۱. یک مدل دسته‌بند دودویی آموزش داده‌اید که دقت بسیار بالایی در داده‌های آموزش دارد، اما در داده‌های اعتبارسنجی[۱] دقت بسیار کمتری دارد. از میان موارد زیر،کدام گزینه‌ها ممکن است درست باشد:

☐ این یک نمونه از بیش‌برازش[۲] است.

☐ این یک نمونه از کم‌برازش[۳] است.

☐ آموزش به خوبی regularized نشده است.

☐ نمونه‌های آموزش و آزمون از توزیع‌های متفاوتی نمونه‌برداری شده‌اند.

پاسخ گزینه های اول سوم و چهارم

۲. بردارهای پشتیبان چیست؟

☐ نمونه‌هایی هستند که دورترین فاصله را از مرز تصمیم مشخص می‌کنند.

☐ تنها نمونه‌های لازم برای محاسبه f(x) در SVM هستند.

☐ مراکز کلاس ها هستند.

☐ تمام نمونه‌هایی هستند که وزن غیرصفر (α_k) در SVM دارند.

پاسخ گزینه های دوم و چهارم

۳. کدام یک از موارد زیر تنها در صورت جداپذیری خطی داده‌های آموزش قابل استفاده است؟

☐ Linear hard-margin SVM

☐ Linear Logistic Regression

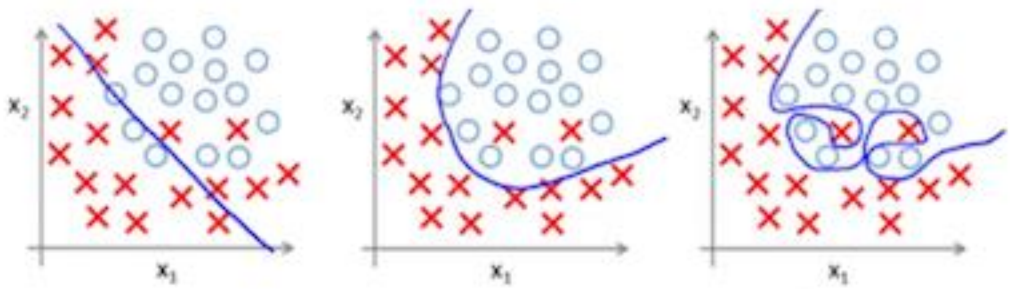
☐ Linear Soft margin SVM

☐ The centroid method

☐ Parzen windows

پاسخ گزینه اول

۴. سه دسته‌بند مختلف بر روی داده‌های یکسان آموزش داده شده‌اند. مرز تصمیم آن‌ها در زیر نشان داده شده است. کدامیک از گزاره‌های زیر درست است؟



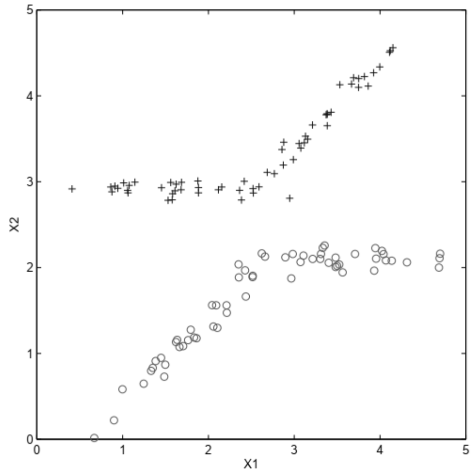
---

[1] validation
[2] overfitting
[3] underfitting

- [ ] دسته‌بند سمت چپ با دقت بالا و انعطاف پذیری کمی آموزش داده شده است.
- [ ] دسته‌بند سمت چپ با دقت پایین و انعطاف پذیری بالا آموزش داده شده است.
- [ ] دسته‌بند سمت راست با دقت پایین و انعطاف پذیری بالا آموزش داده شده است.
- [ ] دسته‌بند سمت راست با دقت بالا و انعطاف پذیری کمی آموزش داده شده است.

