Question 1

(a) For each of the polynomial bases of dimension $k = 1, 2, 3, 4$, the data set $\{(1,3), (2,2), (3,0), (4,5)\}$ is fit by linear regression using least squares, with their plots superimposed is shown in Fig. 1:
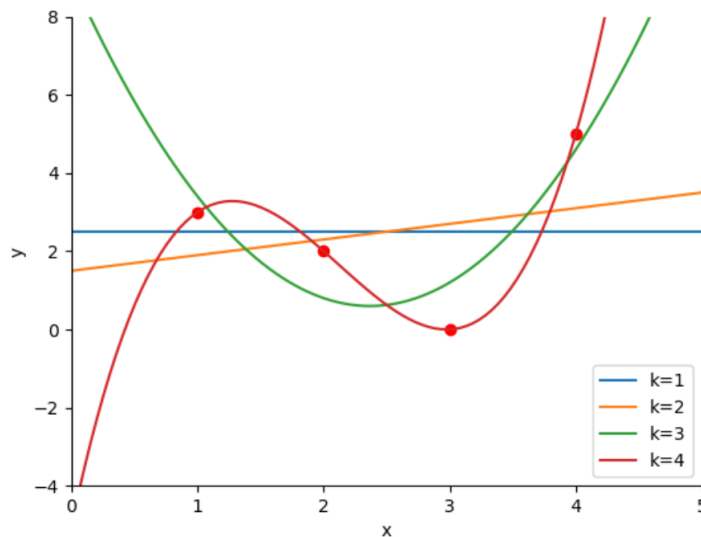

Fig. 1

(b) Using the weights computed for part (a), equations corresponding to the curves fitted for:

k=1 is $y = 2.5$,

k=2 is $y = 1.5 + 0.4x$,

k=3 is $y = 9.0 - 7.1x + 1.5x^2$,

k=4 is $y = -5.0 + 15.2x - 8.5x^2 + 1.3x^3$.

(Coefficients shown to the nearest 1 decimal place.)

(c) For each fitted curve, the mean square error (to the nearest 2 decimal places) for:

k=1 is 3.25,

k=2 is 3.05,

k=3 is 0.8,

k=4 is 0.

Question 2

(a) (i) Plot of function $\sin^2(2\pi x)$ in range $0 \le x \le 1$ with the points of given data set, superimposed (Fig. 2):

