سوال 1

الف) غلط، به طور مثال فرض کنید که سه طبقه بند h_1، h_2 و h_3 را ترکیب کنیم. فرض کنید h_1 و h_2 نمونه ها را دقیقاً مشابه هم دسته بندی کنند. بنابراین، طبقه بند حاصل از رای اکثریت خطای مساوی با h_1 و h_2 دارد، حال آن که ممکن است دقت h_3 بیشتر باشد.

ب) درست، همواره خطای طبقه بند 1-NN روی داده های آموزشی صفر است. لذا خطای بایاس ندارد.

ج) درست، زیرا فقط SV ها روی مرز جداکننده تاثیر دارند. ضرایب لاگرانژ نمونه های غیر SV صفر است.

د) نادرست، علاوه بر محدب بودن باید ناحیه Feasible نقطه ی درونی داشته باشد.

ه) نادرست، زیرا ممکن است یک مجموعه ی K تایی دیگری وجود داشته باشد که این دسته بند نتواند آن را shatter کند.

و) درست، حاصل جمع دو کرنل معتبر همواره یک کرنل معتبر بدست می دهد (البته ضرایب حاصلجمع باید مثبت باشد مانند اینجا)

ز) نادرست، در SVM اگر نمونه های غیر SV را حذف کنیم، تاثیری روی مرز جداساز ندارد. درحالی که در logistic Reg. حذف هر نمونه ای می تواند مرز را تغییر دهد.

سوال ۲   الف)

$$\min_{r,c,e} \quad r^2 + \alpha \sum_{i=1}^{n} e_i$$

$$\text{s.t.} \quad \|x_i - c\|^2 \leq r^2 + e_i \qquad i=1,\dots,n$$

$$e_i \geq 0 \qquad i=1,\dots,n$$

ب)

$$L(r,c,e,\lambda,\mu) = r^2 + \alpha \sum_{i=1}^{n} e_i - \sum_{i=1}^{n} \lambda_i \left(r^2 + e_i - \|x_i - c\|^2\right) - \sum_{i=1}^{n} \mu_i e_i$$

$$\mu_i \geq 0$$

$$\lambda_i \geq 0$$

ج)

$$g(\lambda,\mu) = \min_{r,c,e} L(r,c,e,\lambda,\mu)$$

$$\frac{\partial L}{\partial r} = 2r - \sum_{i=1}^{n} 2r\lambda_i = 0 \Rightarrow \boxed{\sum_{i=1}^{n} \lambda_i = 1} \; ①$$

②

$$\frac{\partial L}{\partial c} = \sum_{i=1}^{n} \lambda_i (2c - 2x_i) = 0 \Rightarrow c \underbrace{\sum_{i=1}^{n} \lambda_i}_{1} = \sum_{i=1}^{n} x_i \Rightarrow \boxed{c = \sum_{i=1}^{n} x_i}$$

$$\frac{\partial L}{\partial e_i} = \alpha - \mu_i - \lambda_i = 0 \Rightarrow \boxed{\mu_i = \alpha - \lambda_i} \; ③$$

با جایگذاری ۳،۲،۱ در لاگرانژین

$$①,②,③ \Rightarrow g(\lambda,\mu) = \cancel{r^2} + \cancel{\alpha \sum_{i=1}^{n} e_i} - \cancel{\sum_{i=1}^{n} \lambda_i r^2} - \cancel{\sum_{i=1}^{n} \lambda_i e_i} + \sum_{i=1}^{n} \lambda_i \|x_i - \sum_{j=1}^{n} x_j\|^2$$

$$- \sum_{i=1}^{n} \cancel{(\alpha - \lambda_i)e_i}$$

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i \left\| x_i - \sum_{j=1}^{n} x_j \right\|^2$$

یا

مسأله دوگان :

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i x_i^T x_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j x_i^T x_j$$

$$\text{s.t.} \quad 0 \leq \lambda_i \leq \alpha \qquad i=1,\dots,n$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

د) نمونه هایی که روی سطح ابرکره یا خارج ابرکره قرار می گیرند، ضرایب لاگرانژ آن ها بزرگتر از صفر است.