پاسخ گزینه های اول و سوم

مجموعه‌ی داده‌ی زیر را برای یک مسأله طبقه‌بندی دو کلاسه در نظر بگیرید. اگر بخواهیم از مدل مخلوط گاوسی (GMM) برای هر یک از دو کلاس استفاده کنیم، چه تعداد مولفه را برای هر کلاس مناسب می‌دانید؟



۱) یک مولفه با ماتریس کوواریانس قطری
۲) یک مولفه با ماتریس کوواریانس دلخواه
۳) دو مولفه با ماتریس کوواریانس قطری
۴) دو مولفه با ماتریس کوواریانس دلخواه

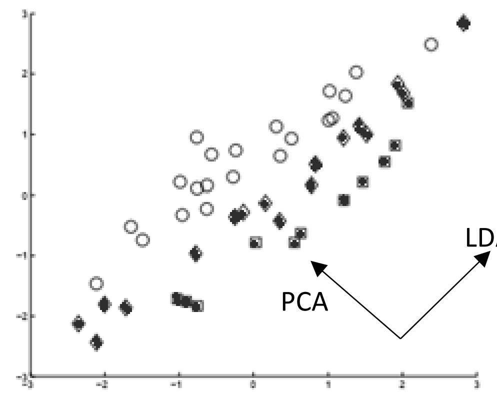
گزینه ۴ درست است.

و) اگر $k_1(x,y)$ و $k_2(x,y)$ دو تابع هسته (*kernel*) معتبر برای روش *kernel-SVM* باشند، آنگاه کدامیک از موارد زیر <u>نمی</u>تواند یک تابع هسته معتبر باشد؟

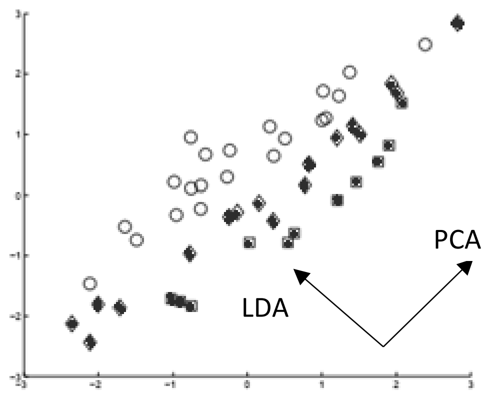
۱) $k_1(x,y)+k_2(x,y)$
۲) $ak_1(x,y)$ به ازای $a>0$
۳) $k_1(x,y)k_2(x,y)$
۴) $k_1(x,x)-k_2(y,y)$

گزینه ۴ درست است.

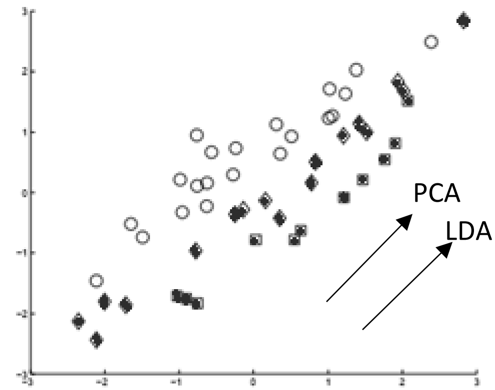
ز) فرض کنید داده‌های دو کلاس در فضای دو بعدی باشند. کدام گزینه جهت درست مولفه اول PCA و LDA را نشان می‌دهد.



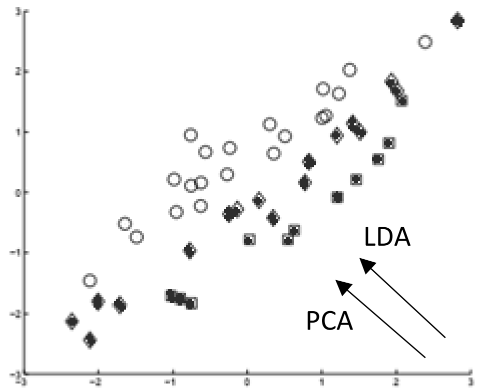
(۱)



(۲)



(۳)



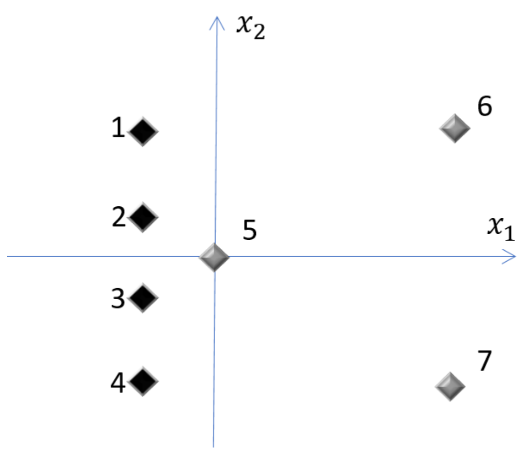
(۴)

گزینه ۲ درست است.

