UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281B / STAT 241B
Advanced Topics in Statistical Learning Theory

Solutions 1
Spring 2009

Solution 1.1

(a) As shown in the lecture, the Bayes rule

$$g^*(x) = \begin{cases} +1 & \text{if } \eta(x) \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

achieves the Bayes risk, i.e., $R^* = R(g^*) = \mathbb{P}[g^*(X) \neq Y]$. Given $X = x$, we note that

$$\begin{aligned} \mathbb{P}[g^*(X) \neq Y \mid X = x] &= \mathbb{P}[g^*(X) = +1, Y = -1 \mid X = x] + \mathbb{P}[g^*(X) = -1, Y = +1 \mid X = x] \\ &= (1 - \eta(x))\mathbb{I}[g^*(x) = +1] + \eta(x)\mathbb{I}[g^*(x) = -1] \\ &= (1 - \eta(x))\mathbb{I}[\eta(x) \geq 1/2] + \eta(x)\mathbb{I}[\eta(x) < 1/2] \\ &= (1 - \eta(x))\mathbb{I}[\eta(x) \geq 1 - \eta(x)] + \eta(x)\mathbb{I}[\eta(x) < 1 - \eta(x)] \\ &= \min\{\eta(x), 1 - \eta(x)\}. \end{aligned}$$

Taking expectation w.r.t. X on both sides gives the result:

$$R^* = \mathbb{P}\left[g^*(X) \neq Y\right] = \mathbb{E}\left[\min\{\eta(X), 1 - \eta(X)\}\right].$$

(b) An elementary result shows that $\min\{\eta(x), 1-\eta(x)\} = \frac{1}{2} - \left|\eta(x) - \frac{1}{2}\right|$ for all $\eta(x) \in [0, 1]$. By part (a), we have

$$R^* = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}] = \mathbb{E}\left[\frac{1}{2} - \left|\eta(X) - \frac{1}{2}\right|\right] = \frac{1}{2} - \frac{1}{2}\mathbb{E}\left|2\eta(X) - 1\right|.$$

(c) Using Bayes' law,

$$\eta(x) = \mathbb{P}[Y = 1 \mid X = x] = \frac{\mathbb{P}[Y = 1]f_+(x)}{\mathbb{P}[Y = 1]f_+(x) + \mathbb{P}[Y = -1]f_-(x)} = \frac{f_+(x)}{f_+(x) + f_-(x)}.$$

Therefore,

$$2\eta(x) - 1 = \frac{2f_+(x)}{f_+(x) + f_-(x)} - 1 = \frac{f_+(x) - f_-(x)}{f_+(x) + f_-(x)}.$$

Substituting it to part(b), we get the desired result:

$$R^* = \frac{1}{2} - \frac{1}{2}\int \left|\frac{f_+(x) - f_-(x)}{f_+(x) + f_-(x)}\right| \left(\frac{1}{2}f_+(x) + \frac{1}{2}f_-(x)\right) dx = \frac{1}{2} - \frac{1}{4}\int \left|f_+(x) - f_-(x)\right| dx.$$

Solution 1.2

(a) Given $(X_1 = x_1, X_2 = x_2)$,

$$\begin{aligned}
\eta(x_1, x_2) &= \mathbb{P}[Y = 1 \mid X_1 = x_1, X_2 = x_2] \\
&= \mathbb{P}[X_1 + X_2 + X_3 \leq 7 \mid X_1 = x_1, X_2 = x_2] \\
&= \mathbb{P}[X_3 \leq 7 - x_1 - x_2 \mid X_1 = x_1, X_2 = x_2] \\
&= \begin{cases} 1 & \text{if } 0 \leq x_1 + x_2 \leq 3 \\ \dfrac{7 - x_1 - x_2}{4} & \text{if } 3 < x_1 + x_2 \leq 7 \\ 0 & \text{if } 7 < x_1 + x_2 \leq 8 \end{cases}
\end{aligned}$$

