

Solutions 2
 Spring 2009

Solution 2.1

For part(a) and (c), we will show that for any natural number n and vectors $\{x^{(i)}, i = 1, \dots, n\}$, the matrix $K \in \mathbb{R}^{n \times n}$ with entries $K(i, j) = \mathbb{K}(x^{(i)}, x^{(j)})$ is symmetric and PSD. Define matrices $K_k \in \mathbb{R}^{n \times n}$ where $K_k(i, j) = \mathbb{K}_k(x^{(i)}, x^{(j)})$, $k = 1, 2$.

For part(d) and (e), we will show that for any natural number n and $\{A_i, i = 1, \dots, n\}$, the matrices $K, \bar{K} \in \mathbb{R}^{n \times n}$ with entries $K(i, j) = \mathbb{K}(A_i, A_j)$ and $\bar{K}(i, j) = \bar{\mathbb{K}}(A_i, A_j)$, respectively, are symmetric and PSD.

- (a) True. In this case, $\mathbb{K}(x^{(i)}, x^{(j)}) = \lambda_1 \mathbb{K}_1(x^{(i)}, x^{(j)}) + \lambda_2 \mathbb{K}_2(x^{(i)}, x^{(j)})$ hence $K = \lambda_1 K_1 + \lambda_2 K_2$. K is apparently symmetric. The property of being PSD follows from the fact that for any given $\alpha \in \mathbb{R}^n$,

$$\alpha^T K \alpha = \lambda_1 \alpha^T K_1 \alpha + \lambda_2 \alpha^T K_2 \alpha \geq 0,$$

because $\lambda_k \geq 0$ and $\alpha^T K_k \alpha \geq 0, k = 1, 2$.

- (b) False. Let $\mathcal{X} = \{0, 1\}$ and $\mathbb{K}(x, y) = |x - y|$, which is elementwise non-negative. However, the matrix formed as

$$K = \begin{bmatrix} \mathbb{K}(0, 0) & \mathbb{K}(0, 1) \\ \mathbb{K}(1, 0) & \mathbb{K}(1, 1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

is not a PSD matrix.

- (c) True. In this case, $\mathbb{K}(x^{(i)}, x^{(j)}) = \mathbb{K}_1(x^{(i)}, x^{(j)}) \mathbb{K}_2(x^{(i)}, x^{(j)})$ hence $K = K_1 \circ K_2$ where \circ denotes the Hadamard product. Since K_1 is symmetric and PSD, it could be the covariance matrix of a multivariate Gaussian random vector $Y = [Y_1, \dots, Y_n]$. Let $Y \sim N(0, K_1)$, which implies $\mathbb{E}[Y_i Y_j] = K_1(i, j)$. Similarly, let $Z = [Z_1, \dots, Z_n]$ be independent with Y and $Z \sim N(0, K_2)$. Now define another random vector $W = [Y_1 Z_1, \dots, Y_n Z_n]$ and denote its covariance matrix as Σ .

$$\begin{aligned} \Sigma(i, j) &= \mathbb{E}[Y_i Z_i Y_j Z_j] - \mathbb{E}[Y_i Z_i] \mathbb{E}[Y_j Z_j] \\ &= \mathbb{E}[Y_i Y_j] \mathbb{E}[Z_i Z_j] - \mathbb{E}[Y_i] \mathbb{E}[Z_i] \mathbb{E}[Y_j] \mathbb{E}[Z_j] \\ &= K_1(i, j) K_2(i, j) \\ &= K(i, j). \end{aligned}$$

Therefore, $\Sigma = K$ and K must be symmetric and PSD.

- (d) In this case, we can write $K(i, j)$ as

$$K(i, j) = \mathbb{K}(A_i, A_j) = \mathbb{P}(A_i, A_j) - \mathbb{P}(A_i) \mathbb{P}(A_j) = \mathbb{E}[\mathbb{I}[A_i] \mathbb{I}[A_j]] - \mathbb{E}[\mathbb{I}[A_i]] \mathbb{E}[\mathbb{I}[A_j]].$$

Therefore, K is the covariance matrix of random vector $[\mathbb{I}[A_1], \dots, \mathbb{I}[A_n]]$, which must be symmetric and PSD.

