دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

پاییز ۱۴۰۳

علیرضا میررکنی - ۴۰۱۱۰۶۶۱۷

تمرین دوم

پاسخ مسئله‌ی ۱.

الف) در ابتدا دقت کنید که برای محاسبه بردار میانگین، داریم:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

پس در اینجا خواهیم داشت:

$$\begin{aligned}\bar{x} &= \frac{1}{4}\sum_{i=1}^{4} x_i \\ &= \frac{1}{4}\left(\begin{bmatrix}4\\1\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}2\\5\end{bmatrix} + \begin{bmatrix}3\\4\end{bmatrix}\right) \\ &= \frac{1}{4}\begin{bmatrix}10\\10\end{bmatrix} = \begin{bmatrix}\frac{5}{2}\\\frac{5}{2}\end{bmatrix}\end{aligned}$$

در ادامه دقت کنید که برای به دست آوردن ماتریس کوواریانس با استفاده از بردار میانگین، داریم:

$$\mathrm{Cov}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

بنابراین در اینجا می توان نوشت:

$$\begin{aligned}\mathrm{Cov}(X) &= \frac{1}{3}\sum_{i=1}^{4}(x_i - \bar{x})(x_i - \bar{x})^T \\ &= \frac{1}{3}\left(\begin{bmatrix}\frac{3}{2}\\-\frac{3}{2}\end{bmatrix}\begin{bmatrix}\frac{3}{2}\\-\frac{3}{2}\end{bmatrix}^T + \begin{bmatrix}-\frac{3}{2}\\-\frac{5}{2}\end{bmatrix}\begin{bmatrix}-\frac{3}{2}\\-\frac{5}{2}\end{bmatrix}^T + \begin{bmatrix}-\frac{1}{2}\\\frac{5}{2}\end{bmatrix}\begin{bmatrix}-\frac{1}{2}\\\frac{5}{2}\end{bmatrix}^T + \begin{bmatrix}\frac{1}{2}\\\frac{3}{2}\end{bmatrix}\begin{bmatrix}\frac{1}{2}\\\frac{3}{2}\end{bmatrix}^T\right) \\ &= \frac{1}{3}\begin{bmatrix}5 & 1\\1 & 17\end{bmatrix} = \begin{bmatrix}\frac{5}{3} & \frac{1}{3}\\\frac{1}{3} & \frac{17}{3}\end{bmatrix}\end{aligned}$$

حال با حل معادله مشخصه، مقادیر ویژه ماتریس $\mathrm{Cov}(X)$ را می یابیم:

$$\begin{aligned}
\det\left(\mathrm{Cov}(X)-\lambda I\right)=۰ &\Longrightarrow \det\left(\begin{bmatrix}\frac{۵}{۳} & \frac{۱}{۳}\\ \frac{۱}{۳} & \frac{۱۷}{۳}\end{bmatrix}-\lambda I\right)=۰\\
&\Longrightarrow \det\left(\begin{bmatrix}\frac{۵}{۳}-\lambda & \frac{۱}{۳}\\ \frac{۱}{۳} & \frac{۱۷}{۳}-\lambda\end{bmatrix}\right)=۰\\
&\Longrightarrow \left(\frac{۵}{۳}-\lambda\right)\left(\frac{۱۷}{۳}-\lambda\right)-\frac{۱}{۹}=۰\\
&\Longrightarrow (۵-۳\lambda)(۱۷-۳\lambda)-۱=۰\\
&\Longrightarrow \lambda_۱=\frac{۱۱+\sqrt{۳۷}}{۳},\quad \lambda_۲=\frac{۱۱-\sqrt{۳۷}}{۳}
\end{aligned}$$

از آنجا که هدف انتقال داده ها به فضای یک بعدی است، در ابتدا مقدار ویژه بزرگ تر (یا همان $\lambda_۱$) را در نظر می گیریم و بردار ویژه متناظر به آن را پیدا می کنیم:

$$\mathrm{Cov}(X)v=\lambda_۱ v,\quad \|v\|_۲=۱$$

از تساوی اول نتیجه می شود:

$$\begin{aligned}
\begin{bmatrix}\frac{۵}{۳} & \frac{۱}{۳}\\ \frac{۱}{۳} & \frac{۱۷}{۳}\end{bmatrix}v=\left(\frac{۱۱+\sqrt{۳۷}}{۳}\right)v &\Longrightarrow \begin{bmatrix}\frac{۵}{۳} & \frac{۱}{۳}\\ \frac{۱}{۳} & \frac{۱۷}{۳}\end{bmatrix}\begin{bmatrix}v_۱\\ v_۲\end{bmatrix}=\frac{۱۱+\sqrt{۳۷}}{۳}\begin{bmatrix}v_۱\\ v_۲\end{bmatrix}\\
&\Longrightarrow \begin{cases}۵v_۱+v_۲=\left(۱۱+\sqrt{۳۷}\right)v_۱\\ v_۱+۱۷v_۲=\left(۱۱+\sqrt{۳۷}\right)v_۲\end{cases}\\
&\Longrightarrow v_۲=\left(۶+\sqrt{۳۷}\right)v_۱
\end{aligned}$$

