DSSV 2023 Data Science, Statistics & Visualisation
EC DA 2023 European Conference on Data Analysis

Interpretable Kernels

Patrick Groenen[1] and Michael Greenacre[2]

[1]Erasmus University Rotterdam, Econometric Institute, P.O.Box 1738, 3000 DR, Rotterdam
[2]Universitat Pompeu Fabra, Barcelona, Spain

The use of kernels for nonlinear prediction is widespread in machine learning. They have been popularized in support vector machines and used in kernel ridge regression, amongst others. These methods share three aspects. First, instead of the original $n \times p$ matrix of predictor variables, each row is mapped into a high dimensional feature space. Second, a ridge penalty term is used to shrink the weights (coefficients) on the predictors in the high-dimensional feature space. Third, the solution is not obtained in this feature space, but through solving a dual problem. A major drawback in the use of kernels is that the interpretation in terms of the original predictor variables is lost.

In this paper, we argue that in the case of a wide $n \times p$ matrix of predictor variables (with $p > n$), the kernel solution can be re-expressed in terms of a linear combination of the original matrix of predictor variables and a ridge penalty that involves a special metric. Consequently, the exact same predicted values can be obtained as a weighted linear combination of the predictor variables in the usual manner and thus can be interpreted. In the case $p \leq n$, we discuss a least-squares approximation of the kernel matrix that still allows the interpretation in terms of a linear combination. It is shown that these results hold for any function of a linear combination that minimizes the weights and has a ridge penalty of these weights such as in kernel logistic regression and kernel Poisson regression. When the objective function is minus the log likelihood, standard likelihood theory can be used to estimate the standard deviations of the weights.

As an extension, it is possible to apply an approximation in a k-dimensional space with $k < p$ thereby enforcing dimension reduction in the predictor space.