یادگیری ماشین

بهار ۱۴۰۳

استاد: فاطمه سیدصالحی

دانشگاه صنعتی شریف

مهلت ارسال: ۲۶ خرداد

تمرین پنجم

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- موضوعات تمرین: **کاهش ابعاد - خوشه‌بندی - EM**

سوالات نظری (۸۰ نمره)

۱. (۲۰ نمره) ۳ داده زیر را در فضای ۲ بعدی در نظر بگیرید.

$$(-۱, -۱), (۰, ۰), (۱, ۱)$$

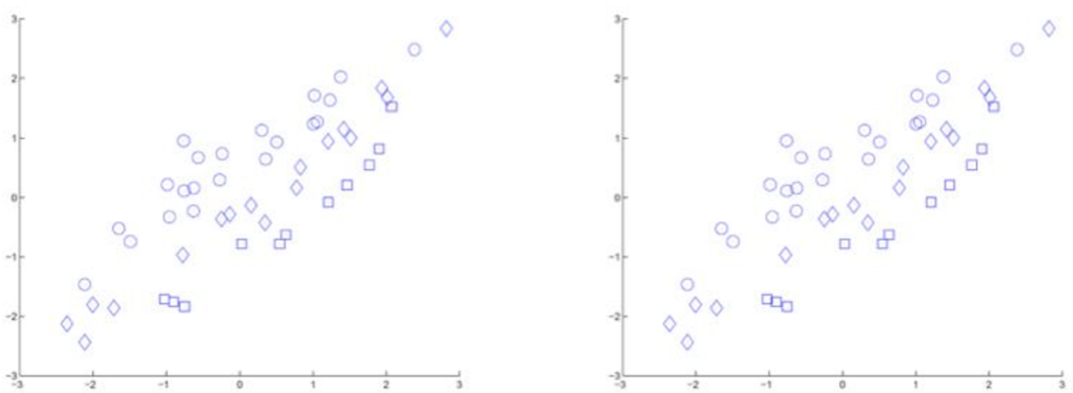
الف) مولفه اصلی اول (first principal component) را به‌دست آورید (بردار آن را بنویسید).

ب) اگر داده‌ها را با مولفه اصلی روی فضای یک بعدی تصویر کنیم، مختصات آن‌ها در فضای یک بعدی چیست؟ واریانس داده‌های تصویر شده را به‌دست آورید.

پ) اگر داده‌های تصویر شده را دوباره به فضای دوبعدی برده و داده‌های اصلی را بازسازی کنیم. خطای بازسازی چه قدر است؟

۲. (۲۰ نمره) برای داده‌های تصویر سمت چپ در شکل ۱، جهت مولفه اصلی اول بدون توجه برچسب‌های داده‌ها رسم کنید (PCA به برچسب‌های داده‌ها توجه نمی‌کند).

برای داده‌های سمت راست در شکل ۱، جهت کاهش خطی فیشر (Fisher Linear Decrement) را رسم کنید (برای اینکار نقاط دایره را به عنوان کلاس مثبت و نقاط مربع و لوزی را به عنوان کلاس منفی در نظر بگیرید).



شکل ۱: داده‌های تصویر شده در دو بعد

۳. (۲۰ نمره) در این مسئله می‌خواهیم به بررسی الگوریتم خوشه‌بندی K-means بپردازیم. فرض کنید $X = \{x_۱, x_۲, \ldots, x_n\}$ داده‌های ما باشند و γ یک ماتریس indicator باشد به این صورت که $\gamma_{ij} = ۱$ اگر x_i متعلق به خوشه j ام باشد و در غیر این صورت برابر ۰ است. فرض کنید $\mu_۱, \ldots \mu_k$ میانگین خوشه‌ها باشند.

اعوجاج J برای داده‌ها به صورت زیر محاسبه می‌شود:

$$J(\gamma, \mu_۱, \ldots, \mu_k) = n \sum_{j=۱}^{k} \gamma_{ij} ||x_i - \mu_j||^۲$$

همچنین $C = ۱, \ldots, k$ را به عنوان مجموعه‌ی خوشه‌ها در نظر بگیرید.

با توجه به معمول‌ترین الگوریتم K-means که در کلاس تدریس شده، به سوالات زیر پاسخ دهید.

الف) نشان دهید که الگوریتم در تعداد متناهی قدم به پایان می‌رسد. (γ چند مقدار متفاوت می‌تواند بگیرد؟)

ب) فرض کنید $\hat{x}$ میانگین داده‌های نمونه باشد. مقادیر زیر را در نظر بگیرید.

$$T(X) = \frac{\sum_{i=۱}^{n} ||x_i - \hat{x}||^۲}{n}$$

$$W_j(X) = \frac{\sum_{i=۱}^{n} \gamma_{ij} ||x_i - \mu_j||^۲}{\sum_{i=۱}^{n} \gamma_{ij}}$$

$$B(X) = \sum_{j=۱}^{k} \frac{\sum_{i=۱}^{n} \gamma_{ij}}{n} ||\mu_j - \hat{x}||^۲$$

در اینجا، $T(X)$ نشان دهنده‌ی انحراف کلی، $W_j(X)$ انحراف درون خوشه‌ای و $B(X)$ انحراف بین خوشه‌ای است.

