Classification and feature selection using a primal-dual method and projection on structured constraints

Michel Barlaud,1 Antonin Chambolle2 and Jean-Baptiste Caillau3

(1) Université Côte d’Azur, CNRS, I3S (2) CEREMADE, CNRS & Paris-Dauphine PSL (3) Université Côte d’Azur, CNRS, Inria, LJAD

Abstract

This work concerns feature selection using supervised classification on high-dimensional datasets. The classical approach to project data onto a low dimensional space and classify by removing an appropriate quadratic cost. We first introduced a matrix of centers in the definition of this cost. Moreover, as quadratic costs are not robust to outliers, we proposed instead to use an ϵ cost (Huber loss to mitigate outliers issues). While control on sparsity is commonly obtained by adding an 1 constraint on the weights vector used for projecting the data, we propose to enforce structured sparsity. To this end we used constraints that take into account the matrix structure of the data based either on the nuclear norm, on the ℓ1 norm, or on the ℓ0 norm for which we provide a new projection algorithm. We optimize simultaneously the projection matrix and the centers with a tailored constrained primal-dual method. The primal-dual framework is general enough to encompass the various robust losses and structural constraints and allows for a convergence analysis. We demonstrate the effectiveness of this approach on three biological datasets. Our primal-dual method with robust losses, adaptive centers and structured constraints does significantly better than classical methods, both in terms of accuracy and computational time.

Problem statement: Minimization of the ℓ1 loss

Let X be the m × d data matrix made of m line samples x1, . . . , xm that belong to the d-dimensional space of features. Let Y ⊆ {0, 1}m be the matrix of labels where k ≥ 2 is the number of clusters. Projecting the data in lower dimension is crucial to be able to separate them accurately. Let W be the d × k projection matrix, where k < d.

\[ \min \| Y - WX \|_2^2 + \frac{1}{2} \| \ell \|_2^2 \text{ s.t. } \| W \|_2^2 \leq \eta \] (1)

where \( I \) denotes the k identity matrix. An ℓ2 regularization is added in order to avoid the trivial solution (W, µ = 0). We use the algorithmic setting described in [2].

A primal dual algorithm

Problem (1) is rewritten in the form of the saddle point problem:

\[ \min_{W \in \mathbb{R}^{d \times k}} \max_{\mu \geq 0} \sum_{i=1}^{m} \left( \langle Y_i, Wx_i \rangle + \frac{1}{2} \| \ell_i \|_2^2 \right) \text{ s.t. } \| W \|_2^2 \leq \eta \] (2)

which can be solved by a primal-dual algorithm as studied in [2] (Algorithm 1).

Algorithm 1 Primal-dual algorithm, ℓ1 loss:

1. Input: X, Y, d, \( \gamma, \eta, \mu, p \in \mathbb{R}, W, \ell, \) do
2. for i = 1, . . . , N do
3. \[ W_{\text{old}} = W \]
4. \[ W = \text{proj}_{W \leq \mu} W + \gamma (X^T Z, \eta) \]
5. \[ \mu = \max \left\{ \frac{1}{\gamma}, \frac{1}{\gamma} (X^T Z, \eta) \right\} \]
6. \[ Z = X (W W^T - W W^T + \gamma I) (X^T Z, \eta) \]
7. \[ \ell = \max (-\mu, 1, \| Z \|_2) \]
8. end
9. Output: W

The convergence condition on the step-sizes \( \gamma \), \( \eta \), and \( \mu \) are given in [1]. The drawback of the term \( \| Y - WX \|_2^2 \) is that it enforces equality of the two matrices out of a sparse set. In order to soften this behaviour, we use the Huber function instead of the ℓ1 norm. In the algorithm, it simply consists in replacing line 7 with the appropriate “prox”, in this case we divide Z by the expression at line 7 with \( \gamma = \infty \) for a small parameter \( \gamma = 0 \) before truncating at \( -\mu \) and \( \mu \).

Group LASSO constraint

Group LASSO was first introduced in [3]. The main idea is to enforce parameters of different classes to share common features. Group sparsity reduces complexity by eliminating entire features. It consists in using the ℓ1,2 norm for the constraint on W, which is defined as follows. The rowwise (ℓ2,1) norm of a d × k matrix \( W \) whose rows are denoted \( w_1, \ldots, w_m \) is:

\[ \| W \|_{2,1} = \sum_{i=1}^{d} \| w_i \|_2 \] (3)

Projecting a matrix \( W \) on this ball is easy as it amount to project first the norms of the rows \( \| w_i \|_2 \), on the ℓ2 ball of radius \( \| w_i \|_2 \) with \( \sum_{i=1}^{d} \| w_i \|_2 \leq \eta \), and then each row \( w_i \) of \( W \) on the ℓ2 ball of radius \( \| w_i \|_2 \). In order to solve Problem (1) with this new constraint, we simply replace at line 4 of Algorithm 1 the projection onto the ℓ1 ball with the appropriate modified projection.

Exclusive LASSO constraint

Exclusive sparsity or exclusive LASSO was first introduced in [4]. The main idea is that if one feature in a class is selected (large weight), the method tends to assign small weights to the other features in the same class. So given a d × k matrix \( X \), the projection on the corresponding ball consists in finding a matrix \( W \) which solves

\[ \min \| W \|_{2,1} \text{ s.t. } \sum_{i=1}^{d} \| w_i \|_2^2 \leq \eta^2 \] (4)

Our approach is to introduce a Lagrange multiplier for the constraint and then compute it by a variant of Newton’s method.

Results and comparison of methods

Ovarian protomic dataset is available on UCI database consists of mass-spectra obtained with the SELDI technique. The dataset is comprised of 216 samples, 1500 features and two clusters. Lung protomic dataset was collected using unbiased liquid chromatography/mass spectrometry. The dataset is comprised of 1085 patients (493 among them with lung cancer and 592 control patients), and 224 features and and two clusters. Zebrafish is a Single cell dataset composed of 3890 cells, 7364 genes and k = 9 clusters.

The table above shows the improvement in accuracy on all biological datasets when using Huber loss instead of ℓ1 or Frobenius loss: ℓ1 loss suffers from overfitting while Frobenius loss is not robust enough. Optimizing the matrix of centers (i.e. fixing \( \mu = 1 \)) also improves accuracy on the three datasets.

Methods

\begin{align*}
\text{F1} & \quad \text{Huber (w) } \quad \text{Huber (f)} \\
\text{Accuracy} & \quad 70.61 \% \quad 70.98 \% \quad 70.98 \% \quad 76.61 \% \quad 76.61 \% \\
\text{AUC} & \quad 89.42 \% \quad 93.54 \% \quad 93.54 \% \quad 97.30 \% \quad 97.60 \% \\
\end{align*}

The second table and the leftmost figure above shows that computational time as a function of the number of samples is linear both for primal-dual and FISTA and that the computational time of ADMM is one or order of magnitude greater than the others because of the linear algebra involved. The figure on the rightmost shows the computational time as a function of the number of features: primal-dual scales much more favorably than FISTA wt. the number of features (a key issue for biological datasets for which the number of genes is large).

References