

Homework Solution 2

Lecturer: Michael I. Jordan

GSI: Chao Chen

1. (a) The optimization problem:

$$\begin{aligned} \min_{c,r} \quad & r^2 \\ \text{s.t.} \quad & \|\phi(x_i) - c\|^2 \leq r^2 \quad i = 1, \dots, m \end{aligned}$$

The Lagrangian:

$$\mathcal{L}(c, r, \alpha) = r^2 + \sum_{i=1}^m \alpha_i (\|\phi(x_i) - c\|^2 - r^2) \quad \alpha_i \geq 0$$

Taking derivatives and setting to zero:

$$\begin{aligned} \frac{\partial \mathcal{L}(c, r, \alpha)}{\partial c} &= 2 \sum_{i=1}^m \alpha_i (\phi(x_i) - c) = 0 \\ \frac{\partial \mathcal{L}(c, r, \alpha)}{\partial r} &= 2r \left(1 - \sum_{i=1}^m \alpha_i \right) = 0 \\ \Rightarrow \quad c &= \sum_{i=1}^m \alpha_i \phi(x_i), \quad \sum_{i=1}^m \alpha_i = 1 \end{aligned}$$

Substituting back into the Lagrangian we get:

$$\begin{aligned} \mathcal{L}(c, r, \alpha) &= r^2 + \sum_{i=1}^m \alpha_i (\|\phi(x_i) - c\|^2 - r^2) \\ &= \sum_{i=1}^m \alpha_i \langle \phi(x_i) - c, \phi(x_i) - c \rangle \\ &= \sum_{i=1}^m \alpha_i \left(k(x_i, x_i) - 2 \sum_{j=1}^m \alpha_j k(x_i, x_j) + \sum_{j,k=1}^m \alpha_j \alpha_k k(x_j, x_k) \right) \\ &= \sum_{i=1}^m \alpha_i k(x_i, x_i) - \sum_{i,j=1}^m \alpha_i \alpha_j k(x_i, x_j) \end{aligned}$$

We therefore get the dualized optimization problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i k(x_i, x_i) - \sum_{i,j=1}^m \alpha_i \alpha_j k(x_i, x_j) \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i = 1 \\ & \alpha_i \geq 0, i = 1, \dots, m \end{aligned}$$

The KKT conditions are satisfied by the optimum solution α^* .

$$\alpha_i^* (\|\phi(x_i) - c^*\|^2 - r^{*2}) = 0. \quad i = 1, \dots, m$$

So the support vectors, for which the α_i^* are non-zero, are the ones that lie on the surface of the optimal hyper-sphere.

(b) We can define slack variables $\xi_i(c, r, x_i) = (\|c - \phi(x_i)\|^2 - r^2)_+$. The optimization problem is now:

$$\begin{aligned} \min_{c, r, \xi} \quad & r^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & \|\phi(x_i) - c\|^2 \leq r^2 + \xi_i \quad i = 1, \dots, m \\ & \xi_i \geq 0, \quad i = 1, \dots, m \end{aligned}$$

The Lagrangian:

$$\mathcal{L}(c, r, \alpha, \xi) = r^2 + C \sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i (\|\phi(x_i) - c\|^2 - r^2 - \xi_i) - \sum_{i=1}^m \beta_i \xi_i \quad \alpha_i, \beta_i \geq 0$$

Taking derivatives and setting to zero:

$$\begin{aligned} \frac{\partial \mathcal{L}(c, r, \alpha, \xi)}{\partial c} &= 2 \sum_{i=1}^m \alpha_i (\phi(x_i) - c) = 0 \\ \frac{\partial \mathcal{L}(c, r, \alpha, \xi)}{\partial r} &= 2r \left(1 - \sum_{i=1}^m \alpha_i \right) = 0 \\ \frac{\partial \mathcal{L}(c, r, \alpha, \xi)}{\partial \xi} &= C - \alpha_i - \beta_i = 0 \\ \Rightarrow \quad & c = \sum_{i=1}^m \alpha_i \phi(x_i), \quad \sum_{i=1}^m \alpha_i = 1, \quad \alpha_i \leq C \end{aligned}$$