(e) Since \mathcal{E} is finite, we can write $\bar{\mathbb{K}}(A, B)$ as

$$\bar{\mathbb{K}}(A, B) = \sum_{x \in A, y \in B} \mathbb{K}(x, y) = \sum_{x, y \in \mathcal{E}} \mathbb{K}(x, y) \mathbb{I}[x \in A] \mathbb{I}[y \in B].$$

Since $\bar{\mathbb{K}}(A, B) = \bar{\mathbb{K}}(B, A)$, \bar{K} is symmetric. Moreover, for any $\alpha \in \mathbb{R}^n$,

$$\begin{aligned} \alpha^T \bar{K} \alpha &= \sum_{i, j=1}^n \alpha_i \alpha_j \bar{\mathbb{K}}(A_i, A_j) = \sum_{i, j=1}^n \alpha_i \alpha_j \sum_{x, y \in \mathcal{E}} \mathbb{K}(x, y) \mathbb{I}[x \in A_i] \mathbb{I}[y \in A_j] \\ &= \sum_{x, y \in \mathcal{E}} \mathbb{K}(x, y) \sum_{i, j=1}^n \alpha_i \alpha_j \mathbb{I}[x \in A_i] \mathbb{I}[y \in A_j] \\ &= \sum_{x, y \in \mathcal{E}} \mathbb{K}(x, y) \left(\sum_i \alpha_i \mathbb{I}[x \in A_i] \right) \left(\sum_j \alpha_j \mathbb{I}[y \in A_j] \right) \\ &\geq 0, \end{aligned}$$

where the last inequality follows from the fact the matrix $K \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{E}|}$ with entries $K(x, y) = \mathbb{K}(x, y)$ is PSD.

Solution 2.2

See script.m for details of implementation.

- We pad additional 1 to each data vector to perform the linear regression with intercept. The sum of squared errors on 10 dimensions is 0.35407.
- The data matrix has been centered first to compute the SVD. The sum of squared errors on the two PCA dimensions is 34.7263.
- There are two subtle steps in the implementation kernel PCA. See B. Schoelkopf et al.'s original paper for more details.

- As in standard PCA, the feature vectors have to be centered first. This corresponds to replacing the original kernel matrix K by

$$K - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' K - \frac{1}{n} K \mathbf{1}_n \mathbf{1}_n' + \frac{1}{n^2} (\mathbf{1}_n' K \mathbf{1}_n) \mathbf{1}_n \mathbf{1}_n'.$$

- After obtaining eigenvector $\alpha^{(j)}$ with eigenvalue $\lambda_j, j = 1, \dots, k$ by solving $K\alpha = \lambda\alpha$, normalize $\alpha^{(j)}$ by requiring $\lambda_j \langle \alpha^{(j)}, \alpha^{(j)} \rangle = 1$.

To choose the bandwidth parameter, 10-fold cross validation is used. Figure 1 shows the average sum of squared errors for different values of σ . For optimal $\sigma = 1.5557$, the sum of squared errors on the test set is 32.4698.

Solution 2.3

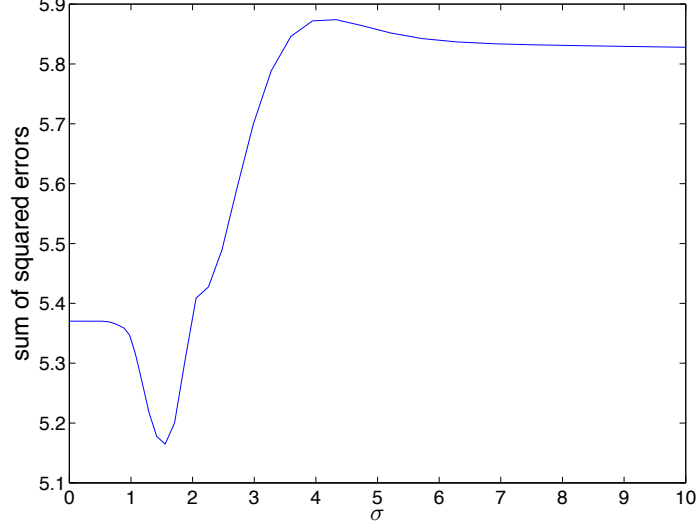


Figure 1: Cross validation for choosing σ .

