عادالین روشنایی ۴۰۳۲۶۰۶۴۷۹

۵) الف) میانگین هر ویژگی را به دست می آوریم:

$$\text{mean } w_1 = \frac{\begin{pmatrix}4\\1\end{pmatrix}+\begin{pmatrix}1\\0\end{pmatrix}}{2} = \begin{bmatrix}2.5\\0.5\end{bmatrix}$$

$$\text{mean } w_2 = \frac{\begin{pmatrix}2\\5\end{pmatrix}\begin{pmatrix}3\\4\end{pmatrix}}{2} = \begin{bmatrix}2.5\\4.5\end{bmatrix}$$

محاسبه کوواریانس :

$$\Sigma_{w_1} = \frac{1}{2}\left[\begin{pmatrix}4-2.5\\1-0.5\end{pmatrix}\begin{pmatrix}4-2.5 & 1-0.5\end{pmatrix} + \begin{bmatrix}1-2.5\\0-0.5\end{bmatrix}\begin{pmatrix}1-2.5 & 0-0.5\end{pmatrix}\right]$$

$$\Sigma_{w_1} = \begin{bmatrix}2.25 & 0.75\\0.75 & 0.25\end{bmatrix}$$

$$\Sigma_{w_2} = \frac{1}{2}\left[\begin{bmatrix}-0.5\\0.5\end{bmatrix}\begin{bmatrix}-0.5 & 0.5\end{bmatrix} + \begin{bmatrix}0.5\\-0.5\end{bmatrix}\begin{pmatrix}0.5 & -0.5\end{pmatrix}\right]$$

$$\Sigma_{w_2} = \begin{bmatrix}0.25 & -0.25\\-0.25 & 0.25\end{bmatrix}$$

محاسبه مقادیر ویژه :

$$\det(\Sigma_{w_1} - \lambda I) = 0$$

$$\longrightarrow (2.25-\lambda)(0.25-\lambda) - (0.75)^2 = 0 \longrightarrow \begin{cases}\lambda = 0\\ \lambda = 2.5 \ ✓\end{cases}$$

$$\begin{bmatrix}2.25-2.5 & 0.75\\0.75 & 0.25-2.5\end{bmatrix}\underbrace{\begin{bmatrix}x_1\\y_1\end{bmatrix}}_{v_1} = 0 \rightarrow v_1 = \begin{bmatrix}3\\1\end{bmatrix} \xrightarrow{\text{Norm}} \frac{1}{\sqrt{10}}\begin{bmatrix}3\\1\end{bmatrix}$$

$$\det(\Sigma_{w_2} - \lambda I) = 0$$

$$\rightarrow \begin{cases}\lambda = 0\\ \lambda = 0.5 \ ✓\end{cases} \longrightarrow \begin{bmatrix}-0.25 & -0.25\\-0.25 & -0.25\end{bmatrix}\overbrace{\begin{bmatrix}x_2\\y_2\end{bmatrix}}^{v_2} = 0 \rightarrow v_2 = \begin{bmatrix}1\\-1\end{bmatrix} \xrightarrow{\text{Norm}} \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$$

$$w_1: \begin{cases}\frac{1}{\sqrt{10}}\begin{bmatrix}3 & 1\end{bmatrix}\begin{bmatrix}4\\1\end{bmatrix} = \frac{13}{\sqrt{10}}\\ \frac{1}{\sqrt{10}}\begin{bmatrix}3 & 1\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} = \frac{3}{\sqrt{10}}\end{cases} \qquad w_2: \begin{cases}\frac{1}{\sqrt{2}}\begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}2\\5\end{bmatrix} = \frac{-3}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}3\\4\end{bmatrix} = \frac{-1}{\sqrt{2}}\end{cases}$$

ج) بردار عمود بر بردارهای ویژه :

$$v_1 = \begin{bmatrix} 3/\sqrt{10} \\ 1/\sqrt{10} \end{bmatrix} \qquad v_1' = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$v_1' \perp v_1 \rightarrow \frac{3}{\sqrt{10}} x_1 + \frac{x_2}{\sqrt{10}} = 0 \rightarrow \begin{cases} x_1 = 1 \\ x_2 = -3 \end{cases} \rightarrow v_1' = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \rightarrow v_2' = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

۲) الف)

$$\text{Bias}(f_i(x)) = \mathbb{E}(f_i(x)) - f(x) = \mu - f(x)$$

$$\text{Bias}(f_{ensemble}(x)) = \mathbb{E}(f_i(x)) - f(x) = \frac{\mu \times m}{m} - f(x) = \mu - f(x)$$

بایاس تغییری نمی‌کند

$$\text{Var}(f_i(x)) = \sigma^2 \rightarrow \text{Var}(f_{ensemble}(x)) = \text{Var}\left(\frac{1}{m}\sum_{i=1}^{m} f_i(x)\right)$$

$$= \frac{1}{m^2}\sum_{i=1}^{m} \text{var}(f_i(x)) = \frac{1}{m^2} \times m\sigma^2 = \frac{\sigma^2}{m} \rightarrow$$

واریانس کاهش می‌یابد!