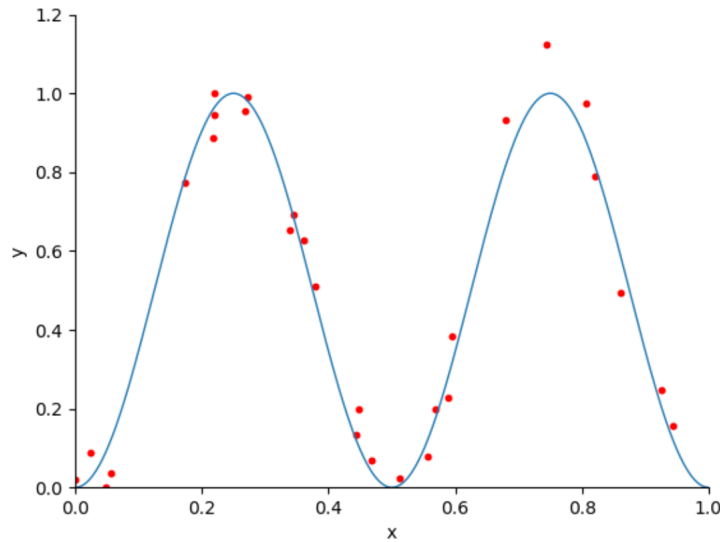
Fig. 2

(a) (ii) Fitting data set with polynomial bases of dimension $k = 2, 5, 10, 14, 18$, the plot of each of these 5 curves superimposed over a plot of data points is shown in Fig. 3:
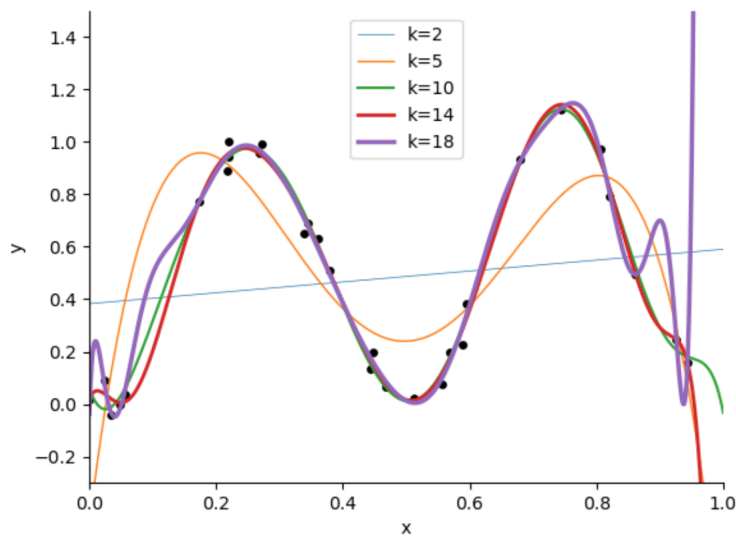

Fig. 3

Question 2 contin.

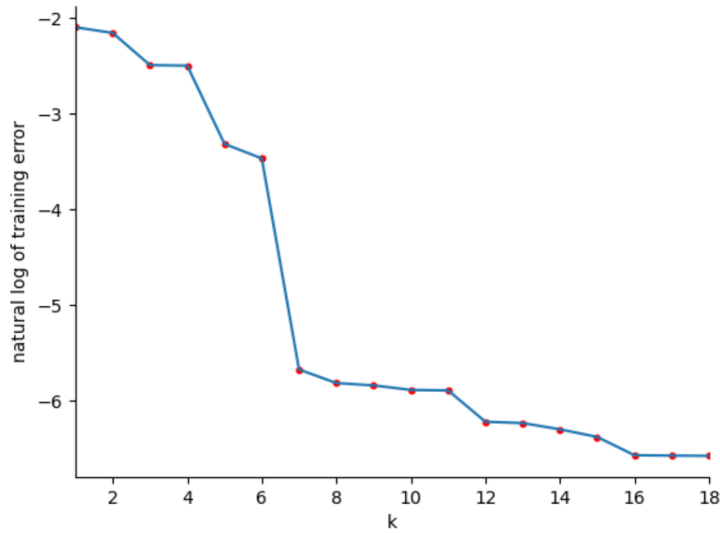(b) Plot of natural log of training error versus polynomial dimension $k = 1, ..., 18$ is shown in Fig. 4:

Fig. 4

(c) Plot of natural log of test error versus the polynomial dimension $k = 1, ..., 18$, displaying overfitting at higher dimensions, is shown in Fig. 5:
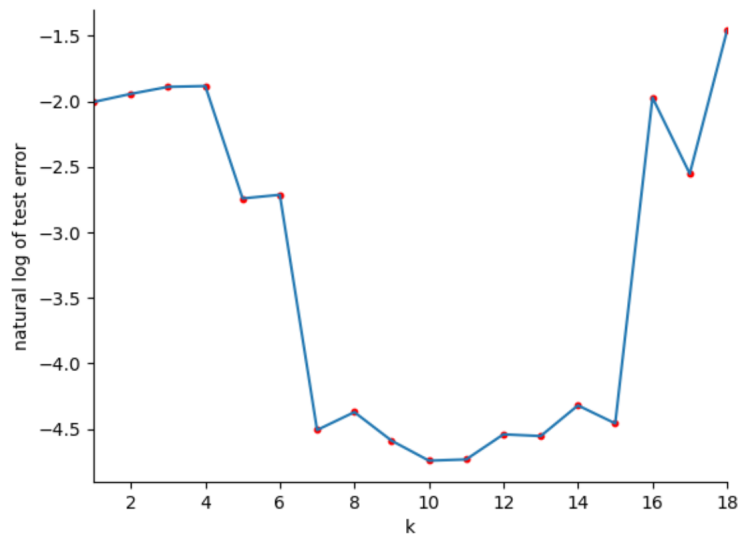

Fig. 5 (Note the y-axis range differs from Fig. 4)

Question 2 contin.