- (a) We claim the set of functions $\{\cos(nx), n \in \mathbb{N}\}$ are the eigenfunctions. To see this, note that

$$\int_0^{2\pi} \mathbb{K}(x, y) \cos(ny) dy = \int_0^{2\pi} \sum_{\ell=0}^{\infty} w_{\ell} \cos(\ell(x-y)) \cos(ny) dy.$$

Dominated convergence theorem allows us to interchange the integral and sum because

$$\left| \sum_{\ell=0}^{\infty} w_{\ell} \cos(\ell(x-y)) \cos(ny) \right| \leq \sum_{\ell=0}^{\infty} |w_{\ell} \cos(\ell(x-y)) \cos(ny)| \leq \sum_{\ell=0}^{\infty} |w_{\ell}| < \infty.$$

Therefore,

$$\begin{aligned} \int_0^{2\pi} \mathbb{K}(x, y) \cos(ny) dy &= \sum_{\ell=0}^{\infty} w_{\ell} \left(\int_0^{2\pi} \cos(\ell(x-y)) \cos(ny) dy \right) \\ &= \sum_{\ell=0}^{\infty} w_{\ell} \left(\int_0^{2\pi} (\cos(\ell x) \cos(\ell y) + \sin(\ell x) \sin(\ell y)) \cos(ny) dy \right) \\ &= \sum_{\ell=0}^{\infty} w_{\ell} \cos(\ell x) (\pi \mathbb{I}[\ell = n]) \\ &= \pi w_n \cos(nx), \end{aligned}$$

which implies that $\cos(nx)$ is an eigenfunction with eigenvalue πw_n .

By the similar calculation, we also identify that $\{\sin(nx), n \in \mathbb{N}\}$ are also eigenfunctions with their corresponding eigenvalues $\{\pi w_n, n \in \mathbb{N}\}$. By the fact that $\{\cos(nx), \sin(nx), n \in \mathbb{N}\}$ is a basis for $L^2([0, 2\pi])$, we conclude that we found a complete list of eigenfunctions.

(b) Let $f(x)$ be an eigenfunction with eigenvalue λ , then

$$\begin{aligned} \int_0^1 \mathbb{K}(x, y) f(y) dy &= \int_0^1 (1 + 2xy + x^2 y^2) f(y) dy \\ &= \int_0^1 f(y) dy + 2x \int_0^1 y f(y) dy + x^2 \int_0^1 y^2 f(y) dy \\ &= \lambda f(x) \end{aligned}$$

– If $\lambda \neq 0$, then the last identity above implies $f(x) = a_0 + a_1 x + a_2 x^2$ for some $a_0, a_1, a_2 \in \mathbb{R}$. Substituting it back to the last identity results in the following system of equations:

$$\begin{bmatrix} 1 & 1/2 & 1/3 \\ 1 & 2/3 & 1/2 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \lambda \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}.$$

Solving the above standard eigenvalue/eigenvector problem gives us 3 eigenfunctions with non-zero eigenvalues:

$$f_1(x) = -0.6330 - 0.7292x - 0.2600x^2, \quad \lambda_1 = 1.7129 \quad (1)$$

$$f_2(x) = -0.5773 + 0.6795x + 0.4527x^2, \quad \lambda_2 = 0.1501 \quad (2)$$

$$f_3(x) = 0.1249 - 0.7105x + 0.6925x^2, \quad \lambda_3 = 0.0036 \quad (3)$$

– If $\lambda = 0$, then any function in $L^2([0, 1])$ that satisfies the following three conditions is an eigenfunction with eigenvalue 0:

$$\int_0^1 f(x) dx = 0, \quad \int_0^1 x f(x) dx = 0, \quad \int_0^1 x^2 f(x) dx = 0.$$

Solution 2.4

The projection of $\Phi(x^{(i)})$ onto f has coordinate

$$\frac{\langle f, \Phi(x^{(i)}) \rangle_{\mathcal{H}}}{\|f\|_{\mathcal{H}}}.$$

The sample variance of $\{y^{(i)}, i = 1, \dots, n\}$ is

$$\begin{aligned} \widehat{\text{Var}}[Y] &= \widehat{\mathbb{E}}[Y^2] - \left(\widehat{\mathbb{E}}[Y]\right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\langle f, \Phi(x^{(i)}) \rangle_{\mathcal{H}}^2}{\|f\|_{\mathcal{H}}^2} - \left(\frac{1}{n} \sum_{i=1}^n \frac{\langle f, \Phi(x^{(i)}) \rangle_{\mathcal{H}}}{\|f\|_{\mathcal{H}}}\right)^2 \\ &= \frac{1}{n\|f\|_{\mathcal{H}}^2} \sum_{i=1}^n \left\langle \sum_{j=1}^n \alpha_j \Phi(x^{(j)}), \Phi(x^{(i)}) \right\rangle_{\mathcal{H}}^2 - \frac{1}{n^2\|f\|_{\mathcal{H}}^2} \left(\sum_{i=1}^n \left\langle \sum_{j=1}^n \alpha_j \Phi(x^{(j)}), \Phi(x^{(i)}) \right\rangle_{\mathcal{H}} \right)^2 \\ &= \frac{1}{n\alpha^T K \alpha} \alpha^T K^2 \alpha - \frac{1}{n^2 \alpha^T K \alpha} (\alpha^T K 1)^2, \end{aligned}$$

which only depends on the kernel matrix K where $K(i, j) = \mathbb{K}(x^{(i)}, x^{(j)})$, $i, j = 1, \dots, n$.

