EQ2415 – Machine Learning and Data Science
HT22

Tutorial 5

A. Honoré, A. Ghosh

1 Sparse Representation

1.1 Norms

Question 1. A norm is used for quantifying a measure of distance between two vectors. But, it should obey certain axioms in order to be called a valid norm. List all such axioms that must be followed by a valid norm.

Solution: (a) Norm of a vector $\mathbf{x}$ should obey non-negativity, homogeneity, positive and the triangle inequality. ∎

Question 2. You have encountered the p-norm in the slides $\|\mathbf{x}\|_p$. For $p < 1$, is this still a norm. If not, which property does this norm violate?

Solution: Triangle inequality is not satisfied. Show by counterexample, let $\mathbf{x} = \begin{bmatrix} a & 0 & \dots & 0 \end{bmatrix}^\top$, $\mathbf{y} = \begin{bmatrix} 0 & a & \dots & 0 \end{bmatrix}^\top$, $p < 1$. ∎

1.2 The P_0 problem

Reducing the l_0 norm of a vector arises in problems where we wish to reconstruct a vector $\mathbf{b}$, from a linear combination of a minimum number of columns of a matrix $\mathbf{A}$. This is called a P_0 problem and is one of our principal problem of interest.

$$P_0 : \quad \mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{x}\|_0 \ \text{ s.t. } \mathbf{Ax} = \mathbf{b} \tag{1}$$

where we have an under-determined problem setup with $\mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{b} \in \mathbb{R}^n$ and $m > n$. This problem is in fact NP-hard.

Question 3. What is the computational complexity for solving such a problem ? Use a simple numerical reasoning to illustrate your point.

Solution: Assume that the matrix size $= m \times n$, if we know that the sparsity level is k_0, to know these k_0 points by a brute force approach, one should sweep over $\binom{n}{k_0}$ possibilities, and at each such possibility test whether the constraint $\mathbf{Ax} = \mathbf{b}$ holds or not. When n is much larger than k_0, the binomial coefficient becomes exponential in n, so exhaustive search will definitely fail. Use some numbers such as $m = 500, n = 2000, k_0 = 20$ to convince yourself! ∎

Question 4. Propose two approximate formulations for the problem P_0.

Solution: e.g. introduce an ϵ for the reconstruction error:

$$\hat{P}_0 : \quad \mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{x}\|_0 \ \text{ s.t. } ||\mathbf{b} - \mathbf{Ax}|| \leq \epsilon \tag{2}$$

or use the l_1 norm instead of l_0.

$$\hat{P}_0 \equiv P_1 : \quad \mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{x}\|_1 \ \text{ s.t. } \mathbf{b} = \mathbf{Ax} \tag{3}$$

There are two main approaches for solving the P_0 problem:

- Finding the support of $\mathbf{x}$: Discrete problem and solved often using greedy algorithms.
- Smoothing penalty schemes: l_1 minimization.

∎

1.3 The spark of a matrix

A special case of P_0 allow us to find a quantity called the *spark*.
Question 5. (a) Define the rank, nullity and spark of a matrix.

Solution: Rank of a matrix $\mathbf{A}$ is the largest number of linearly independent columns of $\mathbf{A}$.
Nullity of a matrix is the dimension of the nullspace i.e. dimension of the set $\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}, \mathbf{x} \neq 0\}$.
The spark of a matrix is the **smallest** number of **linearly dependent** columns.

$$\text{spark}\,(\mathbf{A}) = \min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{x}\|_0 s.t. \mathbf{A}\mathbf{x} = \mathbf{0}, \mathbf{x} \neq \mathbf{0} \tag{4}$$

■

Question 5. (b) Consider a matrix that is constructed as $\mathbf{I}_n - \mathbf{S}_n$, where $\mathbf{S}_n$ is a real, skew-symmetric matrix. Calculate the rank, nullity and spark of $\mathbf{I}_n - \mathbf{S}_n$. *Hint:* Use Schur's determinant identity for a block matrix, and the relation that $\det(A + c^T r) = \det(A) + r^T \text{adj}(A)c$

Solution: The trick here is to show that $\mathbf{I}_n - \mathbf{S}_n$ is non-singular, i.e. it is full rank. This can be shown by mathematical induction:
for $n = 2$, we have

$$\begin{aligned} \mathbf{I}_2 - \mathbf{S}_2 &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & -a_1 \\ a_1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & a_1 \\ -a_1 & 1 \end{bmatrix} \end{aligned} \tag{5}$$

$\det(\mathbf{I}_k - \mathbf{S}_k) = 1 + a_1^2 \geq 0$ as a_1 is real.
for $n = k$, we assume $\mathbf{I} - \mathbf{S}$ is full rank and has a positive determinant. Then for $n = k + 1$, we have:

$$\begin{aligned} \det(\mathbf{I} - \mathbf{S}) &= \det\left(\begin{bmatrix} \mathbf{I}_k - \mathbf{S}_k & \mathbf{a}_{k+1} \\ -\mathbf{a}_{k+1}^\top & 1 \end{bmatrix}\right) \\ &= \det(\mathbf{I}_k - \mathbf{S}_k)\det\left(1 + \mathbf{a}_{k+1}^\top(\mathbf{I}_k - \mathbf{S}_k)^{-1}\mathbf{a}_{k+1}\right) \\ &= \det(\mathbf{I}_k - \mathbf{S}_k) + \mathbf{a}_{k+1}^\top adj\,(\mathbf{I}_k - \mathbf{S}_k)\,\mathbf{a}_{k+1} \geq 0 \end{aligned} \tag{6}$$