و از ترکیب تساوی اخیر و تساوی دوم به دست می آید:

$$\begin{aligned}
v_۱^۲+v_۲^۲=۱ &\Longrightarrow v_۱^۲+\left(۶+\sqrt{۳۷}\right)^۲ v_۱^۲=۱\\
&\Longrightarrow v_۱^۲\left(۱+\left(۶+\sqrt{۳۷}\right)^۲\right)=۱\\
&\Longrightarrow \begin{cases}v_۱=\frac{۱}{\sqrt{۱+\left(۶+\sqrt{۳۷}\right)^۲}}\\ v_۲=\left(۶+\sqrt{۳۷}\right)v_۱=\frac{۶+\sqrt{۳۷}}{\sqrt{۱+\left(۶+\sqrt{۳۷}\right)^۲}}\end{cases}\\
&\Longrightarrow v=\begin{bmatrix}\frac{۱}{\sqrt{۱+(۶+\sqrt{۳۷})^۲}}\\ \frac{۶+\sqrt{۳۷}}{\sqrt{۱+(۶+\sqrt{۳۷})^۲}}\end{bmatrix}
\end{aligned}$$

بنابراین نقاط انتقال داده شده به شکل زیر خواهند بود:

$$x'_i = x_i^T v$$

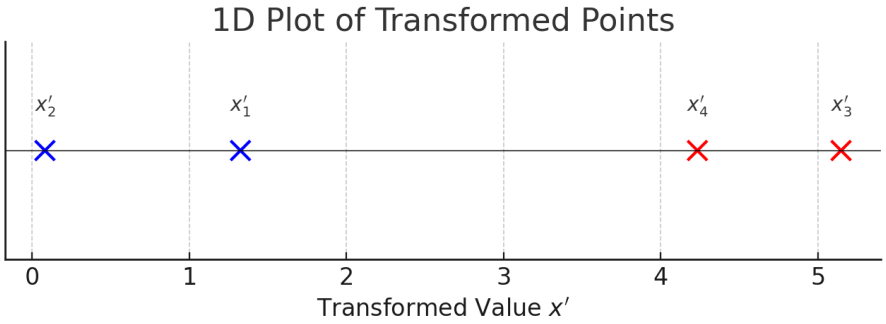
$$\Longrightarrow x'_۱ = \frac{۱۰ + \sqrt{۳۷}}{\sqrt{۱ + (۶ + \sqrt{۳۷})^۲}}$$

$$x'_۲ = \frac{۱}{\sqrt{۱ + (۶ + \sqrt{۳۷})^۲}}$$

$$x'_۳ = \frac{۳۲ + ۵\sqrt{۳۷}}{\sqrt{۱ + (۶ + \sqrt{۳۷})^۲}}$$

$$x'_۴ = \frac{۲۷ + ۴\sqrt{۳۷}}{\sqrt{۱ + (۶ + \sqrt{۳۷})^۲}}$$

حال نقاط به دست آمده را به تفکیک کلاس رسم می کنیم:



همانطور که مشاهده می شود، در این حالت کلاس ها به سادگی و به شکل خطی تمایزپذیرند.

حال مقدار ویژه کوچک تر (یا همان $\lambda_۲$) را در نظر می گیریم و بردار ویژه متناظر به آن را می یابیم:

$$\mathrm{Cov}(X)v = \lambda_۲ v, \quad \|v\|_۲ = ۱$$

کاملا مشابه حالت قبل، می توان نوشت:

$$\begin{bmatrix} \frac{۵}{۳} & \frac{۱}{۳} \\ \frac{۱}{۳} & \frac{۱۷}{۳} \end{bmatrix} v = \left(\frac{۱۱ - \sqrt{۳۷}}{۳}\right) v \Longrightarrow \begin{bmatrix} \frac{۵}{۳} & \frac{۱}{۳} \\ \frac{۱}{۳} & \frac{۱۷}{۳} \end{bmatrix} \begin{bmatrix} v_۱ \\ v_۲ \end{bmatrix} = \frac{۱۱ - \sqrt{۳۷}}{۳} \begin{bmatrix} v_۱ \\ v_۲ \end{bmatrix}$$

$$\Longrightarrow \begin{cases} ۵v_۱ + v_۲ = \left(۱۱ - \sqrt{۳۷}\right) v_۱ \\ v_۱ + ۱۷v_۲ = \left(۱۱ - \sqrt{۳۷}\right) v_۲ \end{cases}$$

