TV-SVM: Support Vector Machine with Total Variational Regularization

Zhendong Zhang & Cheolkon Jung
zhd.zhang.ai@gmail.com zhengzk@xidian.edu.cn
School of Electronic Engineering, Xidian University, Xian 710071, China

Abstract

To leverage the spatial relationship of lattice data, such as images, we introduce total variational (TV) regularization into support vector machines (SVM), called TV-SVM. TV-SVM encourages local smoothness and sparsity in gradient domain of the learned parameters. TV-SVM is optimized via the alternating direction method of multipliers (ADMM) algorithm and is significantly better than (Linear) SVM for image classifications.

The Problem of SVM

Typically, SVM contains a hinge loss function and an L_2 regularization. Given a set of training data $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbf{R}^d$ and $y_i \in \{-1, 1\}$, SVM is formulated as:

$$\min_{w,b} \frac{1}{2}w^T w + C\sum_{i=1}^{n}[1 - y_i(w^T x_i + b)]_+ \quad (1)$$

Notice that each input x_i must be represented as a vector. However, some kinds of real-world data are naturally represented as matrices (images) or even tensors (videos).

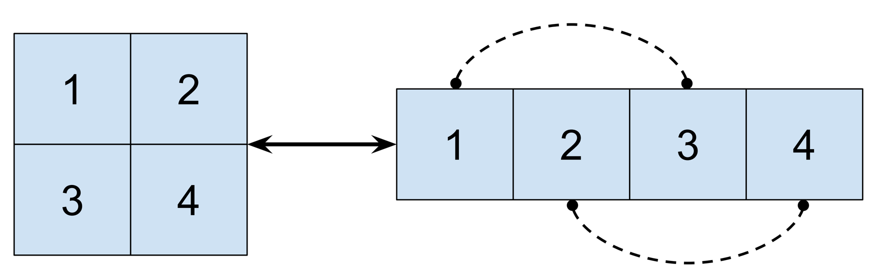


Figure: Reshape a matrix into a vector.

When matrices or tensors are reshaped into vectors, their spatial relationships are partially destroyed. Thus the spatial information is not successfully considered in SVM.

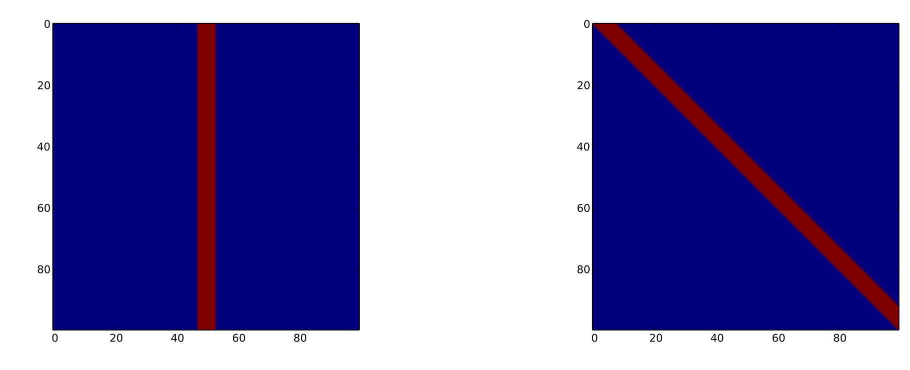
Low-rank Solutions

To deal with the above issue, several methods have been proposed

- rank-k SVM → $\mathbf{W}$ is the sum of k rank-1 matrices
- bilinear SVM → $\mathbf{W}$ is factorized into two low-rank matrices
- SMM → adds the nuclear norm of $\mathbf{W}$ as a regularization term

Nuclear norm introduced by Support Matrix Machine (SMM) is an approximation of matrix rank. Thus, all those methods are based on the low-rank assumption of $\mathbf{W}$, i.e. the rows or columns are highly correlated.

For natural image classification, the low-rank assumption may suffer problems. Rank is sensitive to rotations, but an image classification algorithm is required to be insensitive to rotations.



(a) Low-rank (b) High-rank

Figure: Rank is sesitive to rotations.

Total Variation

In this work, the TV-norm for 2D discrete signal is defined as

$$L_{TV}(X) = \sum_{i,j}|X_{i+1,j} - X_{i,j}| + |X_{i,j+1} - X_{i,j}| \quad (2)$$

where $\mathbf{X} \in \mathbf{R}^{m\times n}$ is the input matrix. L_{TV} can be also represented as linear operators combined with standard L_1 norm.

$$L_{TV}(X) = \|D_x x\|_1 + \|D_y x\|_1 \quad (3)$$

where x is the vector representation of $\mathbf{X}$. $\mathbf{D}_x$ and $\mathbf{D}_y$ are the so-called 2D forward differentiation operators in x-direction and y-direction, respectively. For simplicity, let $D = [D_x; D_y]$, then (3) is rewritten as:

$$L_{TV}(X) = \|Dx\|_1 \quad (4)$$

Then the spatial relationships of images are carried on $\mathbf{D}$.