Substituting back into the Lagrangian we get:

$$\begin{aligned} \mathcal{L}(c, r, \alpha, \xi) &= r^2 + C \sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i (\|\phi(x_i) - c\|^2 - r^2 - \xi_i) - \sum_{i=1}^m \beta_i \xi_i \\ &= \sum_{i=1}^m \alpha_i \langle \phi(x_i) - c, \phi(x_i) - c \rangle \\ &= \sum_{i=1}^m \alpha_i \left(k(x_i, x_i) - 2 \sum_{j=1}^m \alpha_j k(x_i, x_j) + \sum_{j, k=1}^m \alpha_j \alpha_k k(x_j, x_k) \right) \\ &= \sum_{i=1}^m \alpha_i k(x_i, x_i) - \sum_{i, j=1}^m \alpha_i \alpha_j k(x_i, x_j) \end{aligned}$$

We therefore get the dualized optimization problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i k(x_i, x_i) - \sum_{i, j=1}^m \alpha_i \alpha_j k(x_i, x_j) \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C, i = 1, \dots, m \end{aligned}$$

The KKT conditions are satisfied by the optimum solution α^* . So the support vectors, for which the α_i^* are non-zero, are the ones that lie on the surface or outside the optimal hyper-sphere.

(Thanks to Ariel Schwartz for providing Latex solution)

2. Let w be a vector that lies in the span of the vectors $\phi(x_i)$ ($1 \leq i \leq n$), and let $P_w(\phi(x))$ be the projection of $\phi(x)$ on w

$$w = \sum_{i=1}^n \alpha_i \phi(x_i) = X^T \alpha$$

$$P_w(\phi(x)) = \frac{\langle w, \phi(x) \rangle}{\|w\|^2} w$$

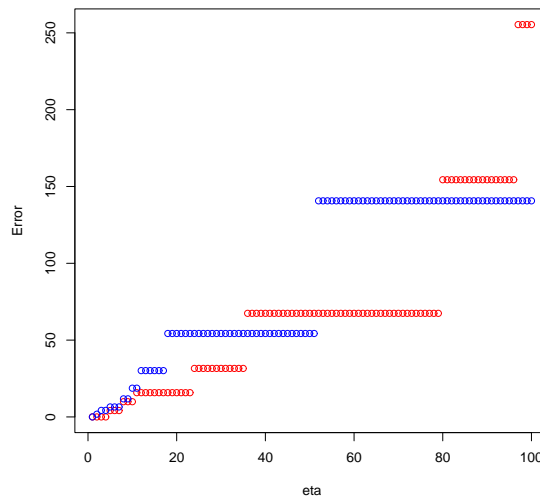
The sample variance of the norms of the projections on w is:

$$\begin{aligned} \sigma_w^2 &= \hat{E} \|P_w(\phi(x))\|^2 - (\hat{E} \|P_w(\phi(x))\|)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\langle w, \phi(x_i) \rangle^2}{\|w\|^2} - \left(\frac{1}{n} \sum_{i=1}^n \frac{\langle w, \phi(x_i) \rangle}{\|w\|} \right)^T \left(\frac{1}{n} \sum_{i=1}^n \frac{\langle w, \phi(x_i) \rangle}{\|w\|} \right) \\ &= \frac{1}{n \|w\|^2} w^T X^T X w - \frac{1}{n^2 \|w\|^2} (w^T X^T \mathbf{1})^T (w^T X^T \mathbf{1}) \\ &= \frac{1}{n \|w\|^2} \alpha^T X X^T X X^T \alpha - \frac{1}{n^2 \|w\|^2} \alpha^T X X^T \mathbf{1} \mathbf{1}^T X X^T \alpha \\ &= \frac{\frac{1}{n} \alpha^T K^2 \alpha - \frac{1}{n^2} (\alpha^T K \mathbf{1})^2}{\alpha^T K \alpha} \end{aligned}$$

3. Use the SVD for a random 10×10 matrix to get orthogonal matrix M to construct both A and B . The eigenvalue spectrum drops off exponentially for A , and according to an inverse power law for B :

$$\begin{aligned} \text{eigen}A &= 1024 \times 2^K, & K = 0, \dots, 9 \\ \text{eigen}B &= 1024 \times K^{-3}, & K = 1, \dots, 10 \end{aligned}$$

Here is the plot of the errors. The red points show the error as the eigenvalue drop exponentially, the blue points show the error as eigenvalue drop according to inverse power law. We can see that the error drops faster for A (exponential decay in eigenvalue) as the parameter η in incomplete Cholesky decomposition decrease.



4. (a) The sum of squared errors for the linear regression on 10 dimensions is 0.3540703.
 (b) The sum of squared errors for the linear regression on the two PCA dimensions 34.72634.
 (c) I used 10-fold cross validation on the training test to choose the bandwidth parameter σ^2 , trying different values, and found 1.5 to be the optimum value. The sum of squared errors for the linear regression on the two kernel PCA dimensions is 32.35.