ب) داشتن مدل‌ها بایاس را تغییر نمی‌دهد

$$\text{Var}(f_{ensemble}(x)) = \frac{1}{m^2}\text{var}\left(\sum_{i=1}^{m} f_i(x)\right)$$

$$= \frac{1}{m^2}\left[\sum_{i=1}^{m}\text{var}(f_i(x)) + \sum_{i \neq j}\text{cov}(f_i(x), f_j(x))\right]$$

$$= \frac{1}{m^2}\left[m\sigma^2 + m(m-1)\rho\sigma^2\right] = \frac{\sigma^2}{m} + \frac{(m-1)}{m}\sigma^2\rho$$

اگر $\rho = 1$ باشد با افزایش تعداد واریانس تغییری نمی‌کند $\text{Var}(f_{ensemble}(x)) = \sigma^2$

پس هرچه ρ کوچکتر باشد تاثیر ensemble learning بیشتر است.

۴) ب) در Adaboost نیازی به مشتق پذیر بودن مدل های ضعیف نیست زیرا بر اساس تراداپان وزن مدل ها تعیین نمی شوند بلکه براساس کارایی آن ها تعیین می‌شود. به طوری که به مدل ها با تعداد misclassification بیشتر، داده ها با وزن بزرگتری اختصاص داده می‌شود.

Boosting از نظر محاسباتی گران تر است. زیرا الگوریتم های Boosting به صورت sequential عمل می کنند در حالی که الگوریتم های Bagging موازی پردازش می‌کنند و وابستگی بین مدل ها وجود ندارد.

