تکلیف ۳ یادگیری ماشین

مهدی کاظمی ۹۹۲۱۰۷۵۳

۱.۱

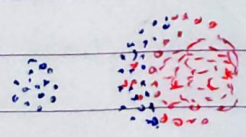
آ) بله، الگوریتم k-means حالت خاصی از خوشه‌بندی با GMM است، طوریکه تمام واریانس‌ها یکسان باشند و کوواریانس‌ها صفر و mixture weight ها هم برابر باشند. پیش‌فرض k-means این است که تمامی کلاستر‌ها چنین هستند و این در بسیاری از موارد پاسخ بهینه را نمی‌دهد. در Gaussian Mixture برای تخمین پارامترها معمولاً از MLE استفاده می‌کنیم که برای داده‌های $D = x_1, x_2, \dots, x_n$ می‌خواهیم پارامترهای (π_k, μ_k, Σ_k) را به ازای هر cluster از ۱ تا k تخمین بزنیم و از رابطه زیر استفاده می‌کنیم:

$$\log P(D \mid \pi, \mu, \Sigma) = \sum_i \log \sum_k P_\pi(z_i = k) P_{\mu,\Sigma}(x_i \mid z_i = k)$$

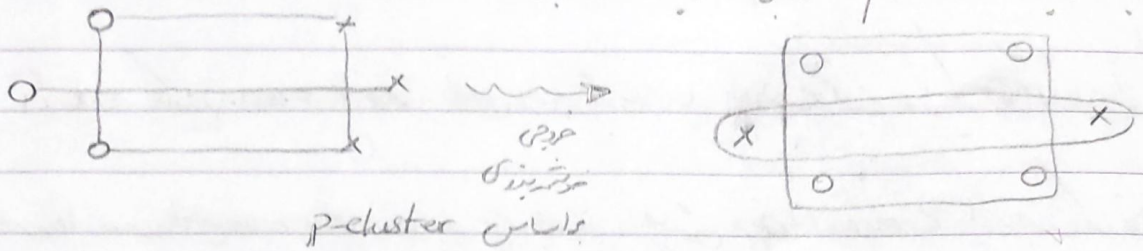
در اینجا اگر واریانس را به صورت $\Sigma_k = \sigma_k I$ در نظر بگیریم یعنی فرض کرده‌ایم که تمامی کلاستر‌ها به شکل دایره باشند و این EM با k خوشه را به k-means تبدیل کرده‌ایم.

۱.۱ ب) در خوشه‌بندی سلسله‌مراتبی با روش محاسبه شباهت به صورت Max یا Complete linkage ممکن است که داده‌ای در یک خوشه به داده‌ای در خوشه دیگر نزدیکتر از داده‌ای در خوشه خودش باشد زیرا که این روش قایل بود که کلاستر‌های بزرگ یا پیچیده و همینطور به outlier ها اهمیت بسیاری می‌دهد؛ دوری به این نوع خطاها در

خوشه‌بندی نشان داده شده است:



۱.۱ ج) این نوع از خوشه بندی با توجه به توضیحات مساله به نظر می رسد که به دلیل استفاده نکردن از معیار فاصله نمی تواند نتیجه خوبی در خروجی برهد زیرا که اگر داده ها به صورت زیر باشند که x ها یک دسته و o ها یک دسته دیگر باشند و بال ها به صورت زیر باشند با P=100 هم خروجی مطلوب نخواهیم بود.



۱.۲)

الف) درصورت Single-linkage در ابتدا نقاط داخل هر دسته با یکدیگر کلاستر تشکیل می دهند زیرا که بسیار به یکدیگر نزدیک اند

سپس نقاط بین دسته های (3و4) و (1و2) باید در کلاستر ها قرار بگیرند که با روش کوتاهترین فاصله دو دسته 3و4 با یکدیگر و دو دسته از 2 نیز با یکدیگر یک کلاستر را تشکیل می دهند ← کلاستر های نهایی: ① (1و2) و ② (3و4)

• درصورت Complete-linkage در ابتدا نقاط درون هر یک از دسته های 1 تا 4 با یکدیگر یک خوشه می شوند و سپس در نهایت ~ حالتی می بینیم که فاصله دورترین نقطه از خوشه 1 تا دورترین نقطه در خوشه 2 بیشتر از بیشترین فاصله بین خوشه 1 و 3 می شود

بنابراین دو خوشه (1و3) و (2و4) ایجاد می شوند ← کلاستر های نهایی: ① (1و3) ، ② (2و4)

• درصورت averge-linkage پس از اینکه اعضای داخل کلاستر ها با یکدیگر تشکیل خوشه دادند میانگین فاصله بین خوشه ها بین (3و1) کمتر از میانگین بین (2و1) خواهد شد بنابراین کلاستر ها به صورت ① (1و3) و ② (2و4) می شوند

ب) در شکل (b) معیار Single-link می‌تواند موفق عمل کند زیرا که در ابتدا نقاطی که در هر یک از خوشه‌ها هستند بسیار به یکدیگر نزدیک با یکدیگر یک خوشه تشکیل داده و سپس این خوشه‌ها گسترش می‌یابند تا کل هر یک از ۲ خوشه را بپوشانند و این صورت به نحوی هر کدام یک خوشه جداگانه می‌شوند.