$$\Longrightarrow v_۲ = \left(۶ - \sqrt{۳۷}\right) v_۱$$

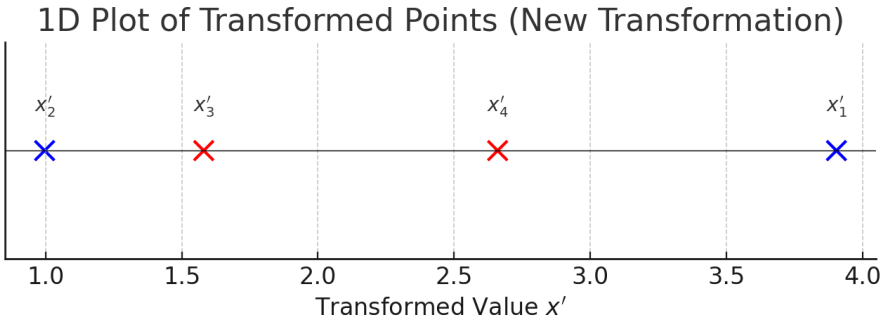
$$v_1^2 + v_2^2 = 1 \Longrightarrow v_1^2 + \left(6 - \sqrt{37}\right)^2 v_1^2 = 1$$
$$\Longrightarrow v_1^2 \left(1 + \left(6 - \sqrt{37}\right)^2\right) = 1$$
$$\Longrightarrow \begin{cases} v_1 = \frac{1}{\sqrt{1+\left(6-\sqrt{37}\right)^2}} \\ v_2 = \left(6 - \sqrt{37}\right) v_1 = \frac{6-\sqrt{37}}{\sqrt{1+\left(6-\sqrt{37}\right)^2}} \end{cases}$$
$$\Longrightarrow v = \begin{bmatrix} \frac{1}{\sqrt{1+(6-\sqrt{37})^2}} \\ \frac{6-\sqrt{37}}{\sqrt{1+(6-\sqrt{37})^2}} \end{bmatrix}$$

بنابراین در این حالت، نقاط انتقال داده شده به شکل زیر خواهند بود:

$$x_i' = x_i^T v$$

$$\Longrightarrow x_1' = \frac{10 - \sqrt{37}}{\sqrt{1 + (6 - \sqrt{37})^2}}$$
$$x_2' = \frac{1}{\sqrt{1 + (6 - \sqrt{37})^2}}$$
$$x_3' = \frac{32 - 5\sqrt{37}}{\sqrt{1 + (6 - \sqrt{37})^2}}$$
$$x_4' = \frac{27 - 4\sqrt{37}}{\sqrt{1 + (6 - \sqrt{37})^2}}$$

حال نقاط به دست آمده را به تفکیک کلاس رسم می کنیم:



در این حالت، دیگر نقاط انتقال یافته به شکل خطی جدایی پذیر نیستند و تمایز کلاس ها نسبت به حالت قبل دشوار تر می باشد. به عبارت دیگر با انتخاب مقدار ویژه کوچک تر، تمایزپذیری کلاس ها کاهش یافت.

ب) نشان می دهیم که اگر بردار دلخواه $x' \in F$ (در اینجا منظور از F، فضای ویژگی ها یا همان Space Feature است) به همه داده ها اضافه شود، مولفه اساسی اول تغییری نخواهد کرد.

در ابتدا دقت کنید که با افزودن x' به همه داده ها، بردار میانگین جدید (که آن را با $\bar{x}'$ نمایش می دهیم) برابر $\bar{x} + x'$ می شود؛ چرا که می توان نوشت:

$$\begin{aligned}
\bar{x}' &= \frac{1}{n}\sum_{i=1}^{n} x_i + x' \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i + nx'\right) \\
&= \frac{1}{n}\left(n\bar{x} + nx'\right) \\
&= \bar{x} + x'
\end{aligned}$$

در نتیجه اگر ماتریس کوواریانس داده های جدید را با $\mathrm{Cov}'(X')$ نمایش دهیم، خواهیم داشت:

$$\begin{aligned}
\mathrm{Cov}'(X') &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i + x' - \bar{x}')(x_i + x' - \bar{x}')^T \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i + x' - \bar{x} - x')(x_i + x' - \bar{x} - x')^T \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \\
&= \mathrm{Cov}(X)
\end{aligned}$$

بنابراین از آنجا که $\mathrm{Cov}'(X')$ برابر $\mathrm{Cov}(X)$ می باشد، بردار های ویژه این دو ماتریس یکسان بوده و بردار ویژه متناظر به بزرگ ترین مقدار ویژه (مولفه اساسی اول برای دو مجموعه داده) نیز در آن ها یکسان خواهند بود. پس با افزودن x' به مجموعه داده اولیه، مولفه اساسی اول تغییر نمی کند و ادعای ما صحیح است.

پاسخ مسئله‌ی ۲.

