Question 4

The ridge regression solves the problem:

$$\min_{\beta}\{\frac{1}{2}\|y - x\beta\|_2^2 + \frac{\lambda}{2}\beta^T\beta\}$$

and it has the optimal solution:

$\widehat{\beta_R} = (X^T X + \lambda I_d)^{-1} X^T y$

$= (\phi^T\phi + \lambda I_d)^{-1}\phi^T y$ (let $\phi(X) \in R^M$ be the feature vector where $X \in R^d$)

Let $B = \phi(X), P = (\lambda I_d)^{-1}, R = I_M$

$\widehat{\beta_R}$

$= (\phi^T\phi + \lambda I_d)^{-1}\phi^T y$

$= (\phi^T I_M X + \lambda I_d)^{-1}\phi^T I_M y$ $(AI = A)$

$= (B^T R B + P^{-1})^{-1} B^T R y$

$= (B^T R B + P^{-1})^{-1} B^T R^{-1} y$ $(I^{-1} = I)$

$= PB^T(BPB^T + R)^{-1} y$ $((P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T(BPB^T + R)^{-1})$

$= (\lambda I_d)^{-1}\phi^T(\phi(\lambda I_d)^{-1}\phi^T + I_M)^{-1} y$

$= \frac{1}{\lambda} I_d \phi^T \left(X \frac{1}{\lambda} I_d \phi^T + I_M\right)^{-1} y$ $((\lambda I_d)^{-1} = (\lambda)^{-1}(I_d)^{-1} = \frac{1}{\lambda} I_d)$

$= \frac{1}{\lambda}\phi^T \left(X \frac{1}{\lambda}\phi^T + I_M\right)^{-1} y$ $(IA = A)$

$= \frac{\frac{1}{\lambda}\phi^T}{\phi\frac{1}{\lambda}\phi^T + I_M} y$

$= \frac{\phi^T}{\phi\phi^T + \lambda I_M} y$

$= \phi^T(\phi\phi^T + \lambda I_M)^{-1} y$

So $\widehat{\beta_R}$ could be written as $\phi(X)^T(\phi(X)\phi(X)^T + \lambda I_M)^{-1} y$

Given x_{new},

y_{new}

$= \widehat{\beta_R}\ \phi(x_{new})$

$= \phi(X)^T(\phi(X)\phi(X)^T + \lambda I_M)^{-1} y \phi(x_{new})$

$= \phi(X)^T\phi(x_{new})(\phi(X)\phi(X)^T + \lambda I_M)^{-1} y$

Define the Gram matrix K to be the $n\,X\,n$ matrix with

$K_{ij} := \phi(x_i)\phi(x_j)^T$

$:= k(x_i, x_j)$

y_{new} will then equal to $\phi(X)^T\phi(x_{new})(K+\lambda I_M)^{-1} y$

Since $\phi(X)^T\phi(x_{new})$ produces $(n\,X\,1)$ vector with the $i-th$ element equal to $k(x_i, x^{new})$, we just need the kernel function but not the feature vector to predict y_{new}.