سوال ۳

الف)

$$\log P(x_1, z_1, \dots, x_n, z_n \mid \lambda_1, \lambda_2, \alpha) \overset{i.i.d}{=} \sum_{i=1}^{n} \log P(x_i, z_i \mid \lambda_1, \lambda_2, \alpha)$$

$$P(x_i, z_i \mid \lambda_1, \lambda_2, \alpha) = \left(\alpha \lambda_1 e^{-\lambda_1 x_i}\right)^{1-z_i} \left((1-\alpha)\lambda_2 e^{-\lambda_2 x_i}\right)^{z_i}$$

$$\Rightarrow \log P(x_i, z_i \mid \lambda_1, \lambda_2, \alpha) = (1-z_i)\left(\log \alpha \lambda_1 e^{-\lambda_1 x_i}\right) + z_i \log\left(\lambda_2 (1-\alpha) e^{-\lambda_2 x_i}\right)$$

$$\Rightarrow$$

$$\text{log complete likelihood} = \sum_{i=1}^{n} (1-z_i)(\log \alpha + \log \lambda_1 - \lambda_1 x_i) + z_i(\log(1-\alpha) + \log \lambda_2 - \lambda_2 x_i)$$

ب)

$$Q(\theta, \theta^t) = E\left[\log P(x_1, z_1, \dots, x_n, z_n \mid \lambda_1, \lambda_2, \alpha)\right]$$

$$= \sum_{i=1}^{n} (1 - E[z_i])(\log \alpha + \log \lambda_1 - \lambda_1 x_i) + E[z_i](\log(1-\alpha) + \log \lambda_2 - \lambda_2 x_i)$$

$$E[z_i] = P(z_i = 1 \mid x_i, \alpha^t, \lambda_1^t, \lambda_2^t) = \frac{P(x_i \mid z_i = 1, \alpha^t, \lambda_1^t, \lambda_2^t)\, P(z_i = 1 \mid \alpha^t, \lambda_1^t, \lambda_2^t)}{P(x_i \mid \alpha^t, \lambda_1^t, \lambda_2^t)}$$

$$= \frac{\lambda_2^t e^{-\lambda_2^t x_i} (1-\alpha^t)}{\alpha^t \lambda_1^t e^{-\lambda_1^t x_i} + (1-\alpha^t)\lambda_2^t e^{-\lambda_2^t x_i}} = \gamma_i^t$$

ج)

$$\frac{\partial Q}{\partial \alpha} = \sum_{i=1}^{n} (1-\gamma_i^t)\frac{1}{\alpha} + \gamma_i^t \frac{-1}{1-\alpha} = 0$$

$$\Rightarrow (1-\alpha)\sum_{i=1}^{n}(1-\gamma_i^t) = \alpha \sum_{i=1}^{n} \gamma_i^t$$

$$\Rightarrow \boxed{\alpha = 1 - \frac{\sum_{i=1}^{n} \gamma_i^t}{n}}$$

$$\frac{\partial Q}{\partial \lambda_1} = \sum_{i=1}^{n} (1-\gamma_i^t)\left(\frac{1}{\lambda_1} - x_i\right) \cancel{+\gamma_i^t} = 0$$

$$\Rightarrow \boxed{\lambda_1 = \frac{n - \sum_{i=1}^{n} \gamma_i^t}{\sum_{i=1}^{n} (1-\gamma_i^t)\, x_i}}$$

سوال ۴

الف)

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left(\alpha h_1(x_i) + \beta h_2(x_i) - y_i\right)^2$$

$$\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{n} 2h_1(x_i)\left(\alpha h_1(x_i) + \beta h_2(x_i) - y_i\right) = 0$$

$$\Rightarrow \alpha \sum_{i=1}^{n} h_1^2(x_i) + \beta \sum_{i=1}^{n} h_1(x_i) h_2(x_i) = \sum_{i=1}^{n} h_1(x_i) y_i$$

به طریق مشابه $\frac{\partial L}{\partial \beta} = 0 \Rightarrow \alpha \sum_{i=1}^{n} h_1(x_i) h_2(x_i) + \beta \sum_{i=1}^{n} h_2^2(x_i) = \sum_{i=1}^{n} h_2(x_i) y_i$