The last equality follows from the fact that X_3 is uniformly distributed over $[0, 4]$. By noticing that $\eta(x_1, x_2) \geq 1/2$ if and only if $0 \leq x_1 + x_2 \leq 5$, we have the Bayes classifier as follows:

$$g^*(x_1, x_2) = \begin{cases} +1 & \text{if } 0 \leq x_1 + x_2 \leq 5 \\ -1 & \text{otherwise.} \end{cases}$$

The Bayes risk is

$$\begin{aligned}
R^* &= \mathbb{P}[g^*(X_1, X_2) \neq Y] \\
&= \mathbb{P}[0 \leq X_1 + X_2 \leq 5, X_1 + X_2 + X_3 > 7] + \mathbb{P}[5 < X_1 + X_2 \leq 8, 0 \leq X_1 + X_2 + X_3 \leq 7] \\
&= \mathbb{P}[3 \leq X_1 + X_2 \leq 5, X_3 > 7 - (X_1 + X_2)] + \mathbb{P}[5 < X_1 + X_2 \leq 7, 0 \leq X_3 \leq 7 - (X_1 + X_2)]
\end{aligned}$$

To compute above probabilities, we define another random variable $Y = X_1 + X_2$ with its density given by

$$f(y) = \begin{cases} \dfrac{y}{16} & \text{if } 0 \leq y \leq 4 \\ \dfrac{8 - y}{16} & \text{if } 4 < y \leq 8. \end{cases}$$

Continue with the computation of Bayes risk,

$$\begin{aligned}
R^* &= \int_3^5 \left[1 - \frac{7 - y}{4}\right] f(y)\, dy + \int_5^7 \left[\frac{7 - y}{4}\right] f(y)\, dy \\
&= \int_3^4 \frac{y - 3}{4} \frac{y}{16}\, dy + \int_4^5 \frac{y - 3}{4} \frac{8 - y}{16}\, dy + \int_5^7 \frac{7 - y}{4} \frac{8 - y}{16}\, dy \\
&= \frac{11}{384} + \frac{31}{384} + \frac{7}{96} \\
&= \frac{35}{192}.
\end{aligned}$$

(b) The intuition is to design a distribution of X_3 such that $\eta(x_1, x_2) = 1/2$ for almost all of x_1 and x_2, and $\eta(x_1, x_2) = 0$ or 1 for other x_1 and x_2. The corresponding Bayes classifier has to randomly toss a fair coin to guess the label for most instances of x_1 and x_2, and makes perfect prediction for other cases. In this way, its risk can be made arbitrarily close to 1/2.

Formally, given any $\epsilon > 0$, we choose $l < u \in \mathbb{R}$ such that for the random variable $Y = 7 - (X_1 + X_2)$, $\mathbb{P}[Y \in [l, u)] \geq 1 - 2\epsilon$. Denote this semi-open interval as $M_\epsilon = [l, u)$. Now we define the distribution of X_3 as $\mathbb{P}[X_3 = l] = \mathbb{P}[X_3 = u] = 1/2$.

$$\begin{aligned}
\eta(x_1, x_2) &= \mathbb{P}[Y = 1 \mid X_1 = x_1, X_2 = x_2] \\
&= \mathbb{P}[X_3 \leq 7 - (x_1 + x_2) \mid X_1 = x_1, X_2 = x_2] \\
&= \begin{cases} 0 & \text{if } 7 - (x_1 + x_2) < l \\ \frac{1}{2} & \text{if } 7 - (x_1 + x_2) \in M_\epsilon \\ 1 & \text{if } 7 - (x_1 + x_2) \geq u. \end{cases}
\end{aligned}$$

