Effective Sparse Coding Algorithms-NISP06

Yunfei Wang

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

November 23, 2012

Fast algorithms for solving two general-purpose convex problems:

1. L1-regularized Least Squares problem solver using the feature-sign search algorithm.
2. L2-constrained Least Squares problem solver using Lagrange dual.

Optimization Problem

Sparse coding represent input vectors approximately as a weighted linear combination of a small number of "basis vectors", which can be applied to learning overcomplete basis sets, in which the number of bases is greater than the input dimension.

$$\min_{S,B} \underbrace{\|X - BS\|_F^2}_{\text{data fitting}} + \underbrace{\beta\phi(S)}_{Sparsity} \tag{1}$$

$$\text{subject to } b_i^2 \leq c, \forall i = 1, \cdots, n.$$

where $X \in \mathbb{R}^{k\times m}$ is input matrix(each column is an input vector),$B \in \mathbb{R}^{k\times n}$ is the basis matrix(each column is a basis vector) and $S \in \mathbb{R}^{n\times m}$ is the coefficient matrix(each column is a coefficient vector).

Sparsity Function

Usual sparsity-inducing regularizers:

$$\phi(s_j) = \begin{cases} \|s_j\|_0 & (L_0 \text{ penalty function}) \\ \|s_j\|_1 & (L_1 \text{ penalty function}) \\ (s_j^2 + \epsilon)^{\frac{1}{2}} & (epsilon L_1 \text{ penalty function}) \\ \log(1 + s_j^2) & (\text{log penalty function}) \end{cases} \tag{2}$$

L_1 regularization is known to produce sparse coefficients and robust to irrelevant features.
Log regularization make gradient-based methods get stuck in local optima.

Strategy in this Paper

The optimization problem is convex in B(while keeping S fixed) and convex is S(while keeping B fixed),but not convex in both simultaneously.
Iteratively optimize the objective by alternatingly optimizing with respect to B(bases) and S(coefficients) while holding the other fixed:

- For learning bases B, the optimization problem is a least squares problem with quadratic constrains. *Lagrange dual* will be used to solve the problem.
- For learning coefficients S, the optimization problem equals a L_1-regularized least squares problem.*Feature-sign search algorithm* will be used to tackle the problem.

Learning the coefficients S

The problem can be solved by optimizing over each s_i individually:

$$\min_{s} \|x - Bs\|^2 + \beta\|x\|_1 \tag{3}$$

Main Idea:If we know the signs at the optimal value,Eq.(3) can be reduced to a standard and unconstrained quadratic problem.

The algorithm firstly guess the signs of coefficients s, then solves the resulting unconstrained QP and finally refines the initially incorrect guess.

Learning the bases B

Given fixed coefficients S,the optimization problem over bases B reduces Eq.(1) to the following least squares problem with quadratic constraints,which is independent of the sparsity function:

$$\min_{S,B} \|X - BS\|_F^2 \tag{4}$$

$$\text{subject to } b_i^2 \leq c, \forall i = 1, \cdots, n.$$

where b_i is a basis vector in B.
Lagrange dual can be more efficient than gradient descent with iterative projection when applied to the above optimization problem.

Lagrange dual I

First,consider the Lagrangian:

$$\mathcal{L}(B,\lambda) = Tr((X-BS)^T(X-BS)) + \sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{k}B_{i,j}^2 - c) \quad (5)$$

where each $\lambda_j \geq 0$ is a dual variable.
Minimizing over B analytically,we obtain the Lagrange dual:

$$\mathcal{D}(\lambda) = \min_{B}\mathcal{L}(B,\lambda) = Tr(X^TX - XS^T(SS^T+\Lambda)^{-1}(XS^T)^T - c\Lambda) \quad (6)$$

where $\Lambda = diag(\lambda)$.

Lagrange dual II

The gradient and Hessian of $\mathcal{D}(\lambda)$ are as follows:

$$\frac{\partial \mathcal{D}(\lambda)}{\partial \lambda_i} = \|XS^T(SS^T + \Lambda)^{-1}e_i\|^2 \tag{7}$$

$$\begin{aligned}\frac{\partial^2 \mathcal{D}(\lambda)}{\partial \lambda_i \partial \lambda_j} = & -2((SS^T + \Lambda)^{-1}(XS^T)^T XS^T(SS^T + \Lambda)^{-1})_{ij} \\ & \times ((SS^T + \Lambda)^{-1})_{ij}\end{aligned} \tag{8}$$

where $e_i \in \mathcal{R}^n$ is the i-th unit vector.

Optimizing Eq.(6) using Newton's method or conjugate gradient,we obtain the optimal bases B as follows:

$$B^T = (SS^T + \Lambda)^{-1}(XS^T)^T \tag{9}$$