الف) در ابتدا دقت کنید که طبق تعریف، بایاس و واریانس مدل f که آن ها را به ترتیب با $b(f)$ و $\mathrm{Var}(f)$ نمایش می دهیم، از روابط زیر به دست می آیند:

$$b(f) = \mathbb{E}[f - y], \quad \mathrm{Var}(f) = \mathbb{E}[(f - \bar{f})^۲]$$

به طوری که در عبارات بالا، $y = y(x)$ مقدار واقعی برچسب برای داده x و $\bar{f} = \mathbb{E}_X[f]$ مدل مورد انتظار (یا Expected Model) می باشد. حال دقت کنید که طبق فرض مسئله، داریم:

$$b(f_۱) = b(f_۲) = \cdots = b(f_M) = b, \quad \mathrm{Var}(f_۱) = \mathrm{Var}(f_۲) = \cdots = \mathrm{Var}(f_M) = \sigma^۲$$

با توجه به خاصیت خطی بودن امید ریاضی، می توان نوشت: (فارغ از مستقل بودن یا نبودن f_i ها)

$$\begin{aligned}
b(f_{\text{ensemble}}) &= b\left(\frac{۱}{M}\sum_{i=۱}^{M} f_i\right) \\
&= \mathbb{E}\left[\frac{۱}{M}\sum_{i=۱}^{M} f_i - y\right] \\
&= \mathbb{E}\left[\frac{۱}{M}\sum_{i=۱}^{M} f_i - \frac{M}{M}y\right] \\
&= \mathbb{E}\left[\frac{۱}{M}\sum_{i=۱}^{M} (f_i - y)\right] \\
&= \frac{۱}{M}\sum_{i=۱}^{M} \mathbb{E}[f_i - y] \\
&= \frac{۱}{M}\sum_{i=۱}^{M} b = b
\end{aligned}$$

در نتیجه بایاس f_{ensemble} با بایاس مدل های ضعیف برابر می باشد و با افزایش تعداد این مدل ها نیز مقدار $b(f_{\text{ensemble}}) = b$ تغییری نخواهد کرد.

از طرف دیگر، با توجه به مستقل بودن مدل های ضعیف از یکدیگر، خواهیم داشت:

$$\begin{aligned}
\mathrm{Var}(f_{\text{ensemble}}) &= \mathrm{Var}\left(\frac{۱}{M}\sum_{i=۱}^{M} f_i\right) \\
&= \frac{۱}{M^۲}\sum_{i=۱}^{M} \mathrm{Var}(f_i) \\
&= \frac{۱}{M^۲}(M\sigma^۲) = \frac{\sigma^۲}{M}
\end{aligned}$$

بنابراین با افزایش مقدار M (تعداد مدل های ضعیف)، $\mathrm{Var}(f_{\text{ensemble}}) = \frac{\sigma^۲}{M}$ کاهش خواهد یافت.

ب) در ابتدا دقت کنید که بایاس f_{ensemble} هیچگونه وابستگی به استقلال یا عدم استقلال مدل های ضعیف ندارد، چرا که خاصیت خطی بودن امید ریاضی، فارغ از مستقل بودن یا نبودن متغیر های تصادفی جمع شونده برقرار است. در نتیجه بایاس در این قسمت برابر بایاس در قسمت قبل می باشد:

$$b(f_{\text{ensemble}}) = b(f_i) = b$$

در نتیجه بایاس مدل نهایی کاملا مستقل از تعداد مدل های ضعیف (M) و وابستگی بین آنها (ρ) می باشد و با تغییر این پارامتر ها، $b(f_{\text{ensemble}}) = b$ تغییری نخواهد کرد.

اما برای واریانس f_{ensemble} در این حالت خواهیم داشت:

$$\begin{aligned}
\operatorname{Var}(f_{\text{ensemble}}) &= \operatorname{Var}\left(\frac{1}{M}\sum_{i=1}^{M} f_i\right) \\
&= \frac{1}{M^2}\operatorname{Var}\left(\sum_{i=1}^{M} f_i\right) \\
&= \frac{1}{M^2}\left(\sum_{i=1}^{M}\operatorname{Var}(f_i) + 2\sum_{i<j}\operatorname{Cov}(f_i, f_j)\right) \\
&= \frac{1}{M^2}\left(\sum_{i=1}^{M}\sigma^2 + 2\sum_{i<j}\rho\sigma^2\right) \\
&= \frac{1}{M^2}\left(M\sigma^2\right) + \frac{2}{M^2}\binom{M}{2}\rho\sigma^2 \\
&= \frac{\sigma^2}{M} + \left(1 - \frac{1}{M}\right)\rho\sigma^2 = \frac{\sigma^2 - \rho\sigma^2}{M} + \rho\sigma^2
\end{aligned}$$

بنابراین با افزایش مقدار ρ (وابستگی بین مدل های ضعیف) با فرض ثابت بودن M، $\operatorname{Var}(f_{\text{ensemble}})$ افزایش می یابد و با افزایش مقدار M (تعداد مدل های ضعیف) با فرض ثابت بودن ρ، این کمیت کاهش خواهد یافت (دقت کنید که $\sigma^2 - \rho\sigma^2 \geqslant 0$ است). البته توجه کنید که اگر $\rho = 1$ باشد، $\operatorname{Var}(f_{\text{ensemble}})$ مستقل از M خواهد شد.

