Vanishing Component Analysis

Krikamol Muandet

Empirical Inference Department
Max Planck Institute for Intelligent Systems

July 8, 2013

Acknowledgement

Vanishing Component Analysis
Roi Livni, David Lehavi, Sagi Schein, Hila Nachlieli, Shai Shalev-Shwartz, and Amir Globerson
International Conference on Machine Learning (ICML) 2013
best paper award

Synopsis

Supervised Learning

Input Data X $\Longrightarrow$ Define Feature $\Longrightarrow$ Learn $f(X)$

Synopsis

Supervised Learning

Input Data X $\Longrightarrow$ Define Feature $\Longrightarrow$ Learn $f(X)$

How to find a "good" feature representation?

Synopsis

Supervised Learning

Input Data X $\Longrightarrow$ Define Feature $\Longrightarrow$ Learn $f(X)$

How to find a "good" feature representation?

- feature selection: select a subset of features.

Supervised Learning

Input Data X $\Longrightarrow$ Define Feature $\Longrightarrow$ Learn $f(X)$

How to find a "good" feature representation?

- feature selection: select a subset of features.
- kernel methods: a high-dimensional feature map $\phi(x)$.

Synopsis

Supervised Learning

Input Data X $\Longrightarrow$ Define Feature $\Longrightarrow$ Learn $f(X)$

How to find a "good" feature representation?

- feature selection: select a subset of features.
- kernel methods: a high-dimensional feature map $\phi(x)$.
- representation/deep learning: learn feature representation from the large-scale dataset.

Supervised Learning

Input Data X $\Longrightarrow$ Define Feature $\Longrightarrow$ Learn $f(X)$

How to find a "good" feature representation?

- feature selection: select a subset of features.
- kernel methods: a high-dimensional feature map $\phi(x)$.
- representation/deep learning: learn feature representation from the large-scale dataset.
- this paper: define the feature as a set of functions that characterizes the dataset.

$$\text{Find } f_1(x), \ldots, f_k(x) \text{ such that } f_i(x) \approx 0 \text{ for all } i \text{ and } x \in S$$

Preliminaries

Find $f_1(x), \dots, f_k(x)$ such that $f_i(x) \approx 0$ for all i and $x \in S$

Vanishing Ideal

The set of *all* polynomials $f(x)$ that attain a value of zero on a set S is known as the *vanishing ideal* of S, denoted by $I(S)$.

Set of Generators

Given an ideal I. A set of polynomials $\{f_1, \dots, f_k\} \subseteq I$ is said to generate I, if $\forall f \in I$ there exist $g_1, \dots, g_k \in \mathbb{R}[x_1, \dots, x_k]$ such that $f = \sum_i g_i f_i$ (HIlbert's basis theorem).

A Simple but Impractical Approach

- **Assumption**: there is a set of generators of $I(S_m)$ of maximal degree D.
- Construct a matrix A of size $(m, |\mathscr{T}_D^n|)$
  - $A_{ij} := t_j(x^{(i)})$ where $t_j(x^{(i)})$ is the j^{th} monomial in $\mathscr{T}_D^n$.
- The null space of A can be used to obtain a set of generators of $I(S_m)$.
  - Let $v_1, \ldots, v_k$ be a basis of the null space of A, i.e., $Av_i = 0$ for all i.
  - Any u such that $Au = 0$ can be written as a linear combination of the v_i.
- The polynomials $f_i(x) = \sum_{j=1}^{n} v_{ij} t_j(x)$ form a set of generators of $I(S_m)$.
- However, the number of columns of A is exponential in D.

Kernels Can't Help!

- Consider the kernel $k(x, y) = (1 + x \cdot y)^d$.
- The projection on the j^{th} principal component is given by $\sum_i \alpha_{ij} k(x, x^{(i)})$.
- The vanishing polynomials cannot be expressed in this fashion.

Theorem

Let k be a reproducing kernel and $f \in span(k(\cdot, x^{(i)}))$ such that f vanishes on all $x^{(i)}$. Then, f is the zero function.

Proof.

