Assignment 4

Seongming Lee (01247436), Yuxuan Luo (01376247) and Nina Hauser (01418616)

Question 2

(a) There are 16 rows with NA values in total. After remove these rows and 'ID' column, the dimension of this dataset is (683,10).

```
## The dimension of BreastCancer dataset is: 699 11
```

```
## The number of rows having 1 or more NA values: 16
```

```
## After removing NA values, the dimension of BreastCancer dataset is: 683 10
```

(b) Linear kernel: the cost parameter chosen by cross-validation is 0.1, with an accuracy rate of 97.08%, sensitivity score of 98.21% and specificity score of 94.92%.

```
## The confusion matrix is:
```

```
##            Reference
## Prediction  benign malignant
##   benign       110         3
##   malignant      2        56
```

```
## The accuracy rate is: 0.9707602
```

```
## The sensitivity score is: 0.9821429
```

```
## The specificity score is: 0.9491525
```

(b) Polynomial kernel of degree 2: the cost and gamma parameters chosen by cross-validation are 0.1 and 0.5 respectively, with an accuracy rate of 97.08%, sensitivity score of 98.21% and specificity score of 94.92%.

```
## The confusion matrix is:
```

```
##            Reference
## Prediction  benign malignant
##   benign       110         3
##   malignant      2        56
```

```
## The accuracy rate is: 0.9707602
```

```
## The sensitivity score is: 0.9821429
```

```
## The specificity score is: 0.9491525
```

(c) Polynomial kernel of degree 3: the cost and gamma parameters chosen by cross-validation are 0.0001 and 3 respectively, with an accuracy rate of 96.49%, sensitivity score of 97.35% and specificity score of 94.83%.

```
## The confusion matrix is:
```

```
##            Reference
## Prediction  benign malignant
##   benign       110         3
##   malignant      3        55
```

```
## The accuracy rate is: 0.9649123
```

```
## The sensitivity score is: 0.9734513
```

```
## The specificity score is: 0.9482759
```

(c) Gaussian kernel: the cost and gamma parameters chosen by cross-validation are 10 and 0.01 respectively, with an accuracy rate of 96.49%, sensitivity score of 97.35% and specificity score of 94.83%.

```
## The confusion matrix is:

##             Reference
## Prediction  benign malignant
##   benign       110         3
##   malignant      3        55

## The accuracy rate is: 0.9649123

## The sensitivity score is: 0.9734513

## The specificity score is: 0.9482759
```

Question 3

Suppose we have a set of data points $X = [d_1, d_2, d_3, ..., d_n]$ with c number of clusters. The K-means algorithm should be, first, random initialize c cluster centers; second, calculate the distance of each data point to its cluster center. Before the calculation, we perform a mapping from the input space X to a high dimensional feature space. The distance calculation can be written as

$$D[(\pi_c)_{c=1}^{k})] = \sum_{c=1}^{k} \sum_{d_i \in \pi_c} ||\phi(d_i) - mean_c||^2,$$

where $mean_c = \frac{\sum_{d_i \in \pi_c} \phi(d_i)}{|\pi_c|}$, which equals to

$$\phi(d_i)\phi(d_i) - \frac{2\sum_{d_j \in \pi_c} \phi(d_i)\phi(d_j)}{|\pi_c|} + \frac{\sum_{d_j, d_l \in \pi_c} \phi(d_j)\phi(d_l)}{|\pi_c|^2}.$$

As we know that every algorithm in which input vectors appear only in dot products with other input vectors can be kernelized, along with formula $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, the distance formula can be re-written as

$$K(d_i, d_i) - \frac{2\sum_{d_j \in \pi_c} K(d_i, d_j)}{|\pi_c|} + \frac{\sum_{d_j, d_l \in \pi_c} K(d_j, d_l)}{|\pi_c|^2}.$$

Question 4

The equation of the kernelized ridge regression can be re-written as follows:

$$minimise(w) : \frac{1}{2}||y - xw||_2^2 + \frac{\lambda}{2} w^T w$$

And the optimal solution(w) mapping to the higher dimension through $\phi(x)$ is

$$w = (\phi^T \phi + \lambda I)^{-1} \phi^T y.$$

Using the hint:

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

after setting $P = \frac{1}{\lambda} I, R = I, B = \phi$, we can firstly plug it in the LHS, which gives

$$(\phi^T \phi + \lambda I)^{-1} \phi^T.$$

The result is the same as w excluding y. And then, we can convert it to $\phi^T(\phi\phi^T + \lambda I)^{-1}$ through the hint formula. Now, we can re-formulate w as $w = \phi^T(\phi\phi^T + \lambda I)^{-1}y$. The decision function can be re-written as

$$f(x) = w^T\phi(x) = y(\phi^T\phi + \lambda I_n)^{-1}\phi^T\phi(x),$$

which includes the kernel function $K(x_1, x_2) = \phi^T(x_1)\phi(x_2)$.