1 Kernel ridge regression

a) Taking partial derivates, we get

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)} + \lambda \theta_j.$$

Writing in matrix notation and setting to zero, this is

$$X^T(X\theta - Y) + \lambda I \theta = 0.$$

Solving for θ, we get

$$\theta = (X^T X + \lambda I)^{-1} X^T Y.$$

b) Let Φ be the matrix we get by applying ϕ to X row-wise. That is, the ith row of Φ is $\phi(x^{(i)})$. Using the hint, we can rewrite θ as

$$\theta = \Phi^T (\lambda I + \Phi \Phi^T)^{-1} Y.$$

The i, jth entry of $\Phi\Phi^T$ is $\phi(x^{(i)})^T \phi(x^{(j)})$, so $\Phi\Phi^T$ is the Kernel matrix, K.

For a new observation x_{new}, the prediction is given by

$$\begin{aligned} y_{new} &= \theta^T \phi(x_{new}) \\ &= Y^T (\lambda I + K)^{-1} \Phi \phi(x_{new}). \end{aligned}$$

We only need to rewrite the expression $\Phi\phi(x_{new})$ in terms of the kernel function. To do so, note that ith entry of $\Phi\phi(x_{new})$ is $\phi(x^{(i)})^T \phi(x_{new}) = K(\phi(x^{(i)}), \phi(x_{new}))$. Finally, we can use the assumption that, for some α, $\theta = \sum_{i=1}^{m} \alpha_i \phi(x^{(i)}) = \Phi^T \alpha$, so $\theta^T = \alpha^T \Phi$. In our case, $\alpha^T = Y^T(\lambda I + K)^{-1}$. Combining, we get

$$y_{new} = \sum_{i=1}^{m} \alpha_i K(x^{(i)}, x_{new}).$$

All terms in the sum are calculated in terms of K, so we're done.

2 ℓ_2 norm soft margin SVMs

a) Permitting negative numbers does not affect the objective function, and the feasibility space corresponding to negative numbers is a strict subset of the space corresponding to positive numbers.

b) The Lagrangian is

$$\mathcal{L}(w,b,\alpha,\xi) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 + \sum_{i=1}^{m}\alpha_i\left[-y^{(i)}(w^Tx^{(i)}+b)+1-\xi_i\right].$$

c) Taking partials with respect to w, b and ξ, and setting to zero, we get

$$\nabla_w\mathcal{L} = w - \sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} = 0 \Longrightarrow w = \sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}$$

$$\frac{\partial\mathcal{L}}{\partial b} = \sum_{i=1}^{m}\alpha_i y^{(i)} = 0$$

$$\nabla_\xi\mathcal{L} = C\xi - \alpha \Longrightarrow C\xi = \alpha$$

d) We want to use the relationships above to rewrite $\mathcal{L}$ as a function of α. Starting with $\frac{1}{2}\|w\|^2$, we have

$$\begin{aligned}\frac{1}{2}\|w\|^2 &= \frac{1}{2}\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}\sum_{j=1}^{m}\alpha_j y^{(j)}x^{(j)} \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}x^{(i)}x^{(j)}\end{aligned}$$

Substituing the formulas for w and αinto the two right-most term, we get

$$\begin{aligned}&\frac{C}{2}\sum_{i=1}^{m}\xi_i^2 + \sum_{i=1}^{m}\alpha_i\left[-y^{(i)}(\sum_{j=1}^{m}\alpha_j y^{(j)}x^{(j)}x^{(i)}+b)+1-\xi_i\right] \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}x^{(i)}x^{(j)} + \sum_{i=1}^{m}\alpha_i - \frac{1}{C}\sum_{i=1}^{m}\alpha_i^2\end{aligned}$$

Combining these results, the dual problem is to maximize

$$-\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}x^{(i)}x^{(j)} + \sum_{i=1}^{m}\alpha_i - \frac{1}{C}\sum_{i=1}^{m}\alpha_i^2$$

with respect to α, such that $\sum_{i=1}^{m}\alpha_i y^{(i)} = 0$.