پ)

- خیر، یادگیرنده‌های ضعیف در AdaBoost نیاز به مشتق‌پذیر بودن ندارند. AdaBoost با تغییر وزن داده‌های نادرست طبقه‌بندی شده در هر تکرار کار می‌کند و به بهینه‌سازی مبتنی بر گرادیان متکی نیست. بنابراین، مدل‌های غیردیفرانسیلی، مانند درخت‌های تصمیم، می‌توانند به عنوان یادگیرنده‌های ضعیف مورد استفاده قرار گیرند.
- boosting به طور کلی از نظر محاسباتی گران‌تر از bagging است. در boosting، هر مدل جدید بر روی نسخه‌ی اصلاح‌شده‌ای از داده‌ها که در آن مثال‌های نادرست طبقه‌بندی شده وزن بیشتری دارند، آموزش می‌بیند، و این فرآیند به صورت ترتیبی انجام می‌شود. این وابستگی بین تکرارها پیچیدگی محاسباتی را افزایش می‌دهد. در مقابل bagging مدل‌ها را به صورت مستقل آموزش می‌دهد که امکان پردازش موازی را فراهم می‌کند.

پاسخ مسئله‌ی ۳.

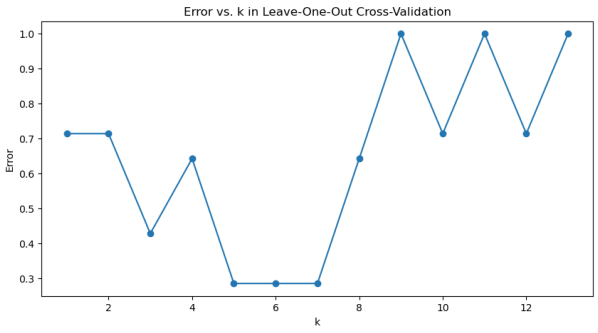
الف) به وضوح اگر فرض کنیم که هر داده همسایه خودش هم می باشد، با انتخاب $k = ۱$ خطا روی مجموعه داده آموزش برابر صفر خواهد شد؛ چرا که نزدیک ترین همسایه هر داده خود آن داده است و در نتیجه مقدار حقیقی برچسب هر داده توسط دسته بند 1-NN برای آن داده پیش بینی خواهد شد که به کمترین میزان خطا (خطای صفر) منجر می شود.

ب) اگر k خیلی کوچک باشد، دسته‌بند به نویزها حساس شده و واریانس آن زیاد می شود و در نتیجه ممکن است دچار بیش‌برازش (overfitting) شود. به طور مثال برای $k = ۱$، دسته‌بند تنها نزدیک‌ترین نقطه را در نظر می‌گیرد که این انتخاب بسیار نسبت به داده های نویزی و outlier ها حساس می باشد.

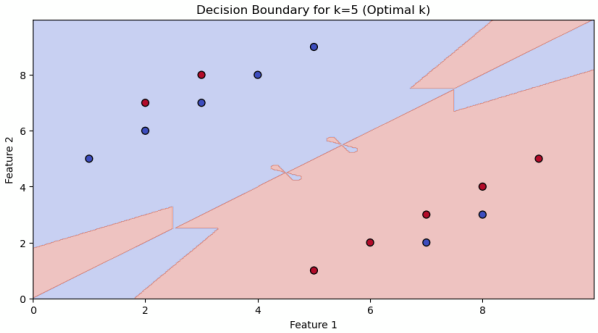
از طرفی اگر k خیلی بزرگ باشد، به طور مثال $k = n$ (تعداد کل نقاط آموزشی)، همه‌ی داده‌ها برای طبقه‌بندی هر نقطه جدید استفاده می‌شوند و دسته‌بند به سمت کم‌برازش شدن (underfitting) می‌رود، زیرا تفاوت بین کلاس‌ها را نادیده می‌گیرد و به هر داده آزمون برچسب بیشینه بین داده های آموزشی را نسبت می دهد. بنابراین انتخاب مقدار بهینه برای k نیازمند تعادلی است که هم الگوهای محلی و هم کلی را در داده‌ها در نظر بگیرد.

پ و ت) در ابتدا فرض می کنیم که برچسب داده هایی که با علامت $(-)$ مشخص شده اند، برابر صفر و برچسب داده هایی که با علامت $(+)$ مشخص شده اند، برابر یک می باشد.

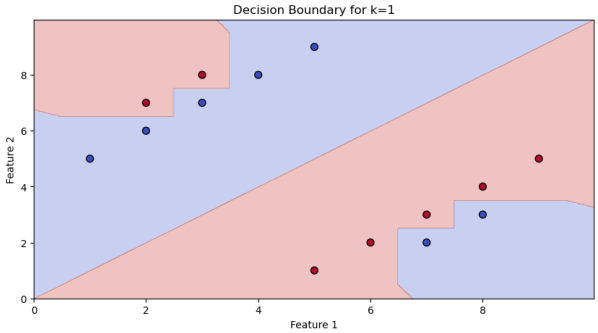
برای حل این قسمت از یک python notebook استفاده شده است که می توانید آن را در این لینک مشاهده کنید. بر اساس این برنامه، بهترین مقدار k با توجه به روش Leave One Out Cross Validation، برابر ۵ است که خطای آن برابر ۰/۲۸۶ می باشد. نمودار های خطا بر حسب مقدار k و مرز تصمیم گیری برای $k = ۱, ۵$ در ادامه آورده شده اند.