One can also show that the adjugate of $\mathbf{I}_k - \mathbf{S}_k$ is in fact positive semidefinite. Hence, it is non-singular and full rank. So, rank is n, nullity is 0, spark is $n + 1$. (think why?). ■

Question 5. (c) Assume that you have an algorithm which is known for giving you the sparsest solution $\mathbf{x}$ for a P_0 problem. It is also assumed that the matrix $\mathbf{A}$ in P_0 is square and full rank. One of your friend also comes and shows you a solution $\mathbf{y}$ for the same problem and claims as well that it is the sparsest. How can you resolve this conflict?

Solution: Check if $\|\mathbf{x}\|_0 < \frac{1}{2}\text{spark}\,(\mathbf{A})$ implying uniqueness through spark. ■

1.3.1 Some useful quantities

1.4 Greedy algorithms for P_0

Finding the spark of a matrix is a combinatorial problem. There exists Greedy algorithms to solve the P_0 problem approximately.

1.4.1 Orthogonal Matching Pursuit

Here, we consider a pre-specified sparsity level of $\mathbf{x}$ is known before the start of the algorithm. We specify this as $\|\mathbf{x}\|_0 = k$. The **support** of $\mathbf{x}$ defined as $\mathcal{S}_\mathbf{x} = \{i : x_i \neq 0\}$, The **non-zero elements** of $\mathbf{x}$ can be referred to as $\mathbf{x}_{\mathcal{S}_\mathbf{x}}$. Also, recall that the ℓ_0 norm of $\mathbf{x}$ is equal to the cardinality of the support set $\mathcal{S}_\mathbf{x}$. We know that the support set is characterized as $\mathcal{S}_\mathbf{x} = \{n : x_n \neq 0\}$. So, now the problem is find the k elements of this support set $\mathcal{S}_\mathbf{x}$, as we have pre-specified cardinality. A greedy solution avoids a brute-force $\binom{m}{k}$ search and instead tries to find an iterative solution. One such greedy algorithm is the

Orthogonal matching pursuit (OMP).

OMP is a greedy solution to the support-finding problem. We are assumed to be given as inputs: $\mathbf{A}$, $\mathbf{b}$ and k_0 (sparsity level of $\mathbf{x}$). OMP consists of the following parts:

1. Initialization
   - Set $k = 0$ (iteration counter)
   - Set initial support set $\mathcal{S}_x^{(0)} = \phi$
   - Set initial residual to be $\mathbf{r}^{(0)} = \mathbf{b}$
   - Set error threshold ε
2. Repeat until either $\|\mathbf{r}^{(k)}\|_2 < \varepsilon$ or max no. of iterations is completed or $\|\mathbf{r}^{(k)}\|_2 > \|\mathbf{r}^{(k-1)}\|_2$
   - Sweep stage: Compute errors $\epsilon(j) = \min_{z_j} \|\mathbf{a}_j z_j - \mathbf{r}^{(k-1)}\|_2^2$ (find the optimal choice) and then finding $i_k^\star = \text{argmin}_j \epsilon(j)$. This can be also done in one single step.
   - Update support $\mathcal{S}_x^{(k)} = \mathcal{S}_x^{(k-1)} \cup i_k^\star$
   - Update residual $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}_{\mathcal{S}_x^{(k)}} \mathbf{A}^\dagger_{\mathcal{S}_x^{(k)}} \mathbf{b}$
   - Update counter $k = k + 1$
3. Finally get $\hat{x} \in \mathbb{R}^N$ with $\hat{\mathbf{x}}_{\mathcal{S}_x^{(k)}} = \mathbf{A}^\dagger_{\mathcal{S}_x^{(k)}} \mathbf{b}$ and remaining part as zeros.

Question 6. Show that the sweep stage is equivalent to finding $i_k^\star = \text{argmax}_j \mathbf{A}^\top \mathbf{r}^{(k-1)}$?

Solution:

$$\begin{aligned} \epsilon(j) &= \min_{z_j} \|\mathbf{a}_j z_j - \mathbf{b}\|_2^2 \\ &= \|\mathbf{b} - \mathbf{a}_j \left(\frac{\mathbf{a}_j^\top \mathbf{b}}{\|\mathbf{a}\|_2^2} \right) \|_2^2 \\ &= \|\mathbf{b}\|_2^2 - \| \frac{\left(\mathbf{a}_j^\top \mathbf{b}\right)^2}{\|\mathbf{a}\|_2^2} \|_2^2 \end{aligned} \tag{7}$$

Replace $\mathbf{b}$ by $\mathbf{r}^{(k-1)}$ and we see that

$$i_k^\star = \text{argmin}_j \epsilon(j) = \text{argmax}_j \| \frac{\left(\mathbf{a}_j^\top \mathbf{r}^{(k-1)}\right)^2}{\|\mathbf{a}\|_2^2} \|_2^2 \tag{8}$$

and this means searching for the index that gives the largest amplitude of $\mathbf{A}^\top \mathbf{r}^{(k-1)}$ where columns of $\mathbf{A}$ are normalized. ■