در شکل (c) به دلیل وجود نویز بین دو دسته معیار Single-link عملکرد خوبی ندارد زیرا ممکن است دو دسته با یکدیگر ترکیب کنند و این دو معیار Complete-link و average-link یا average-linkage بهتر عمل خواهند کرد.

(۱.۴

$$\left.\begin{aligned}\mu &= mean(C)\\ \mu' &= mean(C')\\ \bar{\mu} &= mean(C \cup C')\end{aligned}\right\} \Rightarrow cost(C \cup C') - cost(C) - cost(C')$$

$$= \sum_{x \in C \cup C'} \|x-\bar{\mu}\|^2 - \sum_{x \in C} \|x-\mu\|^2 - \sum_{x \in C'} \|x-\mu'\|^2$$

$$= \sum_{x \in C} (\|x-\bar{\mu}\|^2 - \|x-\mu\|^2) + \sum_{x \in C'} (\|x-\bar{\mu}\|^2 - \|x-\mu'\|^2)$$

$$= |C| \cdot \|\mu-\bar{\mu}\|^2 + |C'| \cdot \|\mu'-\bar{\mu}\|^2$$

$$\bar{\mu} = (|C|\mu + |C'|\mu')/(|C|+|C'|) \Rightarrow = |C| \cdot \left\| \frac{|C'|}{|C|+|C'|}(\mu'-\mu) \right\|^2 + |C'| \cdot \left\| \frac{|C|}{|C|+|C'|}(\mu-\mu') \right\|^2$$

$$= \frac{|C| \cdot |C'|}{|C|+|C'|} \|\mu-\mu'\|^2$$

c1.F

Single-linkage c1

	A	B	C	D	E	F
A	0					
B	√0.12	0				
C	0.57	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

	A,B	C	D	E	F
A,B	0				
C	0.25	0			
D	0.16	√0.14	0		
E	0.28	0.70	0.45	0	
F	0.34	0.93	0.20	0.67	0

	A,B	C,D	E	F
A,B	0			
C,D	√0.16	0		
E	0.28	0.45	0	
F	0.34	0.20	0.67	0

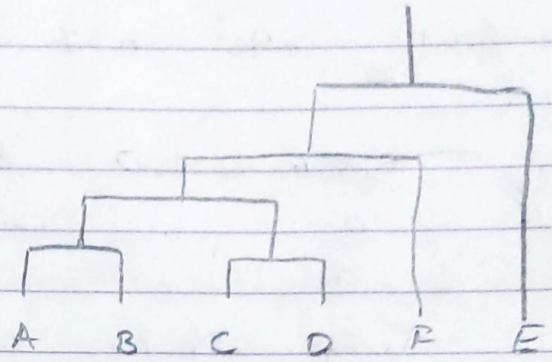
	A,B,C,D	E	F
A,B,C,D	0		
E	0.28	0	
F	√0.20	0.67	0

s.a.m

	A,B,C,D,F	E
A,B,C,D,F	0	
E	0.28	0

	A,B,C,D,F,E
A,B,C,D,F,E	0

Dendrogram:



Complete linkage

	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

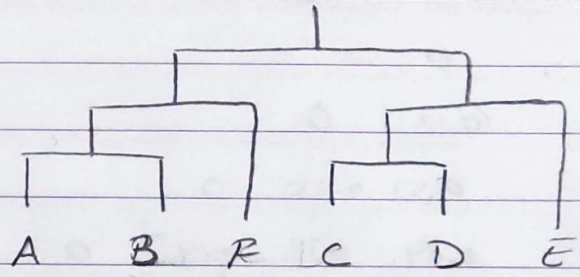
	A,B	C	D	E	F
A,B	0				
C	0.51	0			
D	0.84	0.14	0		
E	0.77	0.70	0.45	0	
F	0.61	0.93	0.20	0.67	0

	A,B	C,D	E	F
A,B	0			
C,D	0.84	0		
E	0.77	0.70	0	
F	0.61	0.93	0.67	0

	A,B,F	C,D	E
A,B,F	0		
C,D	0.93	0	
E	0.77	0.70	0

	A,B,F	C,D,E
A,B,F	0	
C,D,E	0.93	0

Dendrogram:



ج) با تغییر ۲ فاصله A و F از 0.34 به 0.15 و تغییر فاصله D و E از 0.45 به 0.15، dendrogram تغییر می‌کند.

