ELE 535: Machine Learning and Pattern Recognition
Homework 7

Zachary Hervieux-Moore

Wednesday 28th November, 2018

Exercise 1: Sparse Representation in an ON Basis. Let $r \leq n$ and $Q \in \mathbb{R}^{n\times r}$ have orthonormal columns.

a) Find a solution of the following sparse approximation problem and determine if the solution is unique.

$$\min_{w\mathbb{R}^r} \|y - Qw\|_2^2$$
$$\text{s.t. } \|w\|_0 \leq k$$

b) Now let the columns of $X \in \mathbb{R}^{n\times m}$ be a centered set of unlabelled training data and the columns of $Q \in \mathbb{R}^{n\times r}$ be the left singular vectors of a compact SVD of X. In this context, interpret the solution of the above problem.

Answer:

a) Due to the invariance of orthonormal matrices (as shown in a previous homework), we have that the optimization problem in question is equivalent to the following. It is also easily verified by expanding both objectives.

$$\min_{w\mathbb{R}^r} \|Q^T y - w\|_2^2$$
$$\text{s.t. } \|w\|_0 \leq k$$

From here, it is evident that the solution is simply to take that largest k entries of $|Q^T y|$ and set it to w. That is, once you make an entry of w nonzero, it is best to set it equal to the corresponding entry in $Q^T y$. Since we can only have k non zero entries, we pick the k largest in absolute terms. The solution is not unique if $Q^T y$ has entries with the same absolute value for the k and $k+1$ largest entries.

b) If Q is the left singular vectors of an SVD of X, then we can think of w as being the best k combination of the left eigenvectors that approximate a label y. That is, if transmitting between two parties that know X, we can transmit a k-sparse vector w and use the SVD decomposition of X to recover y. This would be of practical use where transmission of bits it costly or error prone but X can be shared beforehand. For example, communication with satellites.

Exercise 2: Let

$$M = \begin{bmatrix} \boldsymbol{e}_1 & \frac{1}{\sqrt{2}}(\boldsymbol{e}_1 + \boldsymbol{e}_2) & \boldsymbol{e}_3 & \frac{1}{\sqrt{3}}(\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3) \end{bmatrix}$$

where $\boldsymbol{e}_i$ denotes the i^{th} standard basis vector in $\mathbb{R}^n$.

a) Show that the columns of M are linearly dependent.

b) Determine spark(M).

c) Determine the mutual coherence $\mu(M)$.

Answer:

a) We have that

$$\sqrt{3}M_4 = \sqrt{2}M_2 + M_3$$

where M_i is the i^{th} column of M.

b) We know that the lower bound for spark(M) is 2 and that part a) has shown an upper bound of spark(M) is 3. Thus, we just have to check if any pairwise combination of the columns are linear dependent. By inspection, this is not the case so spark$(M) = 3$.

c) We have that the columns all have norm 1. Thus, to find out $\mu(M)$, we can simply pick the largest entry in $M^T M$ that is not on the diagonal. This turns out to be $\langle M_2, M_4 \rangle$ which is $\mu(M) = \frac{\sqrt{6}}{3}$.

Exercise 3: Let $A \in \mathbb{R}^{m \times n}$ with $\text{rank}(A) = m < n$, and $y \in \mathbb{R}^m$. We seek the sparsest solution of $Ax = y$:

$$\min_{x \in \mathbb{R}^n} \|x\|_0, \text{ s.t. } Ax = y$$

The convex relaxation of this problem is called Basis Pursuit:

$$\min_{x \in \mathbb{R}^n} \|x\|_1, \text{ s.t. } Ax = y$$

Show that Basis Pursuit is equivalent to the linear program:

$$\begin{gathered} \min_{x, z \in \mathbb{R}^n} \mathbf{1}^T z \\ \text{s.t. } Ax = y \\ x - z \leq \mathbf{0} \\ -x - z \leq \mathbf{0} \end{gathered}$$

Answer: I will show this by contraction. Suppose by that x^*, z^* solves linear program but that $z^* \neq |x^*|$. Then, we have that $x^* - z^* < \mathbf{0}$ and $-x^* - z^* < \mathbf{0}$ for some entries. However, by decreasing the entries of z^* to make all entries either have $x^* - z^* = \mathbf{0}$ and $-x^* - z^* = \mathbf{0}$ will yield a smaller objective in the linear program. This contradicts that z^* is optimal and we conclude that $z^* = |x^*|$. This results in $\mathbf{1}^T z^* = \|x^*\|_1$ which is exactly the Basis Pursuit problem.