Therefore, $\min\{\eta(x_1, x_2), 1 - \eta(x_1, x_2)\} = \frac{1}{2}\mathbb{I}[7 - (x_1 + x_2) \in M_\epsilon]$. By 1.1(a),

$$\begin{aligned}
R^* &= \mathbb{E}[\min\{\eta(X_1, X_2), 1 - \eta(X_1, X_2)\}] \\
&= \frac{1}{2}\mathbb{P}[7 - (X_1 + X_2) \in M_\epsilon] \\
&= \frac{1}{2}\mathbb{P}[Y \in M_\epsilon] \\
&\geq \frac{1}{2}(1 - 2\epsilon) \\
&= \frac{1}{2} - \epsilon.
\end{aligned}$$

On the other hand, $R^* \leq 1/2$. So we can adjust ϵ to make R^* arbitrarily close to $1/2$.

(c) There are many choices of joint distribution that can lead to the same Bayes classifier g^*. The simplest one is to let X_3 deterministically depends on X_1 and X_2:

$$X_3 = h(X_1, X_2) = (-X_1 - X_2)\mathbb{I}[X_1^2 + X_2^2 < 10] + (10 - X_1 - X_2)\mathbb{I}[X_1^2 + X_2^2 \geq 10].$$

The corresponding $\eta(x_1, x_2)$ is

$$\begin{aligned}
\eta(x_1, x_2) &= \mathbb{P}[X_1 + X_2 + X_3 \leq 7 \mid X_1 = x_1, X_2 = x_2] \\
&= \mathbb{P}[x_1 + x_2 + h(x_1, x_2) \leq 7 \mid X_1 = x_1, X_2 = x_2] \\
&= \mathbb{I}[x_1 + x_2 + h(x_1, x_2) \leq 7] \\
&= \mathbb{I}[x_1^2 + x_2^2 < 10].
\end{aligned}$$

Notice that $\eta(x_1, x_2) \geq 1/2$ if and only if $x_1^2 + x_2^2 < 10$, so we obtain the desired Bayes classifier. The joint density of (X_1, X_2, X_3) is

$$f(x_1, x_2, x_3) = e^{-x_1 - x_2}\mathbb{I}[h(x_1, x_2) = x_3].$$

Solution 1.3

(a) There are two versions of implementation for the perceptron algorithm: find a thereshold function without or with the intercept term. The latter one corresponds to augmenting each original data vector $x^{(i)}$ by an additional component 1 and then run the perceptron

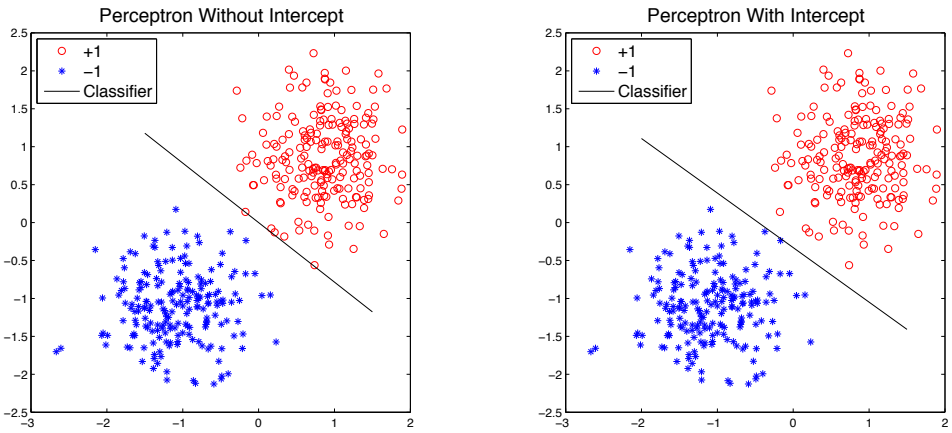

Figure 1: Results of running perceptron algorithms without and with intercept term. When the algorithm still makes mistakes on the data set, the data vector with the smallest index is chosen for update of weight vector.

algorithm without the intercept. See scripts perceptron.m and perceptron_script.m for details.

