Statistical Methods: Assessed Coursework 1

By Henry Bourne

Question 0

0.1

Our prior knowledge says that the temperature changes periodically with time, we would like to to add this behavior to our model. The trigonometric transform is good for approximating a generating function which has a temporal input, the resulting transform also consists of sinusoidal functions which would allow us to model the periodic behavior of temperature with time. Hence for $\phi^{(1)}$ we choose the trigonometric transform.

From our prior knowledge we know that latitude changes linearly with temperature, to be able to model this behavior and incorporate it into our predictive function we should therefore choose a linear transform for $\phi^{(3)}$.

Finally, we have no prior knowledge on how CO2 emissions should affect temperature, this means we would like to select a basis function which isn't restrictive as it could lead to a highly inaccurate model being fit. We would therefore like to choose a very flexible basis function, so a good choice of basis function for $\phi^{(4)}$ is the RBF.

0.2

B

Question 1

1.1

We have,

$$p(\boldsymbol{f}|\boldsymbol{K},\sigma) = N_{\boldsymbol{f}}(\boldsymbol{0},\boldsymbol{K}), \tag{1}$$

$$p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{K},\sigma) = N_{\boldsymbol{y}}(\boldsymbol{f},\sigma^2\boldsymbol{I}) \tag{2}$$

By 2.115 in PRML we have,

$$p(\boldsymbol{y}|\sigma,\boldsymbol{K}) = N_{\boldsymbol{y}}(\boldsymbol{0},\sigma^2\boldsymbol{I}+\boldsymbol{K}) \tag{3}$$

1.2

Let us partition $\boldsymbol{y}$ as follows,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \boldsymbol{y}' \end{bmatrix} \tag{4}$$

Similarly we will partition,

$$\boldsymbol{0} = \begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix} \quad \text{and} \quad \sigma^2\boldsymbol{I}+\boldsymbol{K} = \begin{bmatrix} \sigma^2\boldsymbol{K}_{11} & \boldsymbol{K}_{12} \\ \boldsymbol{K}_{21} & \sigma^2\boldsymbol{I}+\boldsymbol{K}_{22} \end{bmatrix} \tag{5}$$

Where we have partitioned $\boldsymbol{K}$ such that $\boldsymbol{K}_{11} = \boldsymbol{K}^{(1,1)}$, $\boldsymbol{K}_{21} = \boldsymbol{K}^{(2:,1)}$, $\boldsymbol{K}_{12} = \boldsymbol{K}^{(1,2:)}$ and $\boldsymbol{K}_{2,2} = \boldsymbol{K}^{(2:,2:)}$. By 2.81/2.82 in PRML we have that $p(y_1|\boldsymbol{y}',\boldsymbol{K},\sigma)$ has the following probability density function:

$$p(y_1|\boldsymbol{y}',\boldsymbol{K},\sigma) = N_{y_1}(0+\boldsymbol{K}_{12}(\sigma^2\boldsymbol{I}+\boldsymbol{K}_{22})^{-1}(\boldsymbol{y}'-\boldsymbol{0}), \sigma^2\boldsymbol{K}_{11}-\boldsymbol{K}_{12}(\sigma^2\boldsymbol{I}+\boldsymbol{K}_{22})^{-1}\boldsymbol{K}_{22}) \tag{6}$$

$$= N_{y_1}(\boldsymbol{K}_{12}(\sigma^2\boldsymbol{I}+\boldsymbol{K}_{22})^{-1}\boldsymbol{y}', \sigma^2\boldsymbol{K}_{11}-\boldsymbol{K}_{12}(\sigma^2\boldsymbol{I}+\boldsymbol{K}_{22})^{-1}\boldsymbol{K}_{22}) \tag{7}$$

1.3

Our prediction for $\boldsymbol{x}_1$ is,

$$f(\boldsymbol{x}_1; \boldsymbol{w}_{LS}) := \boldsymbol{K}_{21}(\boldsymbol{K}_{22} + \lambda \boldsymbol{I})\boldsymbol{y}'^T \tag{8}$$
$$= \boldsymbol{K}_{12}(\boldsymbol{K}_{22} + \lambda \boldsymbol{I})\boldsymbol{y}' \tag{9}$$

As $\boldsymbol{K}$ is a covariance matrix. Notice that this is exactly the mean of the distribution for $y_1|\boldsymbol{y}', \sigma, \boldsymbol{K}$, if we set the regularization parameter $\lambda = \sigma^2$, hence performing the kernel regression with λ equal to the variance we have that our prediction for a given $\boldsymbol{x}_i$ will be equal to the expected value of the target variable y_i given the rest of our target variables, $\boldsymbol{K}$ and σ.

1.4

We don't assume the data points are independent. If we don't need to assume that the datapoints are independent then it opens us up to be able to work on other types of data such as time series data.

Question 2

2.1

Let $\boldsymbol{h} := \phi(\xi)^T(\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T)^{-1}\phi(\boldsymbol{X})$, then we have,