Exercise 4: One way to create a dictionary is to combine known ON bases. Here we explore combining the standard basis with the Haar wavelet basis. The Haar wavelet basis consists of $n = 2^p$ ON vectors in $\mathbb{R}^n$. These can be arranged into the columns of an orthogonal matrix H_p with

$$H_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \quad H_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$H_3 = \begin{bmatrix} \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{4}} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{4}} & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{4}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{4}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & 0 & \frac{1}{\sqrt{4}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & 0 & \frac{1}{\sqrt{4}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & 0 & -\frac{1}{\sqrt{4}} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & 0 & -\frac{1}{\sqrt{4}} & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

The columns are arranged in groups. The first group consists of the vector $\frac{1}{\sqrt{n}}\mathbf{1}_n$, and second consists of the vector taking the value $\frac{1}{\sqrt{n}}$ in the first half, and $-\frac{1}{\sqrt{n}}$ in the second half. Subsequent group of vectors are derived by subsampling by 2, scaling by $\sqrt{2}$, and translating. This is illustrated above for $p = 1, 2, 3$. Form a dictionary $D \in \mathbb{R}^{n \times 2n}$ by setting $D\ [I_n, H_p]$ with $n = 2^p$. The matrix H_p is the Haar matrix of size $n = 2^p$. Show that

a) For $p = 1$, $\text{spark}(D) = 3$ and $\mu(D) = 1/\sqrt{2}$.

b) For all $p > 1$. determine $\text{spark}(D)$ and $\mu(D)$.

c) For a given $y \in \mathbb{R}^n$, we seek the sparsest solution of $y = Dw$. What condition on y is sufficient to ensure the sparsest solution is unique.

Answer:

a) For $p = 1$, we have that

$$\sqrt{2}D_4 = D_1 - D_2$$

where D_i is the i^{th} column of D. Similar to question 2), visual inspection leaves the other pairwise combinations all linearly independent. Thus spark$(D) = 3$. Since I_n and H_p are orthonormal, then computing $\mu(D)$ will involve one column from I_n and one column from H_p. This is because if you pick two distinct columns in one of them, the dot product is 0. Since the columns of I_n are all 0 except for one entry with 1, the pairwise dot products between I_{ni} and H_{p_j} is simply $H_{p_{ij}}$. Thus, $\mu(D)$ is simply the largest entry in H_p. For $p = 1$, this is $1/\sqrt{2}$ and so we conclude that $\mu(D) = 1/\sqrt{2}$.

b) Using the same logic as part a), $\mu(D)$ is the largest value in H_p. By construction, this will always be $1/\sqrt{2}$. Thus, $\mu(D) = 1/\sqrt{2}$ for all $p > 1$. We know that a lower bound for spark(D) is 2. However, I will argue why that spark$(D) > 2$. Because H_p and I_n are orthonormal, the only way it is possible for spark$(D) = 2$ is to have linear dependence between one of the columns of H_p and I_n. By construction, all columns of H_p have at least 2 entries that are non zero and the columns of I_n have precisely one non zero entry. Thus, it is impossible to pick one one column from H_p and one from I_n that are linearly dependent. Thus spark$(D) > 2$. Now, we can always find the linear dependence

$$\sqrt{2}H_{p_n} = I_{nn} - I_{nn-1}$$

That is, the last column of H_p is always $\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$ which is clearly dependent with $\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}$ which are the last two columns of I_n. Thus spark$(D) = 3$.

c) From theorem 10.5.1 in the class notes, if $\|w\|_0 < \frac{1}{2}\text{spark}(D)$, then