شکل ۱: نمودار خطای validation بر حسب مقدار k



شکل ۲: مرز تصمیم گیری برای $k = ۵$



شکل ۳: مرز تصمیم گیری برای $k = ۱$

پاسخ مسئله‌ی ۴.

الف) جهت پیدا کردن مراکز خوشه ها، به گونه ای که تابع هزینه کمینه شود، از این تابع نسبت به هر کدام از آنها گرادیان می گیریم و تمامی این گرادیان ها را برابر صفر قرار می دهیم (دقت کنید که تابع هزینه محدب است و در نتیجه کمینه سراسری آن در نقطه ای رخ می دهد که گرادیان صفر شود). برای این منظور، در ابتدا تابع L را به شکل زیر بازنویسی می کنیم:

$$L = \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - \mu_j\|^2 = \sum_{j=1}^{k} \sum_{x_i \in S_j} (x_i - \mu_j)^T (x_i - \mu_j) = \sum_{j=1}^{k} \sum_{x_i \in S_j} x_i^T x_i - 2\mu_j^T x_i + \mu_j^T \mu_j$$

حال گرادیان تابع هزینه را نسبت به مرکز l امین دسته (برای $1 \leqslant l \leqslant k$) محاسبه می کنیم و آن را برابر صفر قرار می دهیم:

$$\begin{aligned}
\frac{\partial L}{\partial \mu_l} &= \frac{\partial \left(\sum_{j=1}^{k} \sum_{x_i \in S_j} x_i^T x_i - 2\mu_j^T x_i + \mu_j^T \mu_j \right)}{\partial \mu_l} \\
&= \frac{\partial \left(\sum_{x_i \in S_l} x_i^T x_i - 2\mu_l^T x_i + \mu_l^T \mu_l \right)}{\partial \mu_l} \\
&= \sum_{x_i \in S_l} \frac{\partial \left(x_i^T x_i - 2\mu_l^T x_i + \mu_l^T \mu_l \right)}{\partial \mu_l} \\
&= 2 \sum_{x_i \in S_l} \mu_l - x_i \\
&= 2|S_l|\mu_l - 2 \sum_{x_i \in S_l} x_i
\end{aligned}$$

$$\frac{\partial L}{\partial \mu_l} = 0 \implies 2|S_l|\mu_l - 2 \sum_{x_i \in S_l} x_i = 0 \implies \mu_l = \frac{\sum_{x_i \in S_l} x_i}{|S_l|}$$

و در نتیجه مقدار بهینه برای μ_l، میانگین داده های موجود در دسته l ام می باشد و حکم درست است.

ب) الگوریتم k-means نسبت به مقداردهی اولیه مراکز خوشه ها حساس می باشد. به عنوان مثال، فرض کنید داده هایی که می خواهیم آن ها را به $k = 2$ خوشه تقسیم کنیم، به شکل زیر باشند:

$$x_1 = -2, \quad x_2 = -1, \quad x_3 = 1, \quad x_4 = 2$$

اگر مراکز خوشه ها را به صورت $\mu_1 = 0$ و $\mu_2 = 5$ مقداردهی اولیه کنیم، الگوریتم k-means پس از یک مرحله همگرا می شود و خوشه های زیر را به دست می آورد:

$$S_1 = \{x_1, x_2, x_3, x_4\}, \quad S_2 = \emptyset$$

اما اگر مراکز خوشه ها را به شکل $\mu_1 = -1/5$ و $\mu_2 = 1/5$ مقداردهی اولیه کنیم، الگوریتم k-means مجددا پس از یک مرحله همگرا شده و خوشه های زیر را به دست می آورد:

$$S_1 = \{x_1, x_2\}, \quad S_2 = \{x_3, x_4\}$$

بنابراین مقداردهی اولیه متفاوت، می تواند خروجی نهایی k-means را به کل تغییر دهد و اگر این مقداردهی نامناسب باشد، می تواند منجر به همگرایی k-means به یک کمینه محلی شود. برای رفع این مشکل، راهکار هایی مانند الگوریتم k-means++ معرفی شده اند که احتمال همگرایی به یک خوشه بندی بهتر را افزایش می دهند.

الگوریتم k-means حتما به یک پاسخ (نه لزوما پاسخ بهینه) همگرا خواهد شد.