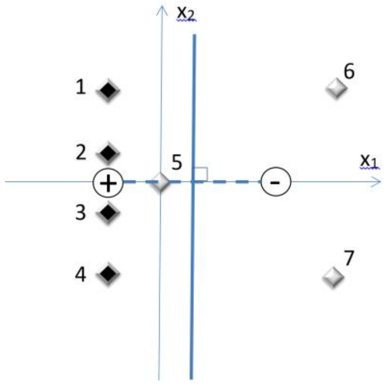
سوال ۲

قسمت اول: دسته بندی به کمک روش مرکز کلاس:(Centroid method)

یک مسئله دسته‌بندی دو کلاسه در یک فضای ویژگی دو بعدی $x=[x_1, x_2]$ با برچسب $y=\pm1$ را در نظر بگیرید. داده‌های آموزشی شامل ۷ نمونه است که در شکل زیر نشان داده شده است (۴ الماس سیاه برای کلاس مثبت و ۳ الماس سفید برای کلاس منفی).



الف) در شکل بالا مراکز[4] دو کلاس را رسم کنید (آن‌ها را با یک دایره با علامت (+) داخل آن، برای کلاس مثبت و یک دایره با علامت (-) برای کلاس منفی مشخص کنید). سپس مراکز را با یک خط چین‌دار ضخیم به‌هم وصل کنید. در نهایت مرز تصمیم روش مرکز کلاس را با یک خط پررنگ صلب رسم کنید. (۲ نمره)



ب) نرخ خطای آموزش چند است (۱نمره) ؟ **پاسخ** ۱/۷

ج) (۲ نمره) آیا نمونه‌ای وجود دارد که با حذف آن، مرز تصمیم به گونه‌ای تغییر می‌کند که نمونه حذف شده به سمت دیگر می‌رود؟ (با پاسخ "بله" یا "خیر" پاسخ دهید) (۲ نمره) پاسخ: **خیر**

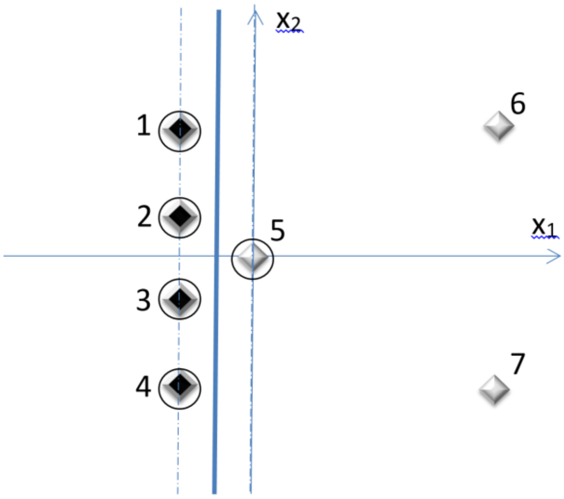
د) نرخ خطای leave-one-out چند است؟ **پاسخ** ۱/۷

سوال ۳ SVM (ماشین بردار پشتیبان)

باز هم داده‌های آموزشی موجود در شکل قبل را درنظر بگیرید

الف) مرز تصمیم حاصل از روش SVM خطی hard margin را با یک خط صلب پهن رسم کنید. حاشیه‌ها را به هر دو طرف با خط‌های چین دار نازک مشخص کنید. بردارهای پشتیبان را با دایره مشخص کنید.

[4] centroids



ب) نرخ خطای آموزش چند است (۱ نمره) ؟ **پاسخ: صفر**

ج) حذف کدام نمونه منجر به تغییر مرز تصمیم می‌شود (۱ نمره)؟ **پاسخ: شماره ۵**

د) نرخ خطای leave-one-out چند است (۲ نمره)؟ **پاسخ** ۱/۷

ه) انعطاف پذیرتری روش مرکز کلاس (قسمت اول) را SVM مقایسه کنید (۲ نمره)؟ **پاسخ: روش مرکز کلاس چرا که تفاوت بین خطای آموزش و خطای leave-one-out آن کمتر است.**

و) قسمت (الف) را در حالتی که از soft-margin SVM استفاده کنیم، پاسخ دهید.