هر دو روش مانند نمودار قسمت ب خواهد شد. (در اینصورت که در صورت برابری ۲ فاصله در ماتریس فواصل، فاصله‌ای که زودتر محاسبه شده است انتخاب شود به عنوان فاصله کمینه برای یکی کردن آن سطر و ستون).

سوال ۲:

۱.۲) TNoM یک روش Non-parametric است و در واقع به صورت پیش فرض داده‌ها از توزیع خاصی نمی‌آیند و روش کار آن به این صورت است که فرض می‌کنیم که متغیرهای ژن را داشته باشیم و بخواهیم ارتباط ژن g با هر یک از label های + و یا - را بسنجیم. در اینصورت اگر فرض کنیم که a تا بافت داریم که label مثبت خورده‌اند و b تا بافت داریم که label منفی خورده‌اند، در اینصورت بردار n را اینصورت تعریف می‌کنیم که میزان بیان ژن g در هر یک از بافت‌ها به صورت مرتب شده از کم به زیاد است $n = \langle n_1, \ldots, n_{a+b} \rangle$ که n_1 کمترین و n_{a+b} بیشترین بیان و است. سپس بردار rank به اینصورت تعریف می‌شود که در همان جایگاه‌های بردار n اگر بیان متعلق به بافت گروه a بود مقدار + و اگر متعلق به b بود مقدار - را قرار می‌دهیم. مثلاً اگر بیان ژن در بافت‌های a به صورت $\{10, 20, 30, 50, 80\}$ و در بافت‌های گروه b به صورت $\{40, 70\}$ باشد، بردار رنک v به صورت روبرو خواهد بود $v = \{+,+,+,-,+,-,+\}$. سپس به ازای تمامی پارتیشن‌های x و y ای که می‌توانیم روی v ایجاد کنیم یک score محاسبه می‌کنیم که در نهایت کمینه این مقادیر برابر score TNoM خواهد بود. حال score‌ها به اینصورت محاسبه می‌شوند که به ازای ۲ پارتیشن x و y یک بار فرض می‌کنیم که x ها کلاس + و y کلاس منفی و بار دیگر برعکس این فرض می‌کنیم و در هر بار تعداد داده‌های از خانواده دیگر در پارتیشن را می‌شماریم و بین دو عددی که برای ۲ نوع تفکیک ایجاد می‌شود min می‌گیریم و سپس بین تمامی مقادیر پارتیشن‌ها min می‌گیریم که همان TNoM Score می‌شود:

$$TNoM(v) = \min_{x;y=v} \min\left([\#_-(x) + \#_+(y)], [\#_+(x) + \#_-(y)]\right)$$

تعیین و پارتیشن کردن ۲ باید انتخاب یک مقدار Threshold برای بیان روش و شمردن تعداد misclassification ها است که Threshold Number of Misclassification پیشنهاد نام این موضوع بود.

۲.۲) با انتخاب ویژگی از بین ویژگی هایی که داریم و برای ما معنی دار بودند نهایت ویژگی هایی بدیع معنی دار و قابل تفسیر هستند و به طور مثال می توانیم متوجه شویم که هرکدام از ویژگی ها چه تغییراتی روی خروجی دارند و این روش های استخراج ویژگی، ترکیبی از ویژگی ها را به دست آوردیم که برای ما معنی دار و قابل تفسیر باشد بنابراین نسبت روش های انتخاب ویژگی، تفسیرپذیر بودن ویژگی ها است نسبت به روش های استخراج ویژگی.

۳.۲)

Pearson correlation : عیب اصلی این روش این است که فقط حساس به روابط خطی است و آن ها را تشخیص می دهد به طریقه ای فرض کنیم این معیار را برای x^2 در نظر بگیریم هرچند که انطباق یک به یک برقرار است ولی P. Correlation مقدار کمی نزدیک به صفر برمی گرداند؟ بنابراین توصیه می شود که تنها به معیار P. correlation بسنده نکنیم و همواره به دنبال راه های دیگر نیز بگردیم.

Mutual Information : دو مشکل اصلی این روش این است که در ابتدا این روش یک metric نیست و همچنین normalized نیست یعنی که مقادیرش محدوده یک بازه خاصی قرار نمی گیرند و بنابراین برای مقایسه بین دو dataset مناسب نیست. مشکل بعدی این است که برای مقادیر پیوسته مناسب نیست و نیاز است که این مقادیر را bin بندی کنیم که در نتیجه معیار بسیار وابسته به bin بندی می شود.

s.a.m

سوال ۳)

۳.۴) بدین روش $V = WH$ ، که V بردارهای ماهیتی W بردارهای تبدیل یافته و ماتریس H ماتریس

ضرایب است؛ بنابراین بردارهای تبدیل یافته ماتریس W حاصل می‌شوند.