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 framework 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 formulate our approach as a constrained optimization problem involving the trace of the conditional covariance operator, and additionally provide some consistency results. We then demonstrate on a variety of synthetic and real data sets that our method compares favorably with other state-of-the-art algorithms.

Two Perspectives of Feature Selection

The problem of feature selection:

Given n i.i.d. samples \( \{(x_i, y_i) : i = 1, 2, \ldots, n\} \) generated from \( P_{X \sim Y} \) 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 \( S \setminus T \) are conditional independent of \( Y \) given \( T \).

Prediction Perspective:
Find the subset of features that minimizes the prediction error:

\[
\min_{|T| \leq m} E_T [L(X_T, Y | T)] = \min_{|T| \leq m} E_T [L(X_T, Y | T, X_{\setminus T})].
\]

where \( E_T [L(X_T, Y | T)] \) is the error of prediction using only the features in \( T \), \( F \) is a function class from \( \mathcal{X} \) to \( \mathcal{Y} \), and \( L \) is a loss.

Conditional Covariance Operator

\((\mathcal{H}_Y, k_Y)\) and \((\mathcal{H}_X, k_X)\): two RKHSs of functions on \( \mathcal{X} \) and \( \mathcal{Y} \), respectively.

\((X, Y)\): a random vector on \( \mathcal{X} \times \mathcal{Y} \) with joint distribution \( P_{X \sim Y} \).

Cross-covariance operator: an operator \( \Sigma_{XY} : \mathcal{H}_X \rightarrow \mathcal{H}_Y \) such that

\[
(g, \Sigma_{XY} h)_Y = E_X [g(X) h(Y)] = E_X [g(X) E_Y [h(Y)|X]].
\]

Conditional covariance operator:

\[
\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}.
\]

Capture conditional variance: For \( g \in \mathcal{H}_Y \),

\[
(g, \Sigma_{Y|X} g)_{Y|X} = E_X [\text{Var}_X [g(Y)|X]].
\]

Capture residual error: For \( g \in \mathcal{H}_X \),

\[
(g, \Sigma_{Y|X} g)_{Y|X} = \inf_{f \in \mathcal{H}_X} E_X [g(Y) - f(X)]^2.
\]

Proposed Method

Feature Selection Criterion:

\[
\min_{T \subseteq \mathcal{X}, |T| = m} Q(T) := \text{Tr}(\Sigma_{Y|X|T}).
\]

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

Property 2. The criterion characterizes prediction error:

\[
\text{Tr}(\Sigma_{Y|X|T}) = E_X [Q_T(X_T)] = \min_{f \in \mathcal{H}_F} E_X [Y - f(X)]^2,
\]

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

The empirical estimate (With a linear kernel on \( X \)):

\[
\min_{|T| \leq m} \hat{Q}_T(Y) := \text{Tr}(Y^T (G_Y + \epsilon n I)^{-1} Y),
\]

where \(G_Y = (I - \frac{1}{n} \mathbf{1} \mathbf{1}^T) K_Y (I - \frac{1}{n} \mathbf{1} \mathbf{1}^T), K_Y = (k_{XY} x_i x_j)_{i,j=1}^n \), \( \mathbf{1} \in \mathbb{R}^n \), \( x_i \in \mathbb{R}^d \) is a vector that \( x_i^T = x_i \) if \( i \in T \) or 0 otherwise.

Optimization

Relaxation of the NP-hard problem:

\[
\min_W y^T (G_{\omega:Y} + \epsilon n I)^{-1} y
\]

subject to \( \omega \geq 0 \), \( \omega \leq 1 \), \( i = 1, \ldots, d \), \( \mathbf{1}^T \omega \leq m \).

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

Using a kernel approximation \( G_{\omega:Y} \approx V_\omega^T V_\omega \):

\[
(G_{\omega:Y} + \epsilon n I)^{-1} \approx \frac{1}{\epsilon n} (I - V_\omega^T V_\omega + \epsilon n I)^{-1} V_\omega^T
\]

Both objectives are optimized using projected gradient descent.

Synthetic Experiments

Synthetic data sets:
Binary classification on Orange Skin.
3-dimensional XOR as 4-way classification.
Additive nonlinear regression.

Comparing algorithms: Recursive feature elimination (RFE), MinimumRedundancy Maximum Relevance (mRMR), BAHSIC, mutual information (MI) and Pearsons correlation (PO).

Metric: The median ranks assigned to the true features.

Real-world Experiments

Data sets Summary:

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Features</th>
<th>Classes</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>synthetic</td>
<td>5,784</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>real-world</td>
<td>20,000</td>
<td>10,000</td>
<td>10</td>
</tr>
</tbody>
</table>

Comparing algorithms: mRMR, BAHSIC, and MI

Metric: Accuracy of a kernel SVM on selected features.

Plots of accuracy vs. number of selected features