Robust classification using $\ell_{2,1}$-norm based regression model

Yunfei Wang

Department of Computer Science & Technology
Huazhong University of Science & Technology

November 16, 2012

Table of contents

Basic Regression Model

N test samples:$Y = [y_1, y_2, \cdots, y_N] \in \mathbb{R}^{d \times N}$
n training samples:$A = [a_1, a_2, \cdots, a_n] \in \mathbb{R}^{n \times d}$
We aim to obtain a coefficient matrix $X \in \mathbb{R}^{n \times N}$ such that

$$Y \approx A^T X$$

This formulation actually indicates that the combination of training samples will be used to approximate each sample.

Properties of $\ell_{2,1}$-norm

The $\ell_{2,1}$-norm is defined as

$$\|A\|_{2,1} = \sum_{i=1}^{j} \sqrt{\sum_{j=1}^{m} a_{ij}^2} = \sum_{i=1}^{n} \|a_i\|^2$$

- Decrease the negative impact of outliers to which ℓ_2-norm is fairly sensitive.
- Rotational invariant
  For any orthogonal matrix R and data transformation:$x_i = Rx_i$,the $\ell_{2,1}$-norm is invariant $\|x_i\|_{2,1} = \|Rx_i\|_{2,1}$.

Minimizing Loss

Here we use $\ell_{2,1}$-norm as a criterion to evaluate approximation:

$$arg\min_{X} J(X) = \|A^T X - Y\|_{2,1} \tag{1}$$

in which $X = [x_1, x_2, \cdots, x_N]$ is the representation matrix and $x_i \in \mathbb{R}^n (i = 1, 2, \cdots, N)$ is the representation of test sample y_i using the entire training samples.

Adding regularization term

The $\ell_{2,1}$-norm regularization term $R(x)$ with a parameter ρ is performed to select class-specific samples with some sparsity. $R(X) = \|X\|_{2,1}$ and ρ is the penalty coefficient,the regularized form of Eq.(1) is equivalent to

$$\begin{aligned} arg\min_X J(X) + \rho R(X) &= arg\min_X \|A^T X - Y\|_{2,1} + \rho\|X\|_{2,1} \\ &= arg\min_X \frac{1}{\rho}\|A^T X - Y\|_{2,1} + \|X\|_{2,1} \end{aligned} \tag{2}$$

Reformulating the Objective function

Suppose $E = \frac{1}{\rho}(Y - A^T X)$,$B = (A^T, \rho I)$ and $U = \begin{pmatrix} X \\ E \end{pmatrix}$

$$
\begin{aligned}
&\frac{1}{\rho}\|A^T X - Y\|_{2,1} + \|X\|_{2,1} \\
&= \|E\|_{2,1} + \|X\|_{2,1} \\
&= \left\| \begin{matrix} X \\ E \end{matrix} \right\|_{2,1} \\
&= \|U\|_{2,1}
\end{aligned}
\tag{3}
$$

subject to $A^T X + \rho E = Y$.

Finally,the objective criterion can be formulated as

$$
U^* = arg \min_U |U|_{2,1} \ s.t. \ BU = Y \tag{4}
$$

Solving the optimization problem

An efficient algorithm was proposed to solve this specific problem,in which gradient-descent approach is implemented on the Lagrangian function to obtain the iterative solution

$$U = D^{-1}B^T(BD^{-1}B^T)^{-1}Y \tag{5}$$

in which D is a diagonal matrix with the i-th diagonal element as $d_{ii} = \frac{1}{2\|u_i\|^2}$

Usually,the matrix $BD^{-1}B^T$ is added by a diagonal matrix ζI to avoid the singular problem,where ζ is a very small positive constant and I is the unit matrix.

Reference

F.Nie,H. Huang,X. Cai,C. Ding.Efficient and robust feature selection via joint $\ell_{2,1}$-norm minimization,in:Twenty-Fourth Annual Conference on Neural Information Processing Systems,2010.

Classification Tactics

1. Extracting the optimal representation $\hat{X}$(the first n rows of matrix U^*).
2. Define C functions δ_s to select only those $\hat{x}$ that correspond to class s,where $\hat{x}$ is the representation for the test sample y.
3. Approximation in terms of the coefficients associated with the sth class is $\hat{y} = A^T\delta_s(\hat{x})$.
4. Assign y to the class that minimizes $\|y - A^T\delta_s(\hat{x})\|_2$

Pseudo-code of the algorithm

Input: $A \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^{d\times N}, \rho$;
Output: ID(Y);

Set $B = \left[A^T, \rho I\right], t = 0$;
Initialize $D_t \in \mathbb{R}^{(n+d)\times(n+d)}$ *as an identity matrix*;
repeat `// Iteratively compute U`
  Calculate $U_{t+1} = D_t^{-1} B^T (B D_t^{-1} B^T)^{-1} Y$;
  Calculate the diagonal matrix D_{t+1},where the i-th diagonal element is $\frac{1}{2\|u_{t+1}^t\|_2}$;
  $t = t + 1$;
until *Convergence*;
for $j = 1$ **to** N **do**
  Get $r_s(y_j) = \|y - A^T \delta_s(x_j)\|_2, s = 1, 2, \cdots, C$;
  $ID(y_i) = \min\limits_{s=1,2,\cdots,C} r_s(y_j)$;
end
return $ID(Y) = ((ID(y_1), ID(y_2), \cdots, ID(y_n))$;

Experiment results of RRC

Classification accuracy on 4 datasets:Yale,YaleB,COIL and ORL.