دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین
پاییز و زمستان ۱۴۰۴
استاد: علی شریفی زارچی
گردآورندگان: حسن‌بیگی - رحمانی - غریبی - هدایی - یوسف‌نیا

تمرین دوم | یادگیری بدون نظارت | مهلت ارسال: ۱۴ آذر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم، برای هر تمرین می‌توانید تا ۵ روز تأخیر مجاز داشته باشید و در مجموع حداکثر ۱۵ روز تأخیر مجاز خواهید داشت. توجه داشته باشید که تأخیر در تمرین‌های عملی و تئوری به صورت جداگانه محاسبه می‌شود و مجموع تأخیر هر دو نباید بیشتر از ۱۵ روز شود. پس از اتمام زمان مجاز، دو روز اضافی برای آپلود غیرمجاز در نظر گرفته شده است که در این بازه، به ازای هر ساعت تأخیر، ۲ درصد از نمره تمرین کسر خواهد شد.
- 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.

سوالات نظری (۱۰۰ نمره)

۱. (۱۵ نمره) در هر مورد انتخاب خود را همراه با استدلال مختصر بیان کنید.

- کدام یک از موارد زیر مزایای معمولی هستند که ممکن است شما را ترغیب کند تا قبل از آموزش یک طبقه‌بندی‌کننده Classifier، داده‌ها را با PCA پیش‌پردازش کنید؟ (گزینه‌های صحیح را انتخاب کنید)

(آ) PCA تمایل دارد بایاس الگوریتم طبقه‌بندی شما را کاهش دهد.
(ب) از PCA می‌توان برای جلوگیری از بیش‌برازش (Overfitting) استفاده کرد.
(ج) PCA تمایل دارد واریانس الگوریتم طبقه‌بندی شما را کاهش دهد.
(د) از PCA می‌توان برای جلوگیری از کم‌برازش (Underfitting) استفاده کرد.

- گزاره‌های صحیح در مورد تحلیل مؤلفه‌های اصلی PCA را انتخاب کنید.

(آ) PCA یک الگوریتم خوشه‌بندی است.
(ب) PCA ویژگی‌هایی تولید می‌کند که ترکیبات خطی از ویژگی‌های ورودی هستند.
(ج) مؤلفه‌های اصلی طوری انتخاب می‌شوند که واریانس داده‌ها را به حداکثر برسانند.
(د) مختصات اصلی همان مقادیر ویژهٔ ماتریس کوواریانس نمونه هستند.

- گزاره‌های صحیح دربارهٔ خوشه‌بندی k-means را انتخاب کنید.

(آ) k-means تضمین می‌کند که خوشه‌هایی را پیدا کند که تابع هزینه را کمینه کنند.
(ب) در خروجی، هر دو خوشه توسط یک مرز تصمیم خطی جدا شده‌اند.
(ج) امکان کرنلایز کردن k-means وجود ندارد.
(د) از دید آماری بهینه‌سازی k-means را با فرض پیشین گوسی روی میانگین‌ها و با استفاده از تخمین حداکثر درست‌نمایی توجیه می‌کنند.

- گزاره‌های صحیح دربارهٔ زمان اجرای الگوریتم k-means برای n نقطه با d ویژگی را انتخاب کنید.

(آ) مرحله به‌روزرسانی میانگین خوشه‌ها می‌تواند در زمان $\mathcal{O}(nd)$ اجرا شود.
(ب) افزایش k همیشه زمان اجرا را افزایش می‌دهد.
(ج) مرحلهٔ تخصیص خوشه‌ها می‌تواند در زمان $\mathcal{O}(nkd)$ اجرا شود.
(د) الگوریتم k-means حداکثر در زمان $\mathcal{O}(nkd)$ اجرا می‌شود.

۲. (۲۰ نمره) به سوالاتی که در ادامه آمده‌اند با بیان توضیحات کامل پاسخ دهید.

(آ) ثابت کنید الگوریتم k-means همگرا می‌شود.

(ب) با یک مثال عددی نشان دهید که الگوریتم k-means نسبت به داده‌های outlier مقاوم نیست. سپس با ارائه یک اصلاح ad-hoc سعی کنید مقاومت این روش را در شرایط حضور outlierها بهبود دهید. در رابطه با مزایا و معایب این تغییر بحث کنید.

(ج) الگوریتم‌های k-means و DBSCAN را از حیث مقاومت نسبت به داده‌های outlier مقایسه کنید.