(d) Plot of average results of a 100 runs for training dataset is shown in Fig. 6:
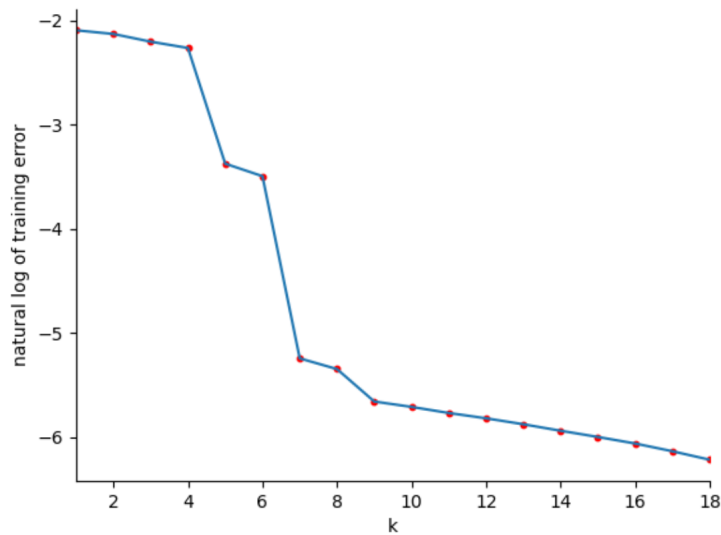

Fig. 6

Plot of average results of a 100 runs for test dataset is shown in Fig. 7:
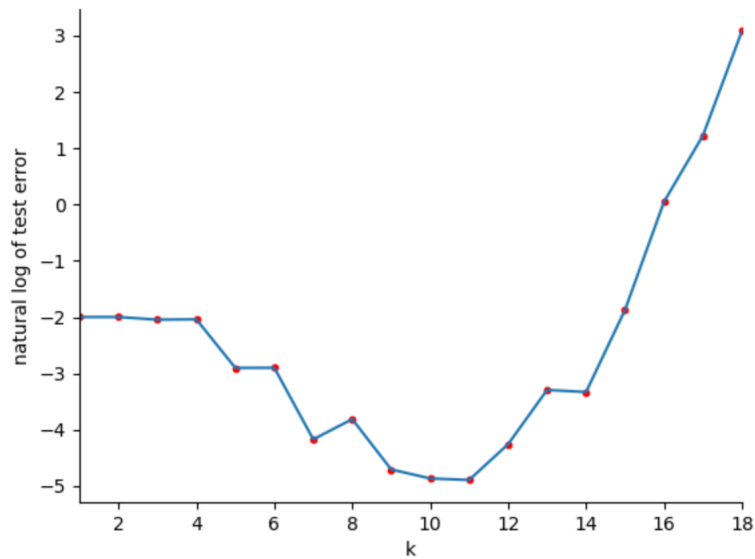

Fig. 7

Question 3

Repeating 2(b) and (c) with basis $\{\sin(\pi x), \sin(2\pi x), \sin(3\pi x), ..., \sin(k\pi x)\}$
(for $k = 1, ..., 18$). Superimposed plots of the training and test datasets is shown in
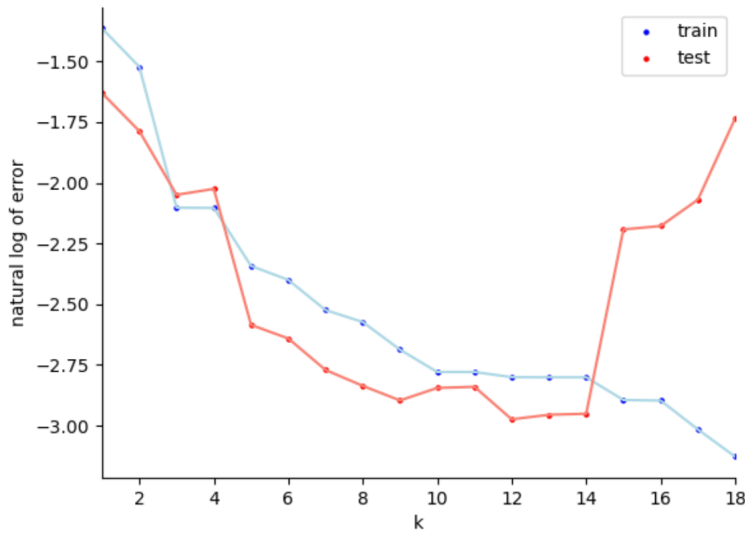Fig. 8:

