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

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

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 $\sum_{i=1}^m \alpha_i y^{(i)} = 0$.s

3 SVM with Gaussian kernel

a) Taking the provided hint and running with it, we've got

$$
\begin{aligned}
\left|f(x^{(j)}) - y^{(j)}\right| &= \left|\sum_{i=1}^{m} y^{(i)} K(x^{(i)}, x^{(j)}) - y^{(j)}\right| \\
&= \left|\sum_{i \neq j}^{m} y^{(i)} K(x^{(i)}, x^{(j)})\right| + \left|y^{(j)} - y^{(j)}\right| \\
&= \left|\sum_{i \neq j}^{m} y^{(i)} K(x^{(i)}, x^{(j)})\right| \\
&\leq \left|\sum_{i \neq j}^{m} y^{(i)} e^{-\frac{z^2}{\tau^2}}\right| \\
&\leq \sum_{i \neq j}^{m} \left|y^{(i)} e^{-\frac{z^2}{\tau^2}}\right| \\
&= \sum_{i \neq j}^{m} e^{-\frac{z^2}{\tau^2}} = (m-1)e^{-\frac{z^2}{\tau^2}} < 1
\end{aligned}
$$

Rearranging the last inequality, we get $\tau > \frac{z}{\sqrt{ln(m-1)}}$.

b) Yes, by design, the resulting classifier will achieve zero training error, though not necessarily zero test error.

c) No, the parameter C regulates the trade-off between bias and variance, or training and test error. When C is small, the objective function may obtain a minimum when w is small but ξ_i terms are potentially large. If the latter are large, the model could have non-zero training error.

4 Naive Bayes and SVMs for Spam Classification

SVMs outperform Naive Bayes on small sample szies, but Naive Bayes has a lower generalization error for sample sizes greater than 1000. See attached graphs.

5 Uniform Convergence

a) With generalization error γ, the probability of not making an error is 1-γ, and the probability of making no errors on m examples is $(1-\gamma)^m$. Using the hint, that $(1-\gamma)^m \leq e^{-\gamma m}$, we set

$$\begin{aligned}
1 - k\exp(-\gamma m) &= 1 - \delta \\
\exp(-\gamma m) &= \frac{\delta}{k} \\
\gamma &= \frac{1}{m}\ln\left(\frac{\delta}{k}\right) \\
\gamma &= \frac{1}{m}\ln\left(\frac{k}{\delta}\right).
\end{aligned}$$

b) Rearranging for m, we get $m = \frac{1}{\gamma}\ln\left(\frac{k}{\delta}\right)$.