Kernel Feature Selection via Conditional Covariance Minimization

Jianbo Chen* Mitchell Stern* Martin J. Wainwright Michael I. Jordan
University of California, Berkeley

Abstract

We propose a method for feature selection that employs kernel-based measures of independence to find a subset of covariates that is maximally predictive of the response. Building on past work in kernel dimension reduction, we show how to perform feature selection via a constrained optimization problem involving the trace of the conditional covariance operator. We prove various consistency results for this procedure, and also demonstrate that our method compares favorably with other state-of-the-art algorithms on a variety of synthetic and real data sets.

Proposed Method

Feature selection criterion:

\[ \min_{T \in \mathcal{M}} \mathcal{Q}(T) = \text{Tr}(\Sigma_{YY|X_T}). \]

Property 1. If \((H_x, K_x)\) is characteristic, then \(\text{Tr}(\Sigma_{YY|X_T}) \leq \text{Tr}(\Sigma_{YY|X_T})\) for any \(T\). Moreover, the equality \(\text{Tr}(\Sigma_{YY|X_T}) = \text{Tr}(\Sigma_{YY|X_T})\) holds if and only if \(Y \perp X|X_T\).

Property 2. The criterion characterizes prediction error:

\[ \text{Tr}(\Sigma_{YY|X_T}) = E_{X_T}(f(T)) = \inf_{f \in F_T} E_{X_T}(Y - f(T))^2, \]

where \(F_T\) is a function space from \(\mathbb{R}^m\) to \(\mathcal{Y}\) defined from \(H_x\).

Empirical estimate (with a linear kernel on \(Y\)):

\[ \min_{T \in \mathcal{M}} \mathcal{Q}(T) = \text{Tr}(Y^T(G_X + n \epsilon)^{-1}Y), \]

where

\[ G_X = (I_n - \frac{1}{n} I_d^T)K_X(I_n - \frac{1}{n} I_d), \]

and \(x^T \in \mathbb{R}^d\) is a vector with \(x^T = x_i^T\) if \(i \in T\) or 0 otherwise.

Theorem 1. [Feature Selection Consistency] Define the set of all optimal feature subsets to be \(\mathcal{A} = \arg\min_{T \in \mathcal{M}} \mathcal{Q}(T)\), and let \(\hat{T}(n) = \arg\min_{T \in \mathcal{M}} \mathcal{Q}(T)\). Then \(\hat{T}(n)\) is a global optimum of the empirical estimate. If \(\epsilon_n \to 0\) and \(\epsilon_n^2 \to \infty\) as \(n \to \infty\), we have \(P(\hat{T}(n) \in \mathcal{A}) \to 1\).

Formulating Feature Selection

The problem of feature selection:

Given \(n\) i.i.d. samples \(\{(x_i, y_j): i = 1, 2, \ldots, n\}\) generated from \(P_{X Y}\) together with an integer \(m \leq d\), select \(m\) of the \(d\) features \(S = \{X_1, X_2, \ldots, X_d\}\) which best predict \(Y\).

Dependence Perspective:

Identify a subset of features \(T\) of size \(m\) such that:

\(X_{S,T}\) is conditionally independent of \(Y\) given \(X_T\).

Prediction Perspective:

Find the subset of features that minimizes the prediction error:

\[ \min_{T \in \mathcal{M}} E_{X_T}(f(T)) = \min_{T \in \mathcal{M}} \inf_{f \in F_T} E_{X_T}(Y(f(T))). \]

where \(E_{X_T}(f(T))\) is the error of prediction using only the features in \(T\), \(F_T\) is a function class from \(X_T\) to \(Y\), and \(L\) is a loss.

Conditional Covariance Operator

\((H_x, K_x)\) and \((H_y, K_y)\): RKHSs of functions on \(X\) and \(Y\); \(X \times Y\): a random vector on \(X \times Y\) with joint distribution \(P_{X Y}\).

Cross-covariance operator: an operator \(\Sigma_{XY} : H_x \rightarrow H_y\) with

\(g \in \Sigma_{XY} \rightarrow E_x y[f(X) - Exy[f(X)]][g(Y) - Ey[g(Y)]]\).

Conditional covariance operator:

\(\Sigma_{YY|X_T} = \Sigma_{YY} - \Sigma_{XY} \Sigma_{X|X_T}^{-1} \Sigma_{XY} \).

\(\Sigma_{YY|X_T}\) captures conditional variance: for \(g \in \mathcal{H}_y\),

\(g \in \Sigma_{YY|X_T} g \rightarrow E_y [\text{Var}_{X_T}[g(Y)|X]]\).

\(\Sigma_{YY|X_T}\) captures residual error: for \(g \in \mathcal{H}_y\),

\(g \in \Sigma_{YY|X_T} g \rightarrow \inf_{f \in H_x} E_{X_T} y[g(Y) - f(X)]^2\).

Optimization

We relax the initial NP-hard formulation to obtain:

\[ \min_w y^T(G_w + n \epsilon)^{-1}y \]

subject to \(0 \leq w_i \leq 1, \ i = 1, \ldots, d, \]

\(1^T w \leq m. \)

where \(\odot\) is the Hadamard product.

We may further use a kernel approximation \(G_w \approx V_w V_w^T:\)

\(G_w + n \epsilon)^{-1} \approx \frac{1}{m}(I - V_w V_w^T + \epsilon n I_d)^{-1}V_w^T. \)

Both objectives are optimized using projected gradient descent.

Synthetic Experiments

Synthetic data sets: binary classification, 4-way classification, additive nonlinear regression.

Other algorithms: Recursive feature elimination (RFE), Minimum Redundancy Maximum Relevance (mRMR), BAHSIC, mutual information (MI) and Pearson’s correlation (PC).

Evaluation: Median rank assigned to true features.

Real-world Experiments

Summary of data sets:

<table>
<thead>
<tr>
<th>Samples</th>
<th>Features</th>
<th>Classes</th>
<th>Other nonlinear algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>171</td>
<td>0.76</td>
<td>2</td>
<td>Other nonlinear algorithms</td>
</tr>
<tr>
<td>100</td>
<td>0.24</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.24</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.24</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.24</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.24</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.24</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Other nonlinear algorithms: mRMR, BAHSIC, and MI.

Evaluation: Accuracy of a kernel SVM on selected features.

Plots of accuracy vs. number of selected features