Exercise about RSM lecture

Abstract

The aim of this exercise is to program your own basic implementation of Radial Basis Functions (RBF) Response Surface Model (RSM) in your favourite programming language (e.g. MATLAB, python, R, ...), to train the RSM on a given training dataset and to assess its predictive accuracy on a given validation dataset.

1 Radial Basis Functions basics

RBF theory is summarised below for your convenience.

Given a training dataset of n points (in a d-dimensional space)

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, n\}\,, \tag{1}$$

the purpose of any RSM is to find a model $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ that fits well the training data (and exhibits a good predictive power).

An RBF interpolant model has the form

$$f(\mathbf{x}) = \sum_{j=1}^{n} c_j \phi\left(\|\mathbf{x} - \mathbf{x_j}\| / \delta\right), \tag{2}$$

where $\|.\|$ is the Euclidean norm in the d-dimensional space and δ a fixed scaling parameter. Let us consider a Gaussian radial function $\phi(r) = \exp(-r^2)$. The RBF interpolant model $f(\mathbf{x})$ is a linear combination of identical spherical symmetric functions, centered at the n different training points sites.

The values of the coefficients c_j are obtained by imposing the interpolation equations

$$f(\mathbf{x}_i) = y_i\,, \quad \forall i = 1, \ldots, n\,. \tag{3}$$

By defining the symmetric collocation matrix $\mathbf{A}$ as

$$A_{ij} = \phi\left(\|\mathbf{x_i} - \mathbf{x_j}\| / \delta\right), \quad i, j = 1, \ldots, n\,, \tag{4}$$

the interpolation equations can be expressed as

$$f(\mathbf{x}_i) = \sum_{j=1}^{n} A_{ij}\, c_j = y_i\,, \quad \forall i = 1, \ldots, n\,, \tag{5}$$

or, in matrix form, as

$$\mathbf{A} \cdot \mathbf{c} = \mathbf{y}\,. \tag{6}$$

The unknown coefficient vector is obtained by solving the linear system of Eq. (6).

2 Problem description

A training dataset is provided in the CSV file *training.csv*. The dataset consists in a set of 50 points. The input variables are x_1 and x_2, while y is the output variable to be modelled.

A validation datased of 150 points is also provided in the CSV file *validation.csv*.

1. Program your own basic implementation of RBF (with a Gaussian basis) in your favourite programming language
2. Train it on the training dataset
3. Assess its predictive accuracy on the validation dataset

The predictive accuracy should be expressed as the Root-Mean-Square Deviation (RMSD):

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2}{n}}\,. \tag{7}$$

The smaller the RMSD value, the better the predictive accuracy.

- Is it possible to achieve a value of RMSD better than 0.05?