Let K be a kernel matrix, we need to prove that for every $\boldsymbol{\alpha}$ in the null space of K, we have $\boldsymbol{\alpha}^\top v = 0$.

1. Given x, define $\tilde{K} = \begin{bmatrix} c & v^\top \\ v & K \end{bmatrix}$, where $c := k(x, x)$ and $v_i := k(x^{(i)}, x)$.
2. The Schur complement of c is $A = K - \frac{1}{c} vv^\top$.
3. For any $\boldsymbol{\alpha}$ in the null space of K we will have $|alpha^\top v|^2 = -c\boldsymbol{\alpha}^\top A \boldsymbol{\alpha}$.

Vanishing Component Analysis

VCA

parameters:
Tolerance ϵ ; "FindRangeNull" procedure
initialize:
$F = \{f(\cdot) = 1/\sqrt{m}\}$, $V = \emptyset$
$C_1 = \{f_1, \ldots, f_n\}$ where $f_i(\boldsymbol{x}) = x_i$
$(F_1, V_1) = \text{FindRangeNull}(F, C_1, S_m, \epsilon)$
$F = F \cup F_1$, $V = V \cup V_1$
for $t = 2, 3 \ldots$
$C_t = \{gh : g \in F_{t-1}, h \in F_1\}$
if $C_t = \emptyset$
break
$(F_t, V_t) = \text{FindRangeNull}(F, C_t, S_m, \epsilon)$
$F = F \cup F_t$, $V = V \cup V_t$
output: F, V

Figure 1. Our Vanishing Component Analysis (VCA) algorithm for finding a generator set for $I(S_m)$.

FindRangeNull (using SVD)

input: F, C, S_m, ϵ
denote $k = |C|$ and $C = \{f_1, \ldots, f_k\}$
for $t = 1, \ldots, k$
let $\tilde{f}_i = f_i - \sum_{g \in F} \langle f_i(S_m), g(S_m) \rangle g$
let $A = [\tilde{f}_1(S_m), \ldots, \tilde{f}_k(S_m)] \in \mathbb{R}^{m,k}$
decompose $A = LDU^\top$ using SVD
for $i = 1, \ldots, k$
let $g_i = \sum_{j=1}^{k} U_{j,i} \tilde{f}_j$
output:
$F_1 = \{g_i / \|g_i(S_m)\| : D_{i,j} > \epsilon\}$
$V_1 = \{g_i : D_{i,j} \leq \epsilon\}$

Figure 2. The implementation of the FindRangeNull function in Figure 1.

Results

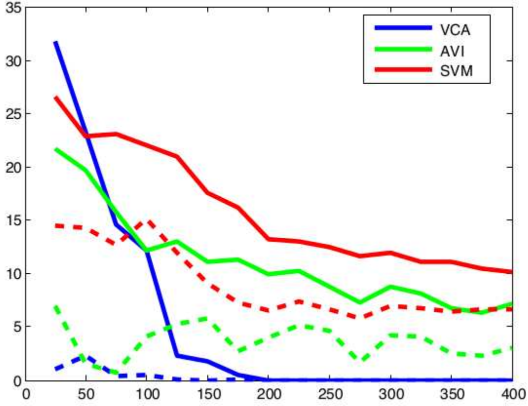

Figure 3. Toy dataset: The error percentage is depicted as a function of the sample size (solid lines are test errors, dashed lines are training errors).

Results

Data Set	Error Rate		Test Runtime	
	VCA	KSVM	VCA	KSVM
Pendigits (5996)	0.42	0.42	2.8e+003	9.6e+003
Letter (12000)	4.8	4.3	1.1e+003	7e+004
USPS (5833)	1.5	1.4	2.6e+003	3.8e+005
MNist (48000)	2.2	2	4e+03	3.1e+06

Table 1. Error rate in (%) and test runtime (in number of operations) for VCA and KernelSVM (training size in brackets). Results are averaged over 10 random 80%/20% train/test partitions.

Conclusions and Discussions

- vanishing component analysis
- find a set of features that describe a set of points S
- ...

Question?