Task 1 - Kernel Ridge Regression (KRR) Example

Consider a data set with two data points, $x_1 = -1, x_2 = 1$ with respective labels $y_1 = 1, y_2 = 1$.

1. We want to fit a simple linear model $f(x) = \omega \cdot x$ to the data using Ordinary Least Squares (OLS). Recall the OLS solution is obtained as

$$\omega = \underset{\omega}{\operatorname{argmin}} \sum_{n=1}^{N} (y_n - f(x_n))^2 = (XX^\top)^{-1} X y^\top$$

   where $N = 2$ is the number of data points, $X = [x_1, x_2]$ and $y = [y_1, y_2]$. Compute ω.

2. We obtain a better fit using the model $g(x) = w_1 + w_2 \cdot x = \mathbf{w}^T \cdot \phi(x)$ where we have defined a mapping $\phi : \mathbb{R} \ni x \mapsto \begin{bmatrix} 1 \\ x \end{bmatrix} \in \mathbb{R}^2$ and a weight vector $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \mathbb{R}^2$. Recall the OLS solution is obtained as

$$\mathbf{w} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{n=1}^{N} (y_n - g(x_n))^2 = (XX^\top)^{-1} X y^\top$$

   where N and y are defined as above and $X = [\phi(x_1), \phi(x_2)] = \begin{bmatrix} 1 & 1 \\ x_1 & x_2 \end{bmatrix}$. Compute $\mathbf{w}$ and the corresponding function $g(x)$.

3. Now, we want to obtain the same solution as above, but by solving the dual representation. Instead of learning $\mathbf{w}$ directly, we learn a linear combination $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \in \mathbb{R}^N$ of the data points, i.e. $\mathbf{w} = \alpha_1 \phi(x_1) + \alpha_2 \phi(x_2)$. In the lecture, we derived the following formula:

$$\alpha = K^{-1} y^\top \tag{1}$$

   where $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the kernel function and $K \in \mathbb{R}^{N \times N}$ is the kernel matrix, with $K_{ij} = k(x_i, x_j)$. Here, the Kernel function is given as $k(x_i, x_j) = \phi(x_i)^\top \phi(x_j) = \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} 1 \\ x_j \end{bmatrix} = 1 + x_i x_j$.

   Using only α and the Kernel function, the predictions for a new data point x are given as

$$g_2(x) = \alpha_1 k(x, x_1) + \alpha_2 k(x, x_2)$$

   Compute α and show that we obtain the same solution as before, i.e. show that $g_2(x) = g(x)$.

4. Now we want to use the Gaussian Kernel defined as

$$k(x_i, x_j) = e^{-\frac{(x_i - x_j)^2}{2\sigma^2}}$$

   with kernel width $\sigma = 0.5$. Compute the Kernel matrix K and the Kernel Ridge Regression coefficients α using Equation (1) . To simplify calculations, approximate $e^{-8} = 0.000335 \approx 0$.

   Sketch the obtained function $h(x) = \alpha_1 k(x, x_1) + \alpha_2 k(x, x_2)$ as a 2D plot (use e.g. $e^{-2} = 0.14$).

5. Suppose we use a linear Kernel function, $k(x_i, x_j) = x_i x_j$. What problem do we get when computing the Kernel Ridge Regression coefficients α using Equation (1)?

Task 2 - Cross-validation

1. You are a reviewer for the International Mega-Conference on Machine Learning of Outrageous Stuff, and you read a paper that selected a small number of features out of a large number of features for a given classification problem. The paper argues as follows:

   (a) We used all our available data to select a subset of "good" features that had fairly strong correlation with the class labels.

   (b) Our final model contained only those features. We evaluate the prediction error of the final model by 10-fold crossvaldiation on all the available data.

   (c) We obtained a low cross-validation error. Thus, we have achieved high classification accurarcy with only few meaningful features. (This is novel and amazing.)

   Would you accept or reject the paper? Why?

2. Suppose you are testing a new algorithm on a data set consisting of 100 positive and 100 negative examples. You plan to use leave-one-out cross-validation (i.e. 200-fold cross-validation) and compare your algorithm to a baseline function, a simple majority classifier. Given a set of training data, the majority classifier always outputs the class that is in the majority in the training set, regardless of the input. You expect the majority classifier to achieve about 50% classification accuracy, but to your surprise, it scores zero every time. Why?