Solution 2.5

- (a) We'd like to find a sphere with center c and radius r containing all the data points such that the radius r is minimum. The optimization problem is:

$$\begin{aligned} \min_{c,r} \quad & r^2 \\ \text{such that} \quad & \left\| \Phi(x^{(i)}) - c \right\|_{\mathcal{H}}^2 \leq r^2, i = 1, \dots, n. \end{aligned}$$

The Lagrangian function is

$$L(c, r; \alpha) = r^2 + \sum_{i=1}^n \alpha_i \left(\left\| \Phi(x^{(i)}) - c \right\|_{\mathcal{H}}^2 - r^2 \right)$$

Taking derivate w.r.t. c and r and setting to zero gives us

$$c = \sum_{i=1}^n \alpha_i \Phi(x^{(i)}), \quad \sum_{i=1}^n \alpha_i = 1$$

Substituting back into the Lagrangian function, we get the dual function

$$\begin{aligned} q(\alpha) &= \inf_{c,r} L(c, r; \alpha) \\ &= \sum_{i=1}^n \alpha_i \left\| \Phi(x^{(i)}) - \sum_{j=1}^n \alpha_j \Phi(x^{(j)}) \right\|_{\mathcal{H}}^2 \\ &= \sum_{i=1}^n \alpha_i \left\langle \Phi(x^{(i)}), \Phi(x^{(i)}) \right\rangle_{\mathcal{H}} - \sum_{i,j=1}^n \alpha_i \alpha_j \left\langle \Phi(x^{(i)}), \Phi(x^{(j)}) \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^n \alpha_i K(i, i) - \sum_{i,j=1}^n \alpha_i \alpha_j K(i, j). \end{aligned}$$

The corresponding dual program is

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \sum_{i=1}^n \alpha_i K(i, i) - \sum_{i,j=1}^n \alpha_i \alpha_j K(i, j) \\ \text{such that} \quad & \alpha_i \geq 0 \quad \forall i = 1, \dots, n \\ & \sum_{i=1}^n \alpha_i = 1, \end{aligned}$$

which only depends on the kernel matrix K .

- (b) As in the extension of a hard-margin SVM to a soft-margin SVM, we introduce the slack variables ξ_i , $i = 1, \dots, n$ and a parameter $C \geq 0$ to control the trade-off between

minimizing the radius and controlling the slack variables. The optimization problem now becomes:

$$\begin{aligned} \min_{c,r,\xi} \quad & r^2 + C \sum_{i=1}^n \xi_i \\ \text{such that} \quad & \left\| \Phi(x^{(i)}) - c \right\|_{\mathcal{H}}^2 \leq r^2 + \xi_i, i = 1, \dots, n \\ & \xi_i \geq 0, i = 1, \dots, n. \end{aligned}$$

The Lagrangian function is

$$L(c, r, \xi; \alpha, \beta) = r^2 + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i \left(\left\| \Phi(x^{(i)}) - c \right\|_{\mathcal{H}}^2 - r^2 - \xi_i \right) - \sum_{i=1}^n \beta_i \xi_i$$

Taking derivate w.r.t. c, r and ξ and setting to zero gives us

$$c = \sum_{i=1}^n \alpha_i \Phi(x^{(i)}), \quad \sum_{i=1}^n \alpha_i = 1, \quad C - \alpha_i - \beta_i = 0.$$

Since $\beta_i \geq 0$, the last equation implies that $\alpha_i \leq C$. Substituting, we obtain the dual function

$$q(\alpha) = \inf_{c,r,\xi} L(c, r; \alpha, \beta) = \sum_{i=1}^n \alpha_i K(i, i) - \sum_{i,j=1}^n \alpha_i \alpha_j K(i, j).$$

