در نظر گرفته شوند، v_1 تنها بردار ویژه مورد نیاز برای محاسبه توزیع احتمالاتی بر روی حالات ماتریس P^{∞} خواهد بود.