رابطه بین این ۳ مقدار به چه صورت است؟

نشان دهید که K-means می‌تواند به عنوان کمینه کننده میانگین وزن‌دار مقادیر درون خوشه‌ای و به طور تقریبی بیشینه کردن انحراف بین خوشه‌ای دیده شود.

پ) نشان دهید که کمینه‌ی J یک تابع غیرافزایشی بر حسب k یا همان تعداد خوشه‌هاست. بنابراین چرا انتخاب تعداد خوشه‌ها با کمینه کردن J کاری بی‌معنیست؟

۴. (۲۰ نمره) دوست شما دو سکه دارد: یک سکه قرمز و یک سکه آبی با بایاس‌های p_r و p_b . (سکه قرمز با احتمال p_r و سکه آبی با احتمال p_b رو می‌آید). او همچنین یک ترجیح π برای انتخاب سکه قرمز دارد. او m پرتاب سکه انجام می‌دهد. برای هر پرتاب او در ابتدا سکه قرمز را با احتمال π یا سکه آبی را با احتمال $۱ - \pi$ انتخاب می‌کند. این پروسه برای هر پرتاب به صورت مستقل از دیگر پرتاب‌ها انجام می‌شود. ما از انتخاب او اطلاعی نداشته و فقط خروجی پرتاب‌ها (رو آمدن یا نیامدن) را مشاهده می‌کنیم.

برای هر پرتاب i ، متغیر تصادفی X_i به صورت زیر تعریف می‌شود.

$$X_i = \begin{cases} ۱ & \text{اگر در i امین پرتاب رو آمده باشد.} \\ ۰ & \text{otherwise.} \end{cases}$$

بنابراین داده قابل مشاهده برای ما مقادیر $x_1, \dots, x_m$ است که از m متغیر تصادفی به‌دست آمده‌اند. با توجه به این داده ما می‌خواهیم مقادیر پارامترهای $\theta = (\pi, p_r, p_b)$ را تخمین بزنیم. برای کمک به این موضوع برای هر پرتاب i یک متغیر تصادفی مشاهده نشده Z_i به صورت زیر در نظر میگیریم:

$$Z_i = \begin{cases} 1 & \text{اگر در i امین پرتاب از سکه قرمز استفاده شده باشد} \\ 0 & \text{otherwise.} \end{cases}$$

الف) یک عبارت برای توزیع احتمال توام X و Z به صورت زیر به‌دست آورید.

$$p(x, z; \theta) = \dots$$

ب) یک عبارت برای complete-data log-likelihood بنویسید. $(L_c(\theta) = \sum_{i=1}^{m} lnp(x_i, z_i; \theta))$

پ) فرض کنید که مقادیر z_i را می‌دانید. تخمین بیشنه درست‌نمایی برای پارامترهای $\hat{\theta}$ را به‌دست آورید. عباراتی برای $\hat{\pi}, \hat{p_r}, \hat{p_b}$ بر حسب x_i, z_i بنویسید.

ت) با ندانستن مقادیر z_i از الگوریتم EM برای تخمین مقادیر پارامترها استفاده می‌کنیم. فرض کنید الگوریتم با مقادیر اولیه θ^0 شروع شده و θ^t نشان‌دهنده مقادیر پارامترها در شروع تکرار t ام است. در E-step الگوریتم نیاز به محاسبه $P(Z_i = 1 | X_i = x_i; \theta^t)$ دارد. این مقدار را به دست آورید.

ث) برای هر پرتاب i مقدار محاسبه شده در بخش قبل را با γ_i^t نمایش می‌دهیم. بنابراین امید ریاضی log-complete-data likelihood به صورت زیر خواهد بود:

$$\sum_{i=1}^{m} (\gamma_i^t . \ln p(x_i, 1; \theta) + (1 - \gamma_i^t) . \ln p(x_i, 0; \theta))$$

در M-step میخواهیم θ^{t+1} را طوری تخمین بزنیم که امید ریاضی complete-data log-likelihood بیشینه شود. پارامترهای آپدیت شده θ^{t+1} را مشخص کنید. (عبارتی برای $\pi^{t+1}, p_r^{t+1}, p_b^{t+1}$ بر حسب x_i و γ_i^t به‌دست آورید).

سوالات عملی (۴۰ نمره)

به نوتبوک های ضمیمه شده تمرین مراجعه کنید.