AM221 Final Project Proposal

Taylor Killian & Leonhard Spiegelberg

April 29, 2016

Abstract

To be filled with an interesting summary of our work and results

1 Dictionary Learning as 2-stage supermodular minimzation

The introduced problem of Dictionary learning, which in its general form can be seen as

$$\min_{D,R,\theta,\lambda} f(D,R,X) + g(D,\theta) + h(R,\lambda)$$

where f describes the objective function used to measure goodness of approximation of X through D, R, g describing suitable constraints on the dictionary, h on the representation respectively. With an input dataset $X = [x_1, \dots, x_k], x_i \in \mathbb{R}^d, X \in \mathbb{R}^{d\times k}$ we wish to find a dictionary $D \in \mathbb{R}^{d\times n}, \mathcal{D} = [d_1, \dots, d_n]$ and a representation $\mathcal{R} = [r_1, \dots, r_k],\ r_i \in \mathbb{R}^n,\ \mathcal{R} \in \mathbb{R}^{n\times k}$, such that both $\|X - \mathcal{D}\mathcal{R}\|_F^2$ is minimized and the representations r_i are "sparse enough". To limit the dictionary becoming infinitely large (or small) we introduce a constraint on the dictionary's columns. For this any sufficient norm can be used. An often used norm is the l_2-norm. Thus for the dictionary learning problem we introduce the problem with constraints

$$\min_{D,R} \|X \ - DR\|_F^2 \tag{1}$$

$$\text{s.t.} \qquad \|d_j\|_2 \leq 1, \forall j = 1, ..., n \tag{2}$$

$$\|r_i\|_0 \leq t, \forall i = 1, ..., k \tag{3}$$

Description of constraints here

Using Lagrange multipliers this can be brought to the general form above

$$\min_{D,R,\theta,\lambda} \|X \ - DR\|_F^2 + \sum_{j=1}^{m} \theta_j(\|d_j\|_2 - 1) + \sum_{i=1}^{k} \lambda_i(\|r_i\|_0 - t)$$

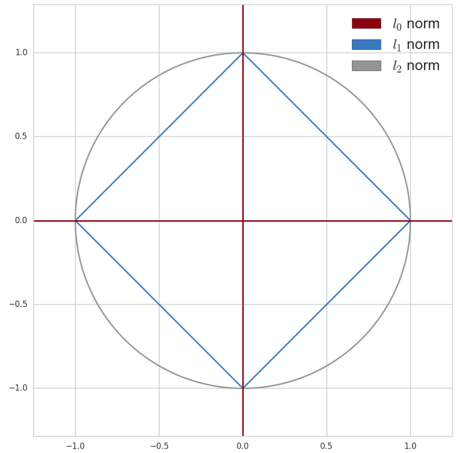

Figure 1: contour plot of the l_0, l_1, l_2 norm at levelset $L_1(f) := \{x \in \mathbb{R}^2 : f(x) = 1\}$. Note that for any space $\mathbb{R}^d$, the image of the l_0 consists of $d + 1$ values $(\{0, ..., d\})$.

I.e. the functions are

$$f(D, R, X) := \|X \;\; - DR\|_F^2$$

$$g(D, \theta) := \sum_{j=1}^{n} \theta_j(\|d_j\|_2 - 1)$$

$$h(R, \lambda) := \sum_{i=1}^{k} \lambda_i(\|r_i\|_0 - t)$$

How to bring this to the penalty function form?

The difficulty in solving this problem comes from the mathematical challenge the $\|.\|_0$ norm yields. To understand this better confer **??**. Let $x \in \mathbb{R}^d$

$$l_0(x) := (\#i : x_i \neq 0)$$

here describe different norms... the goal is to find a norm that sort of converges to the infinite + of the l0 norm

Put here section about possible relaxations of l0 norm (related literature)

1.0.1 Two stage thoughts

We will now relax the problem by formulating it into two stages:

1. solve for fixed D, θ the minimization to obtain optimal R, λ.
2. solve optimization for D, θ with obtained values

