Documentation for klearn, Part I: Derivations

C. R. Schwantes

Here we derive the kernel versions of several methods in supervised and unsupervised learning. We are by no means claiming to have discovered these methods on our own, but this collection may be useful for those not familiar with the derivations. In addition, most of these methods have alternative derivations and interpretations, the proofs disclosed here are simply my interpretation. **These notes are not complete, so there could easily be errors/typos. Let me know if you find one!**

I. RIDGE REGRESSION

A. Goals

Suppose we are given N pairs of variables, $\{X_i \in \mathbb{R}^d\}_{i=1}^N$ and $\{y_i \in \mathbb{R}\}_{i=1}^N$. The goal of linear regression is to find a linear function that fits the observed data. We will optimize the parameters of the function to minimize the sum of the squared residuals:

$$\min_{\mathbf{p},b} \sum_{i=1}^{N} \Big((\mathbf{p}^T X_i + b) - y_i \Big)^2, \tag{1}$$

where $\mathbf{p} \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Importantly, we will regularize the solution, $\mathbf{p}$ by adding an L_2 penalty to the objective function, weighted by a real number, η. In addition, we will assume that both the dependent and independent variables have zero mean, and so the y-intercept, b is equal to zero and the full objective function becomes:

$$\min_{\mathbf{p}} \sum_{i=1}^{N} \Big(\mathbf{p}^T X_i - y_i \Big)^2 + \eta \mathbf{p}^T \mathbf{p}. \tag{2}$$

Let X denote a $d \times N$ matrix with the independent variables in the columns and their features in the rows and $\mathbf{y}$ denote a column vector with the values of the dependent variable in its rows. Then it can be shown that the solution to Eq. (2) is:

$$\mathbf{p} = \left(X X^T + \eta I \right)^{-1} X \mathbf{y} \tag{3}$$

B. Derivation of Kernel Ridge Regression

Consider an unspecified (possibly non-linear) mapping function, $\Phi : \mathbb{R}^d \to V$ that transforms our vectors into a new space, V termed the feature space. We wish to perform the same regression as above, but in the feature space. Now, the optimization problem can be written as:

$$\min_{\mathbf{p}} \sum_{i=1}^{N} \Big(\mathbf{p}^T \Phi(X_i) - y_i \Big)^2 + \eta \mathbf{p}^T \mathbf{p}. \tag{4}$$

We now look to reformulate Eq. (2) and its solution in terms of a gram matrix of inner products. First, note that Eq. (3) shows that $\mathbf{p}$ is in the span of the independent variables. Let, β be the vector of coefficients such that:

$$\mathbf{p} = \sum_{i=1}^{N} \beta_i \Phi(X_i)$$

Now, we need only rewrite Eq. (4) in terms of the vector β. It will be useful to define the $N \times N$ gram matrix, K such that:

$$K_{ij} = \Phi(X_i)^T \Phi(X_j)$$

Notice this matrix is invertible. Now starting with Eq. (2):

$$\begin{aligned}
&\sum_{i=1}^{N} \Big(\mathbf{p}^T \Phi(X_i) - y_i \Big)^2 + \eta \mathbf{p}^T \mathbf{p} \\
&\quad = \sum_{i=1}^{N} \left(\sum_{k=1}^{N} \beta_k \Phi(X_k)^T \Phi(X_i) - y_i \right)^2 \\
&\qquad + \eta \sum_{k=1}^{N} \beta_k \Phi(X_k)^T \sum_{l=1}^{N} \beta_l \Phi(X_l) \\
&\quad = \sum_{i=1}^{N} \left(\sum_{k=1}^{N} \beta_k K_{ki} - y_i \right)^2 \\
&\qquad + \eta \sum_{k=1}^{N} \sum_{l=1}^{N} \beta_k K_{kl} \beta_l \\
&\quad = \left(K\beta - \mathbf{y} \right)^T \left(K\beta - \mathbf{y} \right) + \eta \beta^T K \beta \\
&\quad = \beta^T K K \beta - 2\beta^T K \mathbf{y} + \mathbf{y}^T \mathbf{y} + \eta \beta^T K \beta
\end{aligned}$$

Taking derivatives with respect to the elements of β and setting this equal to zero gives:

$$\begin{aligned}
0 &= 2KK\beta - 2K\mathbf{y} + 2\eta K\beta \\
\beta &= \left(K + \eta I \right)^{-1} \mathbf{y}
\end{aligned}$$

Note, if K is singular, then there are other solutions corresponding to the nonzero vectors in the null space of K.

Interestingly, if we decide to change the regularization and instead penalize the L_2 norm of β then the optimization problem (in terms of β) becomes:

$$\min_{\beta} \beta^T K K \beta - 2\beta^T K \mathbf{y} + \mathbf{y}^T \mathbf{y} + \eta \beta^T \beta, \tag{5}$$

and the solution is given by:

$$\beta = (KK + \eta I)^{-1} K\mathbf{y} \tag{6}$$

Customarily, the L_2 penalty has been applied to the vector $\mathbf{p}$, however, this alternative scheme may be desirable in some contexts.

II. TIME-STRUCTURE BASED INDEPENDENT COMPONENTS ANALYSIS