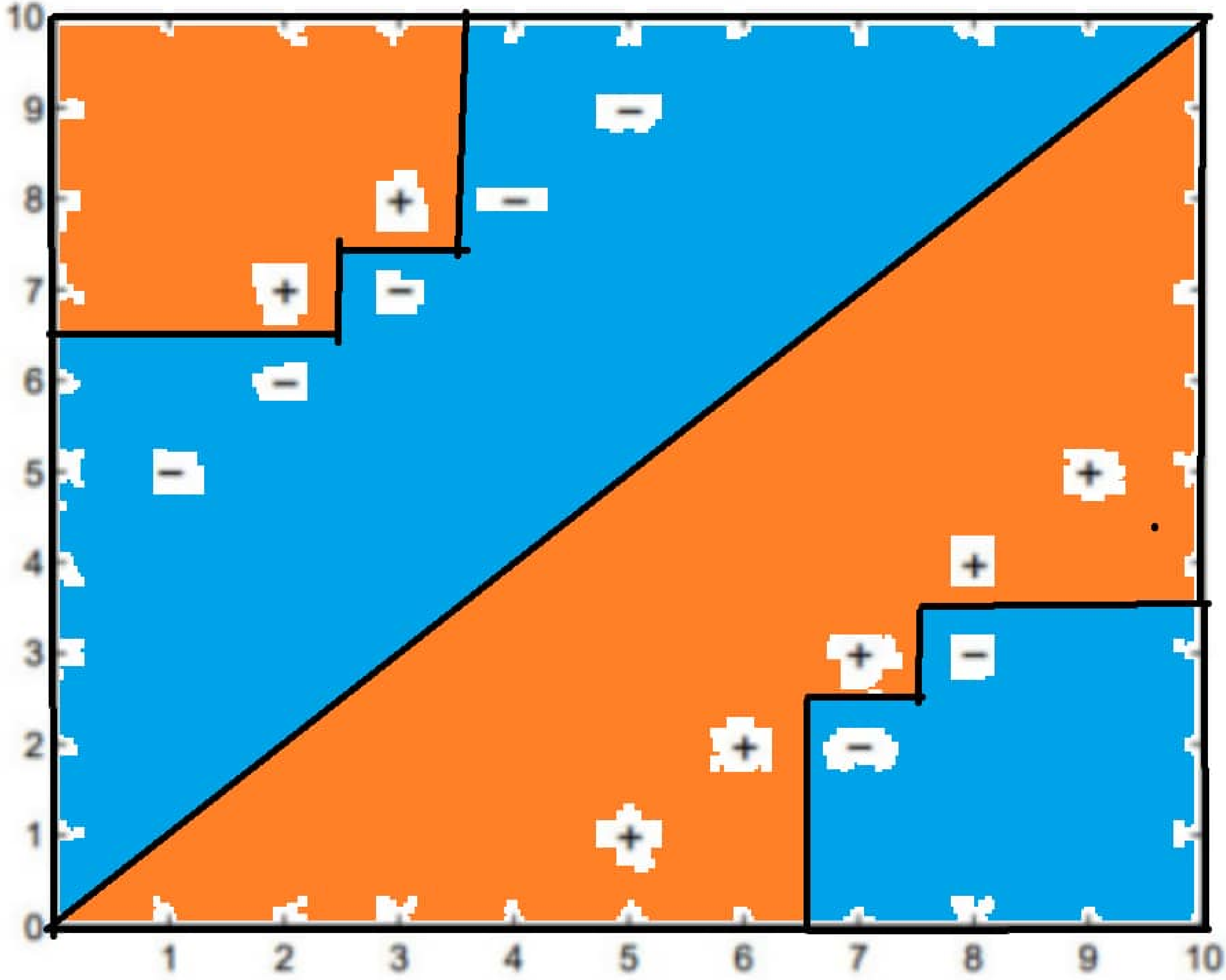
۵) الف) $K=6$

$$error = \frac{\#misclassified}{\#total} = \frac{2}{7} = 28.5\%$$

ب) برای K های کوچک مدل به الگوهای جزئی و غیر عمومی توجه می‌کند و نمی‌تواند به درستی بر روی داده های جدید عملکرد درستی داشته باشد

برای K های خیلی بزرگ، مدل الگوهای محلی را در نظر نمی‌گیرد و نمی‌تواند تفاوت های بین کلاس ها را به درستی تشخیص دهد

پ) $K=5$ و $error = 29\%$ (کد در فایل zip)



0
1
2
3
4
5
6
7
8
9
10
+
−

۴) الف)

$$\frac{\partial L}{\partial \mu_j} = 0 \rightarrow \sum_{j=1}^{k} \sum_{x_i \in S_j} 2\|x_i - \mu_i\| = 0$$

$$\rightarrow \mu_j = \frac{\sum_{x_i \in S_j} x_i}{\| S_j \|} \longrightarrow$$ میانگین هر خوشه

ب) بله با توجه به مقدار دهی های اولیه متفاوت Kmeans پاسخ های متفاوتی می‌تواند تولید کند. و بله Kmeans همواره همگرا می‌شود ولی لزوما جواب درست را پیدا نمی‌کند.

پ) اگر بر فرض x_j به صورت رندوم خوشه‌ای را انتخاب کند ممکن است در بین خوشه‌ها جابه‌جا بشود و این حرکت نوسانی باعث کند شدن همگرایی مدل می‌شود.

۵)

$$f_u(x) = \arg\min_{v \in V} \|x - v\|^2$$

$$V = \{au ;\ a \in \mathbb{R}\} \qquad \Rightarrow f_u(x) = (x \cdot u)u$$

$$\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - f_u(x^{(i)})\|^2 = \frac{1}{m}\sum \|x^{(i)} - (x^{(i)} \cdot u)u\|^2$$

$$\Rightarrow \|x^{(i)} - (x^{(i)} \cdot u)u\|^2 = \|x^{(i)}\|^2 - 2(x^{(i)} \cdot u)^2 + (x^{(i)} \cdot u)^2\|u\|$$

$$\xrightarrow{\|u\|=1} \|x^{(i)}\|^2 - (x^{(i)} \cdot u)^2$$

$$\Rightarrow \frac{1}{m}\sum_{i=1}^{m} \|x^{(i)}\|^2 - (x^{(i)} \cdot u)^2 \xrightarrow{\text{ثابت } \|x^{(i)}\|^2} \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} \cdot u)^2 = Var(\ldots)$$

$$\arg\min_{u : u^T u = 1} \frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - f_u(x^{(i)})\|^2 = \arg\max_{u : u^T u = 1} \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} \cdot u)^2$$

$\Rightarrow$ u باید طوری انتخاب شود که واریانس را بیشینه کند که همان اولین مولفه اساسی است.