The results are visualized in figure 1. Perceptron algorithm without intercept converges in 8 iterations for this data set. However, the theoretical bound is 834155 iterations, which is far from being tight. For perceptron with intercept, the algorithm terminates in 15 iterations and the thereshold bound is 25056.

(b) Similar to part (a), there are also two versions of SVM. See scripts svm.m and svm_script.m for details. The left panel of figure 2 shows that SVM without intercept is not a real maximum-margin classifier in this data set: its boundary lies closer to one class of data than another class. This's just because the boundary has been restricted to go through the origin. On the other hand, the boundary produced by SVM with intercept lies exactly in the middle of two classes (see right panel of figure 2).

Solution 1.4

(a) Assume $X \sim f_1 = N(\mu_1, \Sigma_1)$, together with linearity of expectation, we have

$$\begin{aligned}
\mathbb{E}_{f_1}\left[(X-\mu_1)(X-\mu_1)^T\right] = \Sigma_1 \quad &\Rightarrow \quad \Sigma_1^{-1}\mathbb{E}_{f_1}\left[(X-\mu_1)(X-\mu_1)^T\right] = \Sigma_1^{-1}\Sigma_1 = I_d \\
&\Rightarrow \quad \mathbb{E}_{f_1}\left[\Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^T\right] = I_d \\
&\Rightarrow \quad \mathrm{tr}\left(\mathbb{E}_{f_1}\left[\Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^T\right]\right) = \mathrm{tr}\left(I_d\right) = d \\
&\Rightarrow \quad \mathbb{E}_{f_1}\left[\mathrm{tr}\left(\Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^T\right)\right] = d \\
&\Rightarrow \quad \mathbb{E}_{f_1}\left[(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)\right] = d.
\end{aligned}$$

Similarly, $\mathbb{E}_{f_1}\left[(X-\mu_1)^T\Sigma_2^{-1}(X-\mu_1)\right] = \mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right)$. Since

$$\begin{aligned}
&(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2) = \left((X-\mu_1)+(\mu_1-\mu_2)\right)^T\Sigma_2^{-1}\left((X-\mu_1)+(\mu_1-\mu_2)\right) \\
&= (X-\mu_1)^T\Sigma_2^{-1}(X-\mu_1) + 2(\mu_1-\mu_2)^T\Sigma_2^{-1}(X-\mu_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2),
\end{aligned}$$

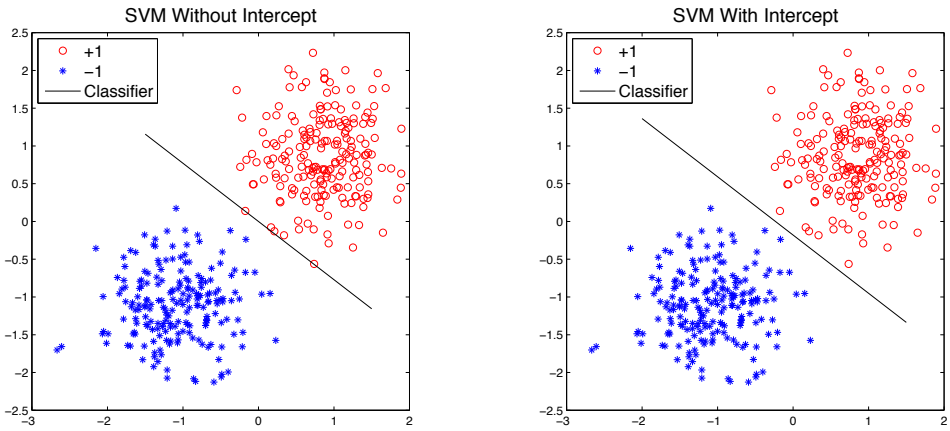

Figure 2: Results of running SVM without and with intercept term.