TV-SVM

To take the advantage of the local smooth assumption and restrain the regression matrix $\mathbf{W}$ to be sparse in the gradient domain, we introduce TV regularization of $\mathbf{W}$ into SVM, instead of low-rank regularization.

$$\min_{\mathbf{W},b} \frac{1}{2}tr(\mathbf{W}^T\mathbf{W}) + \tau L_{TV}(\mathbf{W}) + C\sum_{i=1}^{n}[1 - y_i(tr(\mathbf{W}^T\mathbf{X}_i) + b)]_+ \quad (5)$$

Denote x and w are the vector representations of $\mathbf{X}$ and $\mathbf{W}$. Then, TV-SVM can also be represented as vectorized form:

$$\min_{w,b} \frac{1}{2}w^T w + \tau|Dw|_1 + C\sum_{i=1}^{n}[1 - y_i(w^T x_i + b)]_+ \quad (6)$$

It is equivalent to the following problem:

$$\arg\min_{w,b,z} f(w,b) + g(z) \quad s.t.\ z - w = 0 \quad (7)$$

$$f(w,b) = \frac{1}{2}w^T w + C\sum_{i=1}^{n}[1 - y_i(w^T x_i + b)]_+ \quad (8)$$

$$g(z) = \tau|Dz|_1 \quad (9)$$

Because both f and g are convex, the whole problem is optimized by fast ADAM algorithm. In each iteration, (w_k, b_k) is solved in a similar way as SVM and z_k is solved via an optimized taut-string method.

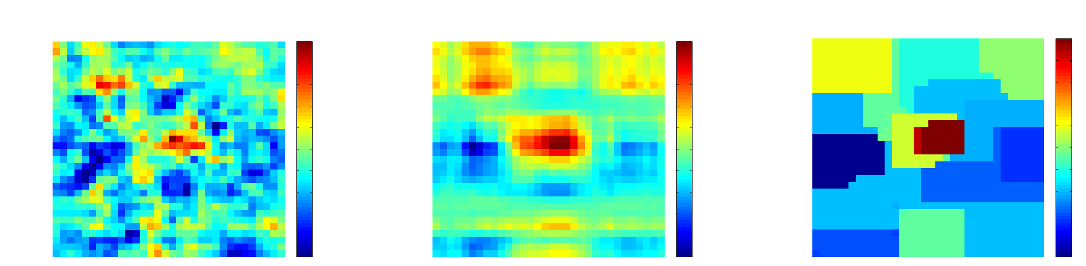
Fast ADMM for TV-SVM

Initialize $z^{-1} = \hat{z}^0$, $u^{-1} = \hat{u}^0$, $\rho = 1$, $t^1 = 1$, $c^0 = 0$ and $\eta \in (0,1)$
for $k = 1, 2, 3...$ **do**
  $(w^k, b^k) = \arg\min_{w,k} f(w,b) + \frac{\rho}{2}\|\hat{z}^k - w + \hat{u}^k\|_2^2$
  $z^k = \arg\min_z g(z) + \frac{\rho}{2}\|z - w^k + \hat{u}^k\|_2^2$
  $u^k = \hat{u}^k + \rho(z^k - w^k)$
  $c^k = \frac{1}{\rho}\|u^k - \hat{u}^k\|_2^2 + \rho\|z^k - \hat{z}^k\|_2^2$
  if $c^k < \eta c^{k-1}$ **then**
    $t^{k+1} = \frac{1 + \sqrt{1 + 4(t^k)^2}}{2}$
    $\hat{z}^{k+1} = z^k + \frac{t^k - 1}{t^{k+1}}(z^k - z^{k-1})$
    $\hat{u}^{k+1} = u^k + \frac{t^k - 1}{t^{k+1}}(u^k - u^{k-1})$
  else
    $t^{k+1} = 1$
    $\hat{z}^{k+1} = z^{k-1}$, $\hat{u}^{k+1} = u^{k-1}$
    $c^k = \frac{c^{k-1}}{\eta}$
  end if
end for

Experiments

Method	INRIA person		CIFAR-10		IG02	
	Time	Accuracy	Time	Accuracy	Time	Accurcy
L-SVM	-	80.12	-	67.60	-	72.91
SMM	**42.21**	81.98	15.40	**68.62**	**11.62**	73.68
TV-SVM	42.87	**82.99**	**14.79**	68.23	11.80	**73.86**

Table: Training time (second) and classification accuracy (%). There 2114 samples (160×90) on INRIA person, 2037 samples (32×32) on CIFAR-10, and 785 samples (90×120) on IG02.



(a) L-SVM (b) SMM (c) TV-SVM

Figure: Normalized $\mathbf{W}$ learned from CIFAR-10.

References

[1] *Support matrix machines.* ICML, 2015.

[2] *Fast alternating direction optimization methods.* Siam Journal on Imaging Sciences, 2014.

[3] *Modular proximal optimization for multidimensional total-variation regularization.* Mathematics, 2014.