برای اثبات این موضوع، در ابتدا به این نکته دقت کنید که در کل حداکثر k^n حالت مختلف برای خوشه بندی داده ها وجود دارد (n برابر تعداد داده ها است)، و در نتیجه k^n حالت برای مقادیر مراکز خوشه ها و مقدار تابع هزینه وجود خواهد داشت. از طرفی دقت کنید که در هر مرحله از الگوریتم k-means، در صورتی که الگوریتم همگرا و متوقف نشود، مقدار تابع هزینه حتما کاهش می یابد و مقدار جدیدی برای تابع هزینه (نسبت به تمامی مراحل قبل) حاصل می شود. در نتیجه پس از حداکثر k^n مرحله، الگوریتم متوقف خواهد شد (چرا که اگر پس از این تعداد مرحله متوقف نشود، حداقل $k^n + 1$ حالت مختلف برای مقدار تابع هزینه مشاهده می شود که تناقض است).

برای نشان دادن اینکه مقدار تابع هزینه در هر مرحله از الگوریتم اکیدا کاهش می یابد، دقت کنید که هر مرحله شامل دو بخش زیر می باشد:

۱. انتقال هر داده به خوشه ای که کمترین فاصله را از مرکزش دارد.

به وضوح اگر حداقل یک داده در این مرحله به خوشه دیگری منتقل شود (که برای متوقف نشدن الگوریتم باید این اتفاق بیفتد)، مقدار تابع L اکیدا کاهش پیدا می کند.

۲. بروزرسانی مرکز خوشه ها به میانگین داده های موجود در آنها.

در این مرحله نیز، مطابق آنچه که در قسمت الف نشان داده شد، مقدار تابع هزینه کم می شود؛ چرا که اگر مرکز هر خوشه را میانگین آن در نظر بگیریم، کمترین میزان تابع هزینه را برای خوشه بندی ایجاد شده خواهیم داشت.

در نتیجه مقدار تابع L در هر مرحله اکیدا کاهش پیدا می کند و حکم درست می باشد.

پ) در الگوریتم k-means، انتخاب همان خوشه قبلی برای X_j (در صورت فاصله مساوی از چند مرکز خوشه) باعث پایداری بیشتر می‌شود. این تصمیم کمک می‌کند که الگوریتم از تغییرات غیرضروری جلوگیری کند و نقاط به‌صورت تصادفی بین خوشه‌ها جابه‌جا نشوند. این پایداری باعث می‌شود الگوریتم سریع‌تر به همگرایی برسد و تعداد تکرارها کاهش یابد.

اگر این اصل رعایت نشود و X_j به‌طور تصادفی به یکی از خوشه‌های دیگر اختصاص یابد، ممکن است الگوریتم در تکرارهای بعدی با مشکل نوسان مواجه شود؛ یعنی نقاط بین خوشه‌ها جابه‌جا شوند بدون اینکه همگرایی رخ دهد. این جابه‌جایی‌ها باعث افزایش تعداد تکرارها می‌شود و ممکن است الگوریتم در چرخه بی‌پایانی از جابه‌جایی‌ها گرفتار شود، به‌خصوص زمانی که چند نقطه به یک اندازه از چند مرکز خوشه فاصله داشته باشند. در نتیجه، الگوریتم نمی‌تواند به‌طور مؤثری واریانس درون‌خوشه‌ای را کاهش دهد و حتی ممکن است هیچ گاه همگرا نشود.

پاسخ مسئله‌ی ۵.

در ابتدا مقدار تابع $f_u(x)$ را برحسب u محاسبه می کنیم. از آنجا که نرم ۲ یک تابع محدب است، برای پیدا کردن کمینه کننده سراسری تابع $f_u(x)$، کافی است گرادیان آن را نسبت به a محاسبه کرده و آن را برابر صفر قرار دهیم:

$$\begin{aligned} f_u(x) &= \arg\min_{a\in\mathbb{R}} \|x - au\|^۲ \\ &= \arg\min_{a\in\mathbb{R}} (x-au)^T(x-au) \\ &= \arg\min_{a\in\mathbb{R}} x^Tx - ۲au^Tx + a^۲u^Tu \end{aligned}$$

$$\frac{\partial f_u(x)}{\partial a} = -۲u^Tx + ۲au^Tu$$

$$\frac{\partial f_u(x)}{\partial a} = ۰ \implies -۲u^Tx + ۲au^Tu = ۰ \implies a = \frac{u^Tx}{u^Tu}$$

$$\implies f_u(x) = au = \frac{u^Txu}{u^Tu}$$

در ادامه تابع MSE تعریغ شده را، به شکل زیر بازنویسی می کنیم:

$$\begin{aligned} \text{MSE} &= \frac{۱}{m}\sum_{i=۱} \left\|x^{(i)} - f_u\left(x^{(i)}\right)\right\| \\ &= \frac{۱}{m}\sum_{i=۱} \left(x^{(i)} - f_u\left(x^{(i)}\right)\right)^T \left(x^{(i)} - f_u\left(x^{(i)}\right)\right) \\ &= \frac{۱}{m}\sum_{i=۱} x^{(i)T}x^{(i)} - ۲f_u\left(x^{(i)}\right)^T x^{(i)} + f_u\left(x^{(i)}\right)^T f_u\left(x^{(i)}\right) \end{aligned}$$