taking expectation on both sides gives us

$$\mathbb{E}_{f_1}\left[(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)\right] \;=\; \mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right)+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2).$$

Note that $D(f_1\|f_2)$ can be written as $\mathbb{E}_{f_1}\left[\log\frac{f_1(X)}{f_2(X)}\right]$, where the expectation is taken over $X\sim f_1$. For $f_1=N(\mu_1,\Sigma_1)$ and $f_2=N(\mu_2,\Sigma_2)$,

$$\begin{aligned}
\mathbb{E}_{f_1}\left[\log\frac{f_1(X)}{f_2(X)}\right] &= \mathbb{E}_{f_1}\left[\log\frac{|\Sigma_2|^{1/2}\exp\left\{-\frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)\right\}}{|\Sigma_1|^{1/2}\exp\left\{-\frac{1}{2}(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)\right\}}\right]\\
&= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}+\frac{1}{2}\mathbb{E}_{f_1}\left[(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)-(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)\right]\\
&= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|}+\mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right)+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)-d\right],
\end{aligned}$$

where in the last equation we plug in the two identities derived earlier.

An alternative way is to express the multivariate Gaussian distribution in the form of exponential family, and then explore the general formula of the Kullback-Leibler divergence for exponential family. Assume $X\sim N(\mu,\Sigma)$, then its density can be reformed as

$$f(x\mid\theta)=\exp\left\{\langle\theta,T(x)\rangle-A(\theta)\right\},$$

where $\theta=(\theta_1,\theta_2)$, $\mathrm{T}(x)=(T_1(x),T_2(x))$ and

$$\begin{aligned}
\theta_1=\Sigma^{-1} \quad &, \quad \theta_2=\Sigma^{-1}\mu\\
T_1(x)=-\frac{1}{2}xx^T \quad &, \quad T_2(x)=x\\
A(\theta) &= \frac{1}{2}\left[d\log(2\pi)-\log|\theta_1|+\theta_2^T\theta_1^{-1}\theta_2\right]\\
&= \frac{1}{2}\left[d\log(2\pi)+\log|\Sigma|+\mu^T\Sigma^{-1}\mu\right].
\end{aligned}$$

The general formula of the Kullback-Leibler divergence for exponential family is well-known:

$$D(\theta^1\|\theta^2) = \mathbb{E}_{f_1}\left[\log\frac{f_1(X)}{f_2(X)}\right] = A(\theta^2) - A(\theta^1) + \left\langle \mu^1, \theta^1 - \theta^2\right\rangle,$$

where $\mu^1 = \mathbb{E}_{f_1}[T(X)]$.

In our case,

$$\theta^1 - \theta^2 = \left(\Sigma_1^{-1} - \Sigma_2^{-1}, \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2\right)$$

and

$$\begin{aligned}\mu^1 = \mathbb{E}_{f_1}[T(X)] &= (\mathbb{E}_{f_1}[T_1(X)], \mathbb{E}_{f_1}[T_2(X)]) \\ &= \left(\mathbb{E}_{f_1}\left[-\frac{1}{2}XX^T\right], \mathbb{E}_{f_1}[X]\right) \\ &= \left(-\frac{1}{2}\left(\Sigma_1 + \mu_1\mu_1^T\right), \mu_1\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\left\langle \mu^1, \theta^1 - \theta^2\right\rangle &= -\frac{1}{2}\mathrm{tr}\left((\Sigma_1 + \mu_1\mu_1^T)^T(\Sigma_1^{-1} - \Sigma_2^{-1})\right) + \mu_1^T(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) \\ &= -\frac{1}{2}d + \frac{1}{2}\mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + \frac{1}{2}\mu_1^T\left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)\mu_1 - \mu_1^T\Sigma_2^{-1}\mu_2.\end{aligned}$$