3. repeat previous steps until done

This is actually the method of optimal directions! In the traditional method of optimal directions (ref here) the subproblem $\min_{R,\lambda} f(D,R,X)+g(D,\theta)+h(R,\lambda)$ is solved either by various relaxation methods (i.e. LASSO, Matching Pursuit(maybe describe here)), however for our approach we want to embedd the problem into a combinatorial optimization framework, which guarantees us bounds on optimality. Recent work done by [?] showed that the subproblem of finding an optimal representation with a given dictionary is equivalent to Sparse Multiple Linear Regression (SMLR) and can be reformulated as a α-weakly supermodular function. We will now show how to obtain this formulation: Remember

$$\min_{R,\lambda} f(D,R,X)+g(D,\theta)+h(R,\lambda)$$
$$=\min_{R,\lambda} \|X-DR\|_F^2+\sum_{j=1}^{n}\theta_j(\|d_j\|_2-1)+\sum_{i=1}^{k}\lambda_i(\|r_i\|_0-t)$$

then let $S\subseteq[n]$ be a set of column indices for the dictionary. From now onwards we will assume that the given, fixed dictionary is normalized, i.e. $\forall d_i : \|d_i\|_2 = 1$. This allows us to remove the term $g(D,\theta)$. Define now D_S as the matrix obtained from D with all columns not indexed by S to be set to zero. D_S^+ is the pseudoinverse which can be obtained i.e. via singular value decomposition when $D_S = U\Sigma V^T$ through $D_S^+ = V\Sigma^+U^T$. Then the problem can be equivalently stated as

$$\min_{R,\lambda} \|X-DR\|_F^2+\sum_{i=1}^{k}\lambda_i(\|r_i\|_0-t)$$
$$\iff \min_{S\subseteq[n],|S|\leq t} \|X-D_SD_S^+X\|_F^2$$

[?] have shown that the objective function $f(S) := \|X-D_SD_S^+X\|_F^2$ of this optimization problem is α-weakly supermodular with

$$\alpha=\max_{\tilde{S}\subseteq[n]}\|D_{\tilde{S}}^+\|_F^2$$

1.1 How to combine it

For the two stage formulation we can generally think of the problem as of adapting the idea of the method of optimal directions but instead of projecting the given representation onto the dictionary we strive for an additional minimization problem: Let $f_1,...,f_m$ be instances of the SMLR problem. Then the dictionary problem can be seen as

$$\min_{D,\theta} q(f_1(D,\theta),...,f_m(D,\theta))+\sum_{j=1}^{n}\theta_j(\|d_j\|_2-1)$$

for some combination function q. What properties should the combination function satisfy?
that is a great question! Can we formulate some ideas, thoughts on what it should look like?
One choice for q could be

$$q(f_1(D,\theta), ..., f_m(D,\theta)) := \sum_{i=1}^{m} f_i(D,\theta)$$

another one to minimize the product of the SMLR instances. Reformulating it with a logarithm might give even better results with increased computationally tractability

$$q(f_1(D,\theta), ..., f_m(D,\theta)) = \sum_{i=1}^{m} \log(f_i(D,\theta))$$

(redwe can omit the logical step of log'ing also the lambda term because it is a constraint...)

2 Introduction

Introduce the problem and it's complexities

3 Background Information

3.1 Background

Here we'll want to add some information about Sparse regression, set functions and definitions of sub modular and super modular, Dictionary Selection, Greedy Methods, etc.

3.2 Related Work

A simple overview of prior work

4 Methods

Place holder section to begin to outline our work

5 Experiments

Hold for applying our methods

5.1 Results

The results of our algorithms against others

Algorithm 1 Notional algorithm

```
Figure out optimal arrangement of rows and columns of input data
while there's still time in the semester do
    if $f(p_k + s_k\lambda_i) < f(p_k)$ for some $\lambda_i$ then
        $p_{k+1} = p_k + s_k\lambda_i$
        $s_{k+1} = s_k$
    else
        $p_{k+1} = p_k$
        $s_{k+1} = \alpha s_k$
    end if
end while
return VICTORIOUS
```

6 Conclusions & Future Work

6.1 Conclusions

We will have done it!

6.2 Future Work

Rule the world

References

[1] Christos Boutsidis, Edo Liberty, and Maxim Sviridenko. Greedy minimization of weakly supermodular set functions. *CoRR*, abs/1502.06528, 2015.