Fig. 8

Repeating 2(d) with sine basis given above, averaging error over 100 runs. Super-
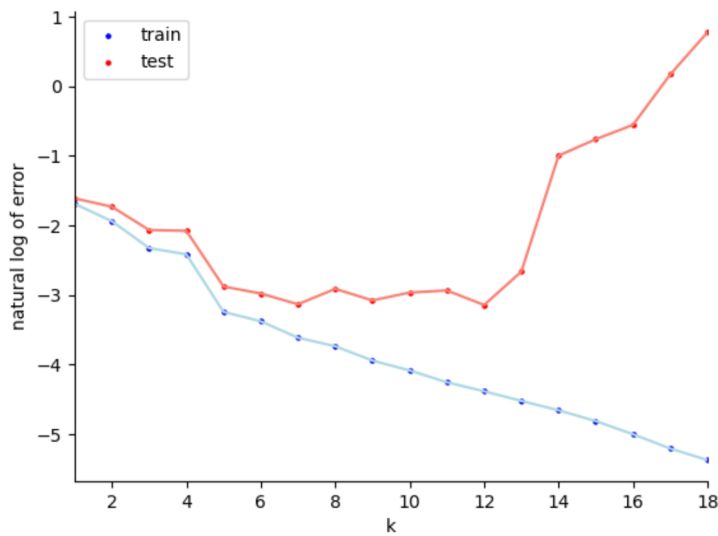imposed plots of training and test datasets is shown in Fig. 9:

Fig. 9

Question 4

(a) The mean MSE of 20 MSEs for training dataset 83.4 (to 3 s.f.).
    The mean MSE of 20 MSEs for test dataset 86.6 (to 3 s.f.).

(b) Fitting the constant function here, $y = b$, where b is the y-intercept, to the data, by least squares linear regression, gives rise to the mean of the y values. Here, this would be the mean house price.

(c) The mean MSE of 20 MSEs, for each of the attributes individually, (to 3 s.f.) is shown in Table 1:

dataset	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT
train	70.7	72.0	64.1	80.3	68.1	43.7	71.4	78.0	71.6	65.3	62.1	37.5
test	74.5	76.8	66.2	85.6	71.1	43.8	74.9	81.8	73.5	67.2	64.2	40.9

Table 1.

(d) The mean MSE of 20 MSEs, using all 12 attributes, (to 3 s.f.) for:
    training dataset is: 25.3,
    test dataset is: 21.7.
    This indicates a significant improvement over the best single predictor which appears to be LSTAT, with an MSE of 37.5 on the training dataset and 40.9 on the test dataset. (The focus in particular should be on comparing the test dataset MSEs).

Question 5

(a) Using 5-fold cross-validation and 2/3 of the full dataset, the γ and σ pair of values that gave the lowest MSE for predictions with kernel ridge regression with a Gaussian kernel, was $\gamma = 2^{-30}$ and $\sigma = 2^{9.5}$. (Although it changed from one 5-fold CV run to the next, presumably because sklearn's KFold performs a random shuffle before dividing a dataset into 5 parts).

(b) Plot of the cross-validation error (natural log of the mean MSE over 5 folds) as a function of γ and σ is shown in Fig. 10. The darker the colour, the lower the MSE. The optimal pair of gamma and sigma (as given in part (a)) is highlighted with grey in the heatmap:
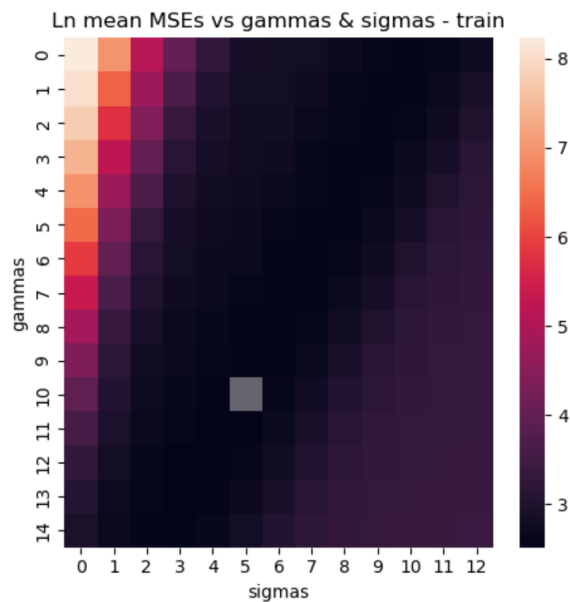

Fig. 10

(c) MSE on training and test sets for the best γ and σ (from part (a)), is 8.38 for the training dataset, and 11.5 for the test dataset (to 3 s.f.).