با توجه به فرض $u^Tu = ۱$، خواهیم داشت $f_u\left(x^{(i)}\right) = u^Tx^{(i)}u$ و در نتیجه می توان نوشت: (دقت کنید که با توجه به خاصیت جابجایی برای ضرب داخلی، $u^Tx^{(i)} = x^{(i)T}u$ برقرار می باشد. همچنین چون $u^Tx^{(i)}$ یک عدد اسکالر است، می توان در ضرب با ماتریس و بردار آن را جابجا نمود)

$$\begin{aligned} \text{MSE} &= \frac{۱}{m}\sum_{i=۱}^{m} x^{(i)T}x^{(i)} - ۲u^Tx^{(i)T}ux^{(i)} + u^Tx^{(i)T}uu^Tx^{(i)}u \\ &= \frac{۱}{m}\sum_{i=۱}^{m} x^{(i)T}x^{(i)} - ۲u^T(u^Tx^{(i)})ux^{(i)} + u^T(u^Tx^{(i)})(u^Tx^{(i)})u \\ &= \frac{۱}{m}\sum_{i=۱}^{m} x^{(i)T}x^{(i)} - ۲\left(u^Tx^{(i)}\right)^۲ + \left(u^Tx^{(i)}\right)^۲ u^Tu \\ &= \frac{۱}{m}\sum_{i=۱}^{m} x^{(i)T}x^{(i)} - ۲\left(u^Tx^{(i)}\right)^۲ + \left(u^Tx^{(i)}\right)^۲ \\ &= \frac{۱}{m}\sum_{i=۱}^{m} x^{(i)T}x^{(i)} - \left(u^Tx^{(i)}\right)^۲ \end{aligned}$$

دقت کنید که اگر تعریف کنیم

$$\Sigma = \frac{۱}{m} \sum_{i=۱}^{m} x^{(i)} x^{(i)^T}$$

می توان نوشت:

$$\begin{aligned}
\text{MSE} &= \frac{۱}{m} \sum_{i=۱}^{m} x^{(i)^T} x^{(i)} - \left(u^T x^{(i)}\right)^۲ \\
&= \frac{۱}{m} \sum_{i=۱}^{m} x^{(i)^T} x^{(i)} - \frac{۱}{m} \sum_{i=۱}^{m} u^T x^{(i)} x^{(i)^T} u \\
&= \frac{۱}{m} \sum_{i=۱}^{m} x^{(i)^T} x^{(i)} - u^T \left(\frac{۱}{m} \sum_{i=۱}^{m} x^{(i)} x^{(i)^T} \right) u \\
&= \frac{۱}{m} \sum_{i=۱}^{m} x^{(i)^T} x^{(i)} - u \Sigma u^T
\end{aligned}$$

حال توجه کنید که با توجه به فرض صورت سوال مبنی بر اینکه داده ها در هر بعد دارای میانگین صفر و واریانس یک هستند، از یک طرف برای ماتریس کوواریانس داده ها داریم:

$$\begin{aligned}
\text{Cov}(X) &= \frac{۱}{m} \sum_{i=۱}^{m} (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T \\
&= \frac{۱}{m} \sum_{i=۱}^{m} x^{(i)} x^{(i)^T} \\
&= \Sigma
\end{aligned}$$

و از طرف دیگر، با توجه به تساوی $\text{trace}(A) = \text{trace}(A^T)$، می توان نوشت:

$$\begin{aligned}
\frac{۱}{m} \sum_{i=۱}^{m} x^{(i)^T} x^{(i)} &= \frac{۱}{m} \text{trace} \left(\sum_{i=۱}^{m} x^{(i)^T} x^{(i)} \right) \\
&= \frac{۱}{m} \text{trace} \left(\sum_{i=۱}^{m} x^{(i)} x^{(i)^T} \right) \\
&= \frac{۱}{m} \text{trace} \left(\text{Cov}(X) \right) \\
&= \frac{۱}{m} \sum_{i=۱}^{m} \sum_{j=۱}^{d} \text{Var}(x_j^{(i)}) \\
&= \frac{۱}{m} \sum_{i=۱}^{m} \sum_{j=۱}^{d} ۱ \\
&= \frac{md}{m} \\
&= d
\end{aligned}$$

به طوری که در تساوی های بالا، d همان بعد فضای ویژگی (Feature Space) می باشد.

بنابراین خواهیم داشت:

$$\text{MSE} = d - u^T \Sigma u$$

$$\implies \arg \min_{u:u^T u=1} \text{MSE} = \arg \min_{u:u^T u=1} d - u^T \Sigma u = \arg \max_{u:u^T u=1} u^T \Sigma u$$

می دانیم که پاسخ مسئله بهنیه سازی $\arg\max_{u:u^T u=1} u^T \Sigma u$ که معادل مسئله بهینه سازی اولیه است، مولفه اساسی اول داده ها می باشد (با توجه به آنچه که در کلاس ها و اسلاید های درسی بیان شده است) و در نتیجه حکم برقرار می باشد.