Put all together,

$$\begin{aligned}D(\theta^1\|\theta^2) &= A(\theta^2) - A(\theta^1) + \left\langle \mu^1, \theta^1 - \theta^2\right\rangle \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} + \mu_2^T\Sigma_2^{-1}\mu_2 - \mu_1^T\Sigma_1^{-1}\mu_1\right] + \left\langle \mu^1, \theta^1 - \theta^2\right\rangle \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} + \mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1 - \mu_2)^T\Sigma_2^{-1}(\mu_1 - \mu_2) - d\right].\end{aligned}$$

(b) Denote $f_+ = N(\mu_{+1}, \Sigma)$ and $f_- = N(\mu_{-1}, \Sigma)$, then $f = pf_+ + (1-p)f_-$ is the marginal density of X. Hence,

$$\begin{aligned}\mathbb{E}[\sqrt{\eta(X)(1-\eta(X))}] &= \int \sqrt{\eta(x)(1-\eta(x))}f(x)\,dx \\ &= \int \sqrt{\frac{pf_+(x)}{f(x)}\,\frac{(1-p)f_-(x)}{f(x)}}f(x)\,dx \\ &= \sqrt{p(1-p)}\int \sqrt{f_+(x)f_-(x)}\,dx \\ &= \sqrt{p(1-p)}\int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{4}h(x)\right\}\,dx,\end{aligned}$$

where $h(x) = (x-\mu_{+1})^T\Sigma^{-1}(x-\mu_{+1}) + (x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})$.

Denote $\bar{\mu} = (\mu_{+1} + \mu_{-1})/2$ and complete the square for $h(x)$,

$$\begin{aligned} h(x) &= 2(x-\bar{\mu})^T\Sigma^{-1}(x-\bar{\mu}) + \frac{1}{2}(\mu_{+1}-\mu_{-1})^T\Sigma^{-1}(\mu_{+1}+\mu_{-1}) \\ &= 2(x-\bar{\mu})^T\Sigma^{-1}(x-\bar{\mu}) + D(\mu_{+1}\|\mu_{-1}). \end{aligned}$$

Plug $h(x)$ back into the computation, we obtain

$$\begin{aligned} \mathbb{E}[\sqrt{\eta(X)(1-\eta(X))}] &= \sqrt{p(1-p)}\exp\left\{-\frac{1}{4}D(\mu_{+1}\|\mu_{-1})\right\} \times \\ &\quad \int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\bar{\mu})^T\Sigma^{-1}(x-\bar{\mu})\right\} dx \\ &= \sqrt{p(1-p)}\exp\left\{-\frac{1}{4}D(\mu_{+1}\|\mu_{-1})\right\}, \end{aligned}$$

where in the first equation we recognize the integrand is the density of $N(\bar{\mu}, \Sigma)$.

Solution 1.5

(a) The Lagrangian function is

$$\begin{aligned} L(\theta, \upsilon; \alpha) &= \frac{1}{2}\|\theta\|_2^2 + C\sum_{i=1}^n \upsilon_i + \sum_{i=1}^n \alpha_i\left(1 - \upsilon_i - y^{(i)}\langle\theta,\, x^{(i)}\rangle\right) \\ &= \underbrace{\frac{1}{2}\|\theta\|_2^2 - \left\langle \theta, \sum_{i=1}^n \alpha_i y^{(i)} x^{(i)} \right\rangle}_{L_1(\theta;\alpha)} + \underbrace{\sum_{i=1}^n \upsilon_i(C-\alpha_i)}_{L_2(\upsilon;\alpha)} + \sum_{i=1}^n \alpha_i \end{aligned}$$

The dual function

$$q(\alpha) = \inf_{\theta\in\mathbb{R}^d, \upsilon\in\mathbb{R}^n_+} L(\theta, \upsilon; \alpha) = \inf_{\theta\in\mathbb{R}^d} L_1(\theta;\, \alpha) + \inf_{\upsilon\in\mathbb{R}^n_+} L_2(\upsilon;\, \alpha) + \sum_{i=1}^n \alpha_i.$$