۳. (۱۵ نمره) مجموعه نمونه زیر از شش نقطه $\mathbf{X}_i \in \mathbb{R}^2$ را در نظر بگیرید:

$$\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}2\\1\end{pmatrix}, \begin{pmatrix}2\\2\end{pmatrix}\right\}$$

(آ) میانگین نقاط نمونه را محاسبه کرده و ماتریس طرح متمرکز (Centered Design Matrix) را بنویسید.

(ب) تمام مؤلفه‌های اصلی را بیابید و آن‌ها را به صورت بردار واحد بنویسید.

(ج) اگر فقط از یک مؤلفه استفاده کنیم، کدام ترجیح دارد؟ الگوریتم PCA از چه اطلاعاتی برای ترجیح یکی بر دیگری استفاده می‌کند؟ از منظر بهینه‌سازی، چرا آن یکی را ترجیح می‌دهیم؟

(د) بردار تصویر هر یک از نقاط (نه نقاط متمرکز شده) را بر روی مؤلفهٔ اصلی ترجیحی محاسبه کنید.

۴. (۲۰ نمره) به سوالاتی که در ادامه آمده‌اند با بیان توضیحات کامل پاسخ دهید.

(آ) در دسته‌ای از مسائل تخمین پارامتر، با مدل‌هایی رو به رو هستیم که خروجی‌های آن، علاوه بر مجموعه‌ای از نمونه‌های مشاهده‌شده ($\mathbf{X}$)، شامل مجموعه‌ای از اطلاعات نهان ($\mathbf{Z}$) نیز می‌باشد. حضور پارامتر جدید $\mathbf{Z}$ به عنوان یک مجهول، باعث می‌شود تکنیک‌هایی نظیر MLE در این مسائل قابل استفاده نباشد. الگوریتم EM برای پاسخ به این چالش، مسئله را به دو مرحله تقسیم می‌کند. در هر گام t:

۱. ابتدا پارامترهای توزیع $\boldsymbol{\theta}^{(t)}$ ثابت در نظر گرفته می‌شود و امید ریاضی تابع log likelihood نسبت به توزیع شرطی $\mathbb{P}(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)})$ به دست می‌آید.
۲. سپس همانند روش‌های پیشین تخمین پارامتر، پارامترهایی که مقدار به‌دست‌آمده را بیشینه کنند، به عنوان پارامترهای جدید انتخاب می‌شوند.

یکی از مهم‌ترین کاربردهای این مدل‌سازی، تخمین پارامترهای توزیع‌های مخلوط است؛ به این صورت که تعلق نمونه‌ها به هر مؤلفه توسط پارامتر نهان $\mathbf{Z}$ مدل‌سازی می‌شود.
توزیع مخلوط زیر را در نظر بگیرید:

$$\mathbb{P}(X) = \alpha\lambda_1 e^{-\lambda_1 X} + (1-\alpha)\lambda_2 e^{-\lambda_2 X}$$

فرض کنید نمونه‌های $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \mathbb{P}(X)$ را در اختیار داریم و می‌خواهیم با EM پارامترهای $\alpha, \lambda_1, \lambda_2$ را تخمین بزنیم. اگر متغیر نهان Z_i به صورت زیر تعریف شود:

$$Z_i = \begin{cases} 0 & X_i \text{ comes from the first component of } \mathbb{P}(X) \\ 1 & X_i \text{ comes from the second component of } \mathbb{P}(X) \end{cases}$$

تابع complete-data log likelihood که به شکل $\log \mathbb{P}(X_1, Z_1, \ldots, X_n, Z_n \mid \lambda_1, \lambda_2, \alpha)$ تعریف می‌شود را بیابید.

(ب) امید ریاضی تابع complete-data log likelihood را نسبت به متغیرهای نهان محاسبه کنید (در محاسبات خود به مقدار $\mathbb{E}\big[Z_i \mid X_i, \alpha^{(t)}, \lambda_1^{(t)}, \lambda_2^{(t)}\big]$ نیاز خواهید داشت و باید آن را محاسبه کنید)

(ج) روابط به‌روزرسانی پارامترهای λ_1 و α را استخراج کنید. (برای سادگی، مقدار مرحلهٔ قبل را $\gamma_i^{(t)}$ بنامید.)

