به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سری چهارم داده کاوی – بخش تئوری

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد و در صورت مشاهده هرگونه تقلب نمره صفر برای کل تمرین منظور خواهد شد.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- گزارش تمرین خود را در قالب یک فایل PDF با نام «**HW4_StudentNumber.pdf**» در سایت درس در مهلت معین بارگذاری نمایید.
- در صورت داشتن اشکال می‌توانید از طریق ایمیل **datamining.fall2020@gmail.com** با تدریس‌یاران درس در ارتباط باشید.
- همچنین لازم بذکر است که اگر مواردی در کلاس تدریس نشده انتظار می رود که خود دانشجویان جستجو کنند و انجام دهند.

نیم‌سال اول ۰۰-۹۹

سوال ۱- بردار ویژگی d بعدی x را داریم که می خواهیم برای دسته بندی آن را به یک بردار ویژگی به نام y با ابعاد کمتر تبدیل کنیم. شما برای این کار سه گزینه در دست دارید: PCA، LDA و انتخاب زیر مجموعه ویژگی[۱] با استفاده از درخت تصمیم (که یدین صورت عمل می کند که یک دسته بند درخت تصمیم بر روی x تعریف می شود که داده را براساس سوال/سوالاتی دسته بندی می کند که منجر به یافتن y می شود) فرض کنید درخت تصمیم فقط از یک سوال استفاده می کند(تک متغیره است). برای هر یک از سناریو های زیر بیان کنید که کدام یک از این سه روش بهترین است و چرا. (همچنین ممکن است روش بهینه، ترکیبی از سه روش گفته شده باشد)

توجه کنید: m تعداد کلاسها ، d بعد ویژگی و n تعداد داده های آموزشی(برچسب دار) در هر کلاس است.

الف) $m = 50, d = 20, n = 10, x \in \Re^d$

ب) $m = 5, d = 200, n = 1000$ در این حالت x شامل ویژگی های عددی و categorical است

ج) $m = 10, d = 50, n = 100, x \in \Re^d$ همچنین ده هزار داده بدون برچسب وجود دارد

سوال ۲- درباره الگوریتم kmeans++ تحقیق کنید و تفاوت آن را با الگوریتم kmeans شرح دهید. همچنین توضیح دهید که این الگوریتم چگونه سرعت همگرایی و کیفیت خوشه‌بندی را بهتر می‌کند.

سوال ۳- کدام یک از موارد زیر در مورد الگوریتم DBSCAN صحیح می‌باشد؟ علت درست یا غلط بودن را ذکر کنید.

الف) برای اینکه نقاط داده در یک خوشه قرار بگیرند، باید در یک فاصله آستانه ای از نقطه مرکزی (core point) قرار داشته باشند.

---

[۱] Feature subset selection

ب) پیچیدگی زمانی این الگوریتم از اوردر $O(n^3)$ می‌باشد.

ج) این الگوریتم نسبت به داده های پرت (outliers) مقاوم است.

د) این الگوریتم نیازی به دانستن تعداد خوشه ها برای خوشه بندی ندارد.

سوال ۴- جدول زیر نمایانگر ماتریس فواصل ۶ شی می باشد.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

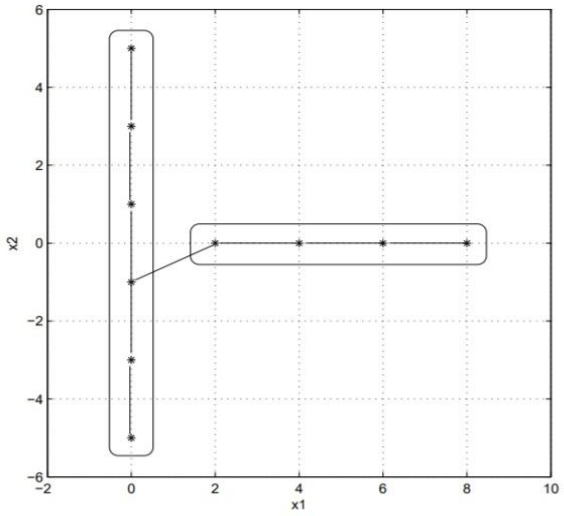
الف) نتیجه نهایی خوشه بندی سلسله مراتبی (Hierarchical clustering) با single link را با کشیدن dendrogram نمایش دهید.

ب) نتیجه نهایی خوشه بندی سلسله مراتبی با(complete link (MAX را با کشیدن dendrogram نمایش دهید.

توجه- در dendrogram رسم شده باید ترتیب ترکیب شدن نقاط و خوشه ها واضح باشد.

سوال ۵-

داده های نمایش داده شده در شکل زیر را در نظر بگیرید.



قصد داریم این نقاط را به وسیله الگوریتم خوشه بندی طیفی به دو دسته اصلی تقسیم بندی کنیم. در الگوریتم مورد استفاده، گراف همسایگی بر اساس اتصال هر داده به دو نزدیک ترین همسایه اش به دست می آید.
همچنین وزن دهی یالهای نتیجه شده میان نقاط xi و xj به صورت زیر می باشد:

$$Wij = exp(-||xi - xj||^2)$$

الف) نشان دهید خوشه های نهایی به چه صورت خواهند بود.

ب) آیا این نتیجه به وسیله الگوریتم خوشه بندی k-means نیز قابل مشاهده خواهد بود؟ چرا؟