$$\mathrm{Var}(f(\boldsymbol{x}_i; \boldsymbol{w}_{LS})|\boldsymbol{x}_i) = \mathbb{E}(f(\boldsymbol{x}_i; \boldsymbol{w}_{LS})^2|\boldsymbol{x}_i) - \mathbb{E}(f(\boldsymbol{x}_i; \boldsymbol{w}_{LS})|\boldsymbol{x}_i)^2 \tag{10}$$
$$= \mathbb{E}(\langle \boldsymbol{h}, \boldsymbol{y}\rangle^2|\boldsymbol{x}_i) - \mathbb{E}(\langle \boldsymbol{h}, \boldsymbol{y}\rangle|\boldsymbol{x}_i)^2 \tag{11}$$
$$= \mathbb{E}(\langle \boldsymbol{h}, g(\boldsymbol{x}) + \boldsymbol{\epsilon}\rangle^2|\boldsymbol{x}_i) - \mathbb{E}(\langle \boldsymbol{h}, g(\boldsymbol{x}) + \boldsymbol{\epsilon}\rangle|\boldsymbol{x}_i)^2 \tag{12}$$
$$= \langle \boldsymbol{h}, g(\boldsymbol{x})\rangle^2 + \mathbb{E}(\langle \boldsymbol{h}, \boldsymbol{\epsilon}\rangle^2) - \mathbb{E}(\langle \boldsymbol{h}, g(\boldsymbol{x})\rangle)^2 \tag{13}$$
$$= \mathbb{E}(\langle \boldsymbol{h}, \epsilon\rangle^2) \tag{14}$$
$$= \mathbb{E}(\boldsymbol{h}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{h}) \tag{15}$$
$$= \boldsymbol{h}\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\boldsymbol{h} \tag{16}$$
$$= \boldsymbol{h}^T\boldsymbol{h}\dot{\sigma}^2 \tag{17}$$
$$= \langle \boldsymbol{h}, \boldsymbol{h}\rangle \cdot \sigma^2 \tag{18}$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(f(\boldsymbol{x}_i; \boldsymbol{w}_{LS})|\boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{h}, \boldsymbol{h}\rangle \cdot \sigma^2 \tag{19}$$

Note that,

$$\langle \boldsymbol{h}, \boldsymbol{h}\rangle = \mathrm{tr}(\boldsymbol{h}\boldsymbol{h}^T) \tag{20}$$
$$= \mathrm{tr}(\phi(\boldsymbol{x}_i)^T(\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T)^{-T}\phi(\boldsymbol{x}_i)) \tag{21}$$
$$= \mathrm{tr}(\phi(\boldsymbol{x}_i)^T(\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T)^{-1}\phi(\boldsymbol{x}_i)) \tag{22}$$

Hence,

$$\frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{h}, \boldsymbol{h}\rangle \cdot \sigma^2 = \mathrm{tr}(\sum_{i=1}^{n} \phi(\boldsymbol{x}_i)^T(\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T)^{-1}\phi(\boldsymbol{x}_i)) \cdot \frac{\sigma^2}{n} \tag{23}$$
$$= \mathrm{tr}((\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T)(\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T)^{-1}) \cdot \frac{\sigma^2}{n} \tag{24}$$
$$= \frac{b\sigma^2}{n} \tag{25}$$

Hence, as b increases $\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(f(\boldsymbol{x}_i; \boldsymbol{w}_{LS})|\boldsymbol{x}_i)$ grows.

2.2

The in-sample error is,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}((y_i - f_{LS}(\boldsymbol{x}_i))^2|\boldsymbol{x}_i) \tag{26}$$

Using bias-variance decomposition we know this equals,

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\epsilon_i) + [g(\boldsymbol{x}_i) - \mathbb{E}(f_{LS}(\boldsymbol{x}_i)|\boldsymbol{x}_i)]^2 + \mathrm{Var}(f_{LS}(\boldsymbol{x}_i)|\boldsymbol{x}_i) \tag{27}$$

Looking at the bias we have that its equal to,

$$g(\boldsymbol{x}_i)^2 - 2g(\boldsymbol{x}_i)\cdot\mathbb{E}(f_{LS}(\boldsymbol{x}_i)|\boldsymbol{x}_i) + \mathbb{E}(f_{LS}(\boldsymbol{x}_i)|\boldsymbol{x}_i)^2 \tag{28}$$

$$= g(\boldsymbol{x}_i)^2 - 2g(\boldsymbol{x}_i)\cdot\mathbb{E}(\phi(\boldsymbol{x}_i)^T(\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T)^{-1}\phi(\boldsymbol{X})(g(\boldsymbol{x}) + \epsilon_i)|\boldsymbol{x}_i) \tag{29}$$

$$+ \mathbb{E}(\phi(\boldsymbol{x}_i)^T(\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T)^{-1}\phi(\boldsymbol{X})(g(\boldsymbol{x}) + \epsilon_i)|\boldsymbol{x}_i)^2 \tag{30}$$

$$= g(\boldsymbol{x}_i)^2 - 2g(\boldsymbol{x}_i)\cdot\mathbb{E}(\phi(\boldsymbol{x}_i)^T\boldsymbol{w}^*) + \mathbb{E}(\phi(\boldsymbol{x}_i)^T\boldsymbol{w}^*)^2 \tag{31}$$

$$= g(\boldsymbol{x}_i)^2 - 2g(\boldsymbol{x}_i)\cdot\mathbb{E}(g(\boldsymbol{x}_i)) + \mathbb{E}(g(\boldsymbol{x}_i))^2 \tag{32}$$

$$= 0 \tag{33}$$

We also have that $\mathrm{Var}(\epsilon_i)$ is constant as n increases, finally we have from the previous question that,

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(f(\boldsymbol{x}_i;\boldsymbol{w}_{LS})|\boldsymbol{x}_i) = \frac{b\sigma^2}{n} \tag{34}$$

Hence as n increases, this above decreases and the in-sample error decreases

2.3

Because the least squares estimator is unbiased if we reduce the variance associated with our prediction we obtain an estimate closer to the true data generating function. Therefore, if we increase n (by question 2.1/2.2) we have that the average variance of the prediction function given the dataset decreases, this means that we are less likely to overfit as our prediction is likely to stray from the expected value of the prediction function which is the data generating function. Conversely if we decrease n then we are more likely to overfit the data as the variance of our prediction given our dataset increases.

If we increase b, then the variance increases and using the same argument as before we are more likely to overfit, conversely if we decrease b the variance is more likely to decrease and we are less likely to overfit.