Question 5d

(d) MSE means and sample standard deviations for naive regression, linear regression with single attributes, all 12 attributes and KRR are shown (to 2 d.p) in Table 2.

| Method	MSE train	MSE test
Naive Regression	83.44±5.25	86.63±10.59
LR (CRIM)	70.7±5.43	74.49±10.81
LR (ZN)	72.04±4.93	76.79±9.9
LR (INDUS)	64.05±5.3	66.16±10.65
LR (CHAS)	80.35±5.07	85.57±10.31
LR (NOX)	68.12±5.25	71.11±10.45
LR (RM)	43.68±2.52	43.81±5.13
LR (AGE)	71.37±5.52	74.94±11.01
LR (DIS)	78.04±5.53	81.8±10.97
LR (RAD)	71.56±5.64	73.53±11.29
LR (TAX)	65.32±5.62	67.24±11.25
LR (PTRATIO)	62.09±4.19	64.18±8.6
LR (LSTAT)	37.48±2.5	40.88±5.21
LR (All 12)	21.7±1.35	25.27±2.85
KRR	6.16±1.38	13.2±2.16

Table 2.

Question 6
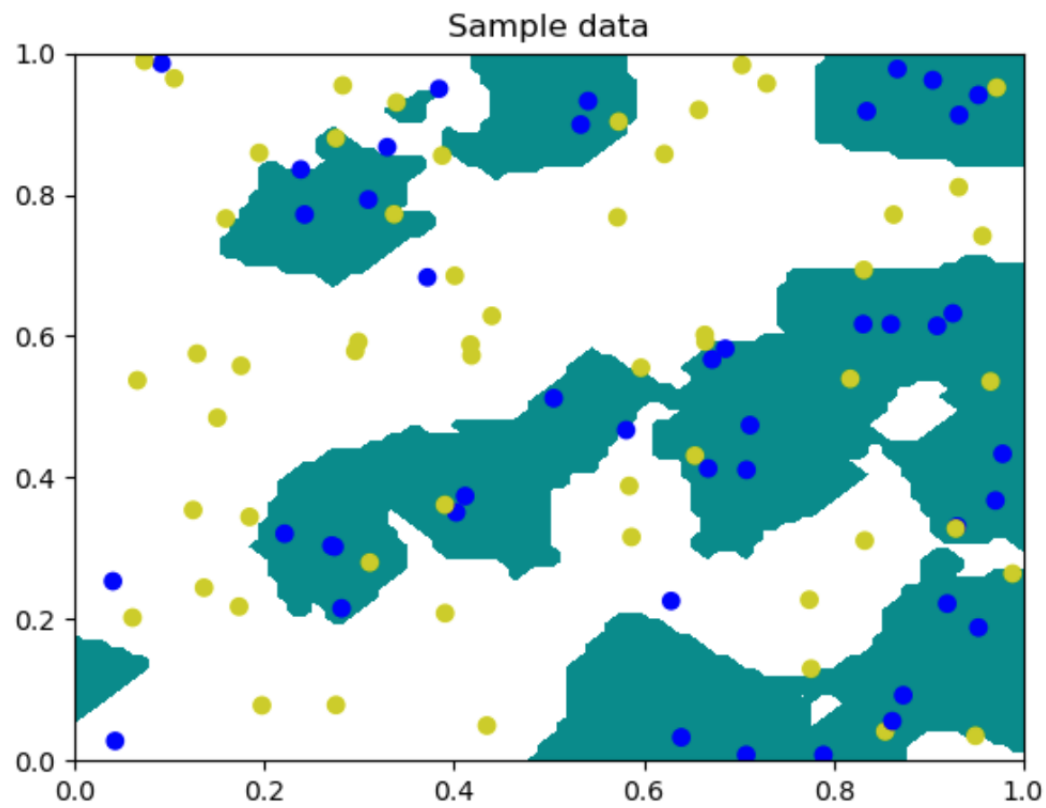
Fig. 11

(a)

(b)

Question 7

(a)

(b)

Question 8

(a)

(b)