همه نمونه‌ها SV می‌شوند. حاشیه‌ی جدا کننده (خط‌چین‌ها) دو خط عمودی می‌شوند که یکی از داده‌ی ۶ و ۷ عبور می‌کند و دیگری از داده‌های ۱ تا ۴ عبور می‌کند. مرز جدا کننده، یک خط عمودی می‌شود که دقیقا وسط دو خط‌چین قبلی قرار می‌گیرد.

سوال ۴

الف)

$$e_g = \frac{1}{n}\sum_{i=1}^{n}\left(g(x_i) - y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} w_k f_k(x_i) - y_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} w_k f_k(x_i) - \sum_{k=1}^{K} w_k y_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} w_k \left(f_k(x_i) - y_i\right)\right)^2$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_k \left(f_k(x_i) - y_i\right)^2$$

$$= \sum_{k=1}^{K} w_k \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(f_k(x_i) - y_i\right)^2}_{e_k}$$

$$= \sum_{k=1}^{K} w_k e_k$$

ب)

این کاهش می‌تواند ناشی از کاهش هر دو خطای بایاس و واریانس باشد. در واقع مدل g از تک‌تک مدل‌های f_k پیچیده‌تر است و لذا خطای بایاس کمتری دارد. از طرفی اگر مدل‌های f_k روی داده‌های آموزشی مختلفی آموزش دیده باشند، این معادل این است که مدل g روی مقدار داده‌های بیشتری آموزش دیده است و این باعث کاهش خطای واریانس می‌شود.

سوال ۵

الف)

$\theta = (\alpha, \beta, \mu, \lambda_1, \lambda_2)$

$$L(\theta) = \log P(x_1, y_1, z_1, \dots, x_n, y_n, z_n|\theta) \overset{i.i.d.}{=} \sum_{i=1}^{n} \log P(x_i, y_i, z_i|\theta)$$

$$= \sum_{i=1}^{n} (1-z_i)\left[\log(1-\alpha) - \overbrace{\log 2\pi}^{const.} - \frac{(x_i-\mu)^2}{2}\right] +$$

$$z_i \left[\log\alpha + (1-y_i)\left(\log(1-\beta) + \log\lambda_2 - \lambda_2 x_i\right)\right] +$$

$$z_i y_i \left(\log\beta + \log\lambda_1 - \lambda_1 x_i\right)$$

$$= \sum_{i=1}^{n} (1-z_i)\left(\log(1-\alpha) - \frac{(x_i-\mu)^2}{2}\right) + z_i\log\alpha + (z_i - y_i z_i)\left(\log(1-\beta) + \log\lambda_2 - \lambda_2 x_i\right)$$

$$+ y_i z_i \left(\log\beta + \log\lambda_1 - \lambda_1 x_i\right) + const.$$

ب)

$$Q(\theta) = E[L(\theta)] = \sum_{i=1}^{n} (1 - E[z_i])\left(\log(1-\alpha) - \frac{(x_i-\mu)^2}{2}\right) + E[z_i]\log\alpha$$

$$+ \left(E[z_i] - E[y_i z_i]\right)\left(\log(1-\beta) + \log\lambda_2 - \lambda_2 x_i\right)$$

$$+ E[z_i y_i]\left(\log\beta + \log\lambda_1 - \lambda_1 x_i\right) + const.$$

بنابراین باید $E[z_i]$ و $E[y_i z_i]$ را حساب کنیم.

$$E[z_i \mid x_i, \theta^t] = P(z_i = 1 \mid x_i, \theta^t) = 1 - P(z_i = 0 \mid x_i, \theta^t)$$

$$= 1 - \frac{P(x_i \mid z_i = 0, \theta^t)\; P(z_i = 0 \mid \theta^t)}{P(x_i \mid \theta^t)}$$

$$= 1 - \frac{\frac{1}{2\pi} e^{-\frac{1}{2}(x_i - \mu^t)^2} (1 - \alpha^t)}{\sum_{z=0}^{1} \sum_{y=0}^{1} P(x_i, y, z \mid \theta^t)} = \gamma_i^t$$