۵. (۱۵ نمره) برای هر یک از خوشه‌بندی‌های ارائه‌شده روی نقاط دوبعدی (هر کدام با $k = ۲$ خوشه) که در آن نقاط قرمز مربوط به خوشهٔ اول و نقاط آبی مربوط به خوشهٔ دوم هستند، مشخص کنید کدام‌یک از الگوریتم‌های زیر می‌توانستند منجر به این تخصیص خوشه‌ای شده باشند:

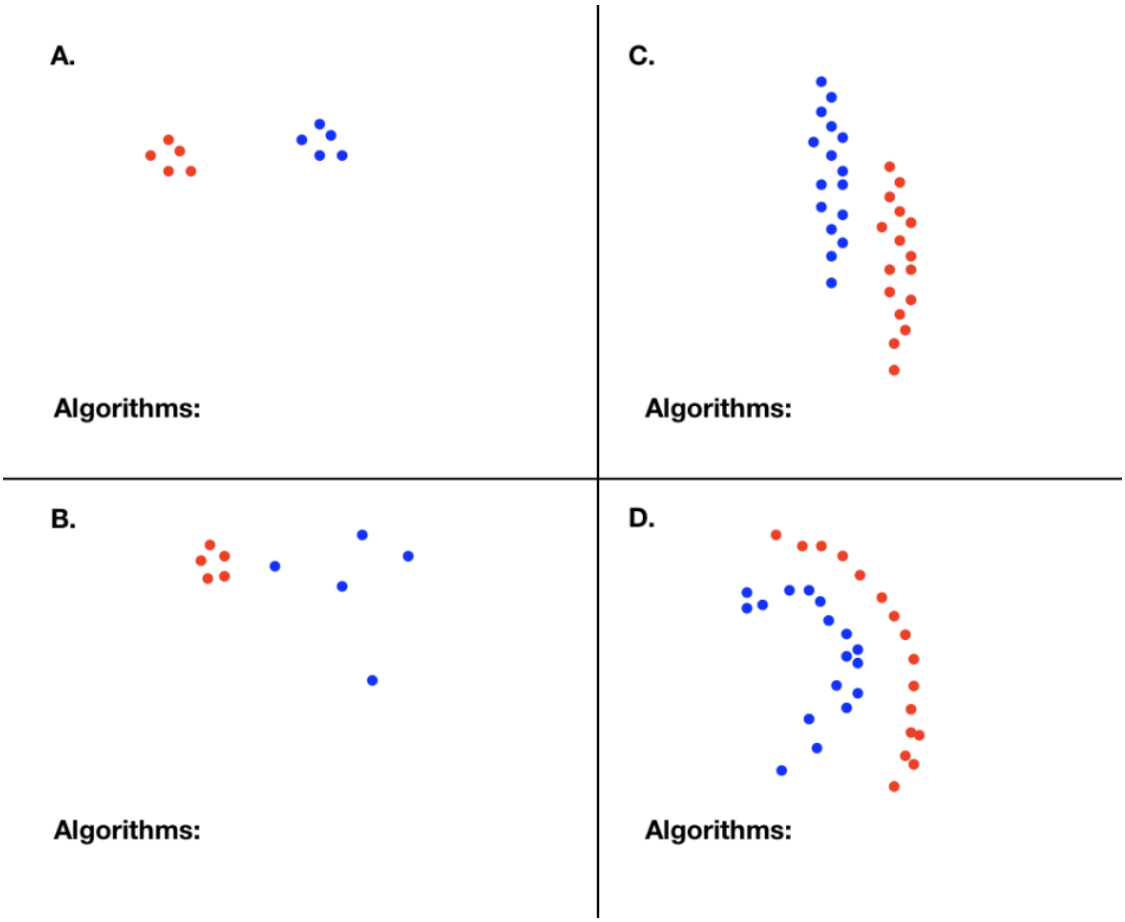
(آ) الگوریتم k-means

(ب) خوشه‌بندی single-link

(ج) مدل‌های مخلوط گاوسی Gaussian Mixture Models (GMM)

(د) هیچ‌یک از موارد فوق

ممکن است بیش از یک پاسخ صحیح باشد؛ تمام گزینه‌هایی را که ممکن است منجر به این خوشه‌بندی شده باشند با بیان استدلال ذکر کنید.



۶. (۱۵ نمره) می‌دانیم که نمایش‌های کم‌بعد برای مجموعه داده $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$ در PCA به صورت زیر تعریف می‌شوند:

$$\mathbf{z}_i = \mathbf{U}^\top (\mathbf{x}_i - \boldsymbol{\mu})$$

که در آن $\mathbf{U}$ ماتریسی است که سطرهای آن شامل k بردار ویژهٔ بالایی ماتریس کوواریانس هستند و

$$\boldsymbol{\mu} = \frac{۱}{n} \sum_{i=۱}^{n} \mathbf{x}_i$$

می‌باشد. نشان دهید که میانگین نمایش‌های کم‌بعد برابر صفر است:

$$\bar{\mathbf{z}} = \frac{۱}{n} \sum_{i=۱}^{n} \mathbf{z}_i = ۰$$