ب)

$$e = \sum_{i=1}^{n} \left(\frac{h_1 + h_2}{2} - y_i \right)^2 = \frac{1}{4} \sum_{i=1}^{n} \left(h_1 - y_i + h_2 - y_i \right)^2$$

$$= \frac{1}{4} \sum_{i=1}^{n} (h_1 - y_i)^2 + (h_2 - y_i)^2 + 2(h_1 - y_i)(h_2 - y_i)$$

$$= \frac{1}{4}(e_1 + e_2) + \frac{1}{2} \sum_{i=1}^{n} (h_1 - y_i)(h_2 - y_i)$$

$$\leq \frac{1}{4}(e_1 + e_2) + \frac{1}{2} \sum_{i=1}^{n} |h_1 - y_i| \, |h_2 - y_i|$$

$$= \frac{1}{4}(e_1 + e_2) + \frac{1}{2} \sqrt{\left(\sum_{i=1}^{n} |h_1 - y_i| \, |h_2 - y_i| \right)^2}$$

Cauchy-Schwarz

$$\leq \frac{1}{4}(e_1 + e_2) + \frac{1}{2} \sqrt{\left(\sum_{i=1}^{n} (h_1 - y_i)^2 \right) \left(\sum_{i=1}^{n} (h_2 - y_i)^2 \right)}$$

$$= \frac{1}{4}(e_1 + e_2) + \frac{1}{2} \sqrt{e_1 e_2}$$

سوال ۵

$D=\{x_1,\dots,x_n\}$

$$P(\theta|D) \propto P(D|\theta)\,P(\theta) = P(D|\theta)\sum_{d=1}^{D}\lambda_d P(\theta|\alpha_d)$$

$$= \sum_{d=1}^{D}\underbrace{\lambda_d P(D|\theta)\,P(\theta|\alpha_d)}_{\text{طبق فرض مسأله به صورت } P(\theta|\alpha'_d) \text{ می‌شود}} = \sum_{d=1}^{D}\lambda_d P(\theta|\alpha'_d)$$

یعنی به صورت یک توزیع مخلوط شد که مشابه توزیع پیشین است.

سوال ۶

ماتریس داده‌ها را به صورت زیر تشکیل می‌دهیم:

$$X=\begin{bmatrix}10 & 8\\ 8 & 2\\ 2 & 0\\ 4 & 6\end{bmatrix} \qquad \mu=\begin{bmatrix}6\\ 4\end{bmatrix}$$

با کم کردن میانگین از نمونه‌ها داریم:

$$X=\begin{bmatrix}4 & 4\\ 2 & -2\\ -4 & -4\\ -2 & 2\end{bmatrix} \qquad S=\frac{1}{4}X^TX=\frac{1}{4}\begin{bmatrix}40 & 24\\ 24 & 40\end{bmatrix}=\begin{bmatrix}10 & 6\\ 6 & 10\end{bmatrix}$$

$$\det(S-\lambda I)=0 \Rightarrow \lambda^2-20\lambda+64=0 \Rightarrow \lambda_1=16,\ \lambda_2=4$$

$$u=\begin{bmatrix}u_1\\ u_2\end{bmatrix}$$

$$Su=\lambda_1 u \Rightarrow 10u_1+6u_2=16u_1 \Rightarrow \boxed{u_1=u_2}$$

$$\|u\|=1 \Rightarrow u_1^2+u_2^2=1 \Rightarrow 2u_1^2=1 \Rightarrow u_1=\pm\sqrt{2}/2$$

پس جهت اصلی به صورت $u=\begin{bmatrix}\frac{\sqrt2}{2}\\ \frac{\sqrt2}{2}\end{bmatrix}$ یا معکوس آن یعنی $\begin{bmatrix}-\frac{\sqrt2}{2}\\ -\frac{\sqrt2}{2}\end{bmatrix}$ خواهد شد

هرکدام از دو بردار بالا اگر بدست آمده باشد نمره کامل تعلق می‌گیرد. اگر روند حل درست باشد ولی محاسبات اشتباه ۷۵٪ نمره داده می‌شود.

سوال ۷

فرض کنید $k=2$ است و داده ها در فضای دو بعدی باشند. حال فرض کنید m نمونه در نقطه‌ی $(0,-1)$ و m نمونه در نقطه‌ی $(0,1)$ قرار دارند و یک نقطه مثلاً در مکان $(0,100)$ قرار دارد. انتخاب بهینه‌ی مرکز دو خوشه این است که نقطه‌ی $(0,1)$ و $(0,-1)$ به عنوان دو مرکز انتخاب شوند. اما الگوریتم پیشنهادی (اگر $m \to \infty$) حتماً نقطه‌ی $(0,100)$ را به عنوان یکی از مراکز انتخاب می‌کند. که هزینه‌ی آن با بزرگ شدن m به سمت بی‌نهایت می‌رود.