$$E[y_i z_i \mid x_i, \theta^t] = P(y_i z_i = 1 \mid x_i, \theta^t) = P(y_i = 1, z_i = 1 \mid x_i, \theta^t)$$

$$= \frac{P(x_i \mid y_i = 1, z_i = 1, \theta^t)\; P(y_i = 1, z_i = 1 \mid \theta^t)}{P(x_i \mid \theta^t)}$$

$$= \frac{\lambda_1^t e^{-\lambda_1^t x_i}\; P(y_i = 1 \mid \theta^t)\; P(z_i = 1 \mid \theta^t)}{P(x_i \mid \theta^t)}$$

$$= \frac{\lambda_1^t e^{-\lambda_1^t x_i}\; \beta^t \alpha^t}{\sum_{z=0}^{1} \sum_{y=0}^{1} P(x_i, y, z \mid \theta)} = \eta_i^t$$

برای محاسبه $E[z_i|x_i,\theta^t]$ به صورت زیر می‌توان عمل کرد:

$$E[z_i|x_i,\theta^t] = P(z_i=1|x_i,\theta^t) = \frac{P(x_i|z_i=1,\theta^t)\,P(z_i=1|\theta^t)}{P(x_i|\theta^t)}$$

$$= \frac{\left(\beta^t \lambda_1^t e^{-\lambda_1^t x_i} + (1-\beta^t)\lambda_2^t e^{-\lambda_2^t x_i}\right)\alpha^t}{\sum_{z=0}^{1}\sum_{y=0}^{1} P(x_i, y, z|\theta^t)}$$

ج)

$$\frac{dQ(\theta)}{d\beta} = \frac{d}{d\beta} \sum_{i=1}^{n} (\gamma_i^t - \eta_i^t) \log(1-\beta) + \eta_i^t \log\beta + \text{const}$$

$$= \frac{-\sum_{i=1}^{n}(\gamma_i^t - \eta_i^t)}{1-\beta} + \frac{\sum_{i=1}^{n} \eta_i^t}{\beta} = 0$$

$$\Rightarrow \boxed{\beta = \frac{\sum_{i=1}^{n} \eta_i^t}{\sum_{i=1}^{n} \gamma_i^t}}$$

$$\frac{dQ(\theta)}{d\mu} = \frac{d}{d\mu} \sum_{i=1}^{n} (1-\gamma_i^t) \frac{-(x_i-\mu)^2}{2}$$

$$= +\sum_{i=1}^{n} (1-\gamma_i^t)(x_i-\mu) = 0 \Rightarrow \boxed{\mu = \frac{\sum_{i=1}^{n}(1-\gamma_i^t)x_i}{n}}$$

$$\frac{dQ(\theta)}{d\lambda_1} = \frac{d}{d\lambda_1} \sum_{i=1}^{n} \eta_i^t \left(\log \lambda_1 - \lambda_1 x_i\right) + \text{const.}$$

$$= \sum_{i=1}^{n} \frac{\eta_i^t}{\lambda_1} - \eta_i^t x_i = 0 \Rightarrow \boxed{\lambda_1 = \frac{\sum_{i=1}^{n} \eta_i^t}{\sum_{i=1}^{n} \eta_i^t x_i}}$$

سوال ۶

الف) تابع لاگرانژین را تشکیل می‌دهیم:

$$L(w) = w^T S_B w + \alpha \|w\|_2^2 - \lambda\left(w^T S_w w - k\right)$$

$$\nabla_w L(w) = 2 S_B w + 2\alpha w - 2\lambda S_w w = 0$$

$$\cancel{S_B w = \lambda(S_w + \alpha I) w}$$

$$\Rightarrow (S_B + \alpha I) w = \lambda S_w w$$

$$\Rightarrow \begin{cases} \boxed{S_w^{-1} (S_B + \alpha I) w = \lambda w} \\ \text{OR} \\ \boxed{(S_B + \alpha I)^{-1} S_w w = \frac{1}{\lambda} w} \end{cases}$$

ب) همان طور که در قسمت قبل دیده شد، باید معکوس ماتریس S_w یا S_B در LDA معمولی حساب شود که ممکن است معکوس پذیر نباشد ولی در اینجا به دلیل وجود αI ماتریس $(S_B + \alpha I)$ معکوس پذیر می‌شود.