$L_1(\theta;\, \alpha)$ is a quadratic function of θ so it is minimized at $\theta = \sum_{i=1}^n \alpha_i y^{(i)} x^{(i)}$.

$$\inf_{\upsilon\in\mathbb{R}^n_+} L_2(\upsilon;\, \alpha) = \begin{cases} 0 & \text{if } \alpha_i \le C, \forall i \\ -\infty & \text{otherwise.} \end{cases}$$

Put them together,

$$q(\alpha) = \begin{cases} -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y^{(i)}y^{(j)}\left\langle x^{(i)}, x^{(j)}\right\rangle + \sum_{i=1}^n \alpha_i & \text{if } \alpha_i \le C, \forall i \\ -\infty & \text{otherwise.} \end{cases}$$

The dual program is

$$\begin{aligned} &\max_{\alpha\in\mathbb{R}^n} \quad -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y^{(i)}y^{(j)}\left\langle x^{(i)}, x^{(j)}\right\rangle + \sum_{i=1}^n \alpha_i \\ &\text{such that} \quad 0 \le \alpha_i \le C \qquad \forall i = 1, \ldots, n \end{aligned}$$

(b) The modified primal program is

$$\begin{aligned} \min_{\theta\in\mathbb{R}^d,\upsilon\in\mathbb{R}^n_+,b\in\mathbb{R}} \quad & \frac{1}{2}\|\theta\|_2^2 + C\sum_{i=1}^n \upsilon_i \\ \text{such that} \quad & y^{(i)}\left(\langle\theta,\, x^{(i)}\rangle + b\right) \geq 1-\upsilon_i \qquad \forall i = 1,\ldots,n \end{aligned}$$

Its Lagrangian function is

$$\begin{aligned} L(\theta,\upsilon,b;\alpha) &= \frac{1}{2}\|\theta\|_2^2 + C\sum_{i=1}^n \upsilon_i + \sum_{i=1}^n \alpha_i\left(1-\upsilon_i - y^{(i)}\langle\theta,\, x^{(i)}\rangle - y^{(i)}b\right) \\ &= L(\theta,\upsilon;\alpha) + \underbrace{\left(-b\sum_{i=1}^n \alpha_i y^{(i)}\right)}_{L_3(b;\alpha)} \end{aligned}$$

Therefore, the dual function $q(\alpha) = \inf_{\theta\in\mathbb{R}^d,\upsilon\in\mathbb{R}^n_+} L(\theta,\upsilon;\alpha) + \inf_{b\in\mathbb{R}} L_3(b;\alpha)$. Note that

$$\inf_{b\in\mathbb{R}} L_3(b;\alpha) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \alpha_i y^{(i)} = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Combined with the result in part (a), the dual function associated with this modified program is

$$q(\alpha) = \begin{cases} -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y^{(i)}y^{(j)}\left\langle x^{(i)}, x^{(j)}\right\rangle + \sum_{i=1}^n \alpha_i & \text{if } \sum_{i=1}^n \alpha_i y^{(i)} = 0 \text{ and } \alpha_i \leq C, \forall i \\ -\infty & \text{otherwise.} \end{cases}$$

The corresponding dual program is

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^n} \quad & -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y^{(i)}y^{(j)}\left\langle x^{(i)}, x^{(j)}\right\rangle + \sum_{i=1}^n \alpha_i \\ \text{such that} \quad & 0 \leq \alpha_i \leq C \qquad \forall i = 1,\ldots,n \\ & \sum_{i=1}^n \alpha_i y^{(i)} = 0 \end{aligned}$$