The corresponding dual program is

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \sum_{i=1}^n \alpha_i K(i, i) - \sum_{i,j=1}^n \alpha_i \alpha_j K(i, j) \\ \text{such that} \quad & 0 \leq \alpha_i \leq C \quad \forall i = 1, \dots, n \\ & \sum_{i=1}^n \alpha_i = 1. \end{aligned}$$

Solution 2.6

- (a) An elementary result shows that $t\mathbb{I}(z \geq t) \leq z$ for any $z \geq 0$. Taking expectation on both sides gives us $t\mathbb{P}[Z \geq t] \leq \mathbb{E}[Z]$. If $t > 0$, we can divide both sides by t and preserve the direction of inequality: $\mathbb{P}[Z \geq t] \leq \mathbb{E}[Z]/t$.
- (b)

$$\begin{aligned} \mathbb{E}[\exp(sX)] &= \int e^{sx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \int \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2-2\sigma^2sx}{2\sigma^2}} dx \\ &= e^{\frac{\sigma^2s^2}{2}} \int \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\sigma^2s)^2}{2\sigma^2}} dx \\ &= e^{\frac{\sigma^2s^2}{2}}, \end{aligned}$$

where in the last equation we recognize the integrand is the density of $N(\sigma^2 s, \sigma^2)$. Therefore, X is sub-Gaussian with parameter σ .

- (c) Note that $\mathbb{E}[X^n] = 0$ for each odd n and $\mathbb{E}[X^n] = 1$ for each even n . By Taylor expansion,

$$\begin{aligned}\mathbb{E}[e^{sX}] &= \sum_{n=0}^{\infty} \frac{s^n \mathbb{E}[X^n]}{n!} \\ &= \sum_{n=0}^{\infty} \frac{s^{2n}}{(2n)!} \\ &\leq \sum_{n=0}^{\infty} \frac{s^{2n}}{2^n n!} \\ &= e^{s^2/2}.\end{aligned}$$

Hence the Bernoulli random variable is sub-Gaussian with parameter $\sigma = 1$. The above argument can be easily generalized to any symmetric and bounded random variable. In this case, we have $\mathbb{E}[X^n] = 0$ for each odd n and $\mathbb{E}[X^n] \leq B^n$ for each even n .

$$\mathbb{E}[e^{sX}] \leq \sum_{n=0}^{\infty} \frac{s^{2n} B^{2n}}{(2n)!} \leq \sum_{n=0}^{\infty} \frac{(sB)^{2n}}{2^n n!} = e^{s^2 B^2/2},$$

which implies it is sub-Gaussian with parameter $\sigma = B$. However, the symmetric property is not necessary if we apply the more powerful tool—Hoeffding’s inequality: if the zero-mean random variable X satisfies $a \leq X \leq b$ almost surely, then for any $s \in \mathbb{R}$,

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8},$$

showing X is sub-Gaussian with parameter $\sigma = (b-a)/2$. Applying it to Bernoulli case immediately leads us to $\sigma = (1+1)/2 = 1$, as claimed above.

- (d) By Chernoff bound,

$$\begin{aligned}\mathbb{P}[X > t] &\leq \mathbb{P}[X \geq t] \\ &\leq \inf_{s>0} \{\mathbb{E}[e^{sX}]/e^{st}\} \\ &\leq \inf_{s>0} \{e^{-st} e^{\sigma^2 s^2/2}\} \\ &= e^{-t^2/2\sigma^2},\end{aligned}$$

where the infimum is achieved at $s = t/\sigma^2$. By similar argument, $\mathbb{P}[-X < -t] \leq e^{-t^2/2\sigma^2}$. Putting them together gives us the two-sided bound: $\mathbb{P}[|X| > t] \leq 2e^{-t^2/2\sigma^2}$ for all $t \geq 0$.

(e)

$$\begin{aligned}\mathbb{P}\left[\max_{i=1,\dots,n} X_i > \sqrt{(2+\delta)\sigma^2 \log n}\right] &= \mathbb{P}\left[\bigcup_{i=1,\dots,n} \{X_i > \sqrt{(2+\delta)\sigma^2 \log n}\}\right] \\ &\leq n\mathbb{P}\left[X_1 > \sqrt{(2+\delta)\sigma^2 \log n}\right] \\ &\leq n \exp\left(-((2+\delta)\sigma^2 \log n) / 2\sigma^2\right) \\ &= n^{-\delta/2} \\ &\rightarrow 0 \quad \text{as } n \rightarrow +\infty.\end{aligned}$$

The first inequality is the union bound and the second one follows from the Chernoff bound derived in part (d).