Solution 1.6

(a) By the definition of PSD kernel function, the matrix $K \in \mathbb{R}^{2\times 2}$ formed as

$$K = \begin{bmatrix} \mathbb{K}(x^{(1)},x^{(1)}) & \mathbb{K}(x^{(1)},x^{(2)}) \\ \mathbb{K}(x^{(2)},x^{(1)}) & \mathbb{K}(x^{(2)},x^{(2)}) \end{bmatrix}$$

is symmetric and PSD. Hence its determinant is nonnegative,

$$\begin{aligned} |K| &= \mathbb{K}(x^{(1)},x^{(1)})\mathbb{K}(x^{(2)},x^{(2)}) - \mathbb{K}(x^{(1)},x^{(2)})\mathbb{K}(x^{(2)},x^{(1)}) \\ &= \mathbb{K}(x^{(1)},x^{(1)})\mathbb{K}(x^{(2)},x^{(2)}) - \mathbb{K}(x^{(1)},x^{(2)})^2 \\ &\geq 0, \end{aligned}$$

which immediately implies the desired inequality.

(b) It boils down to show that given any $x^{(1)}, x^{(2)}, \ldots, x^{(n)} \in \mathcal{S}$, the matrix $K \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathbb{K}(x^{(i)}, x^{(j)}) = \left\langle \Phi(x^{(i)}), \Phi(x^{(j)}) \right\rangle$ is symmetric and PSD. Since $K_{ij} = \left\langle \Phi(x^{(i)}), \Phi(x^{(j)}) \right\rangle = \left\langle \Phi(x^{(j)}), \Phi(x^{(i)}) \right\rangle = K_{ji}$, K is symmetric. The property of being PSD follows from the fact that for any given $\alpha \in \mathbb{R}^n$,

$$\begin{aligned}
\alpha^T K \alpha &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \left\langle \Phi(x^{(i)}), \Phi(x^{(j)}) \right\rangle \\
&= \left\langle \sum_{i=1}^{n} \alpha_i \Phi(x^{(i)}), \sum_{j=1}^{n} \alpha_j \Phi(x^{(j)}) \right\rangle \\
&= \left\| \sum_{i=1}^{n} \alpha_i \Phi(x^{(i)}) \right\|_{\mathbb{R}^D}^2 \\
&\geq 0
\end{aligned}$$

(c)

$$\begin{aligned}
\mathbb{K}(x, y) &= (\langle x, y \rangle)^3 \\
&= \left(\sum_{i=1}^{d} x_i y_i \right) \left(\sum_{j=1}^{d} x_j y_j \right) \left(\sum_{k=1}^{d} x_k y_k \right) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} x_i x_j x_k y_i y_j y_k \\
&= \sum_{i,j,k=1}^{d} (x_i x_j x_k)(y_i y_j y_k).
\end{aligned}$$

If we define a mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^{d^3}$ where $\Phi(x) = (x_i x_j x_k)_{i,j,k=1}^{d}$, then $\mathbb{K}(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Therefore, $\mathbb{K}$ is PSD by part (b).

(d) Note that the kernel function $\mathbb{K}(A, B) = 2^{|A \cap B|} = |\mathcal{P}(A \cap B)|$ counts the number of common subsets A and B share. It motivates us to define a mapping $\Phi : \mathcal{S} \to \{0, 1\}^{|\mathcal{S}|}$ where the image of a set A is indexed by all subsets of Ω:

$$\Phi(A)_U = \mathbb{I}[U \subseteq A], \text{ where } U \in \mathcal{S}.$$

It follows that

$$\begin{aligned}
\mathbb{K}(A, B) &= |\mathcal{P}(A \cap B)| \\
&= \sum_{U \in \mathcal{S}} \mathbb{I}[U \subseteq A \cap B] \\
&= \sum_{U \in \mathcal{S}} \mathbb{I}[U \subseteq A] \mathbb{I}[U \subseteq B] \\
&= \sum_{U \in \mathcal{S}} \Phi(A)_U \Phi(B)_U \\
&= \langle \Phi(A), \Phi(B) \rangle.
\end{aligned}$$

Therefore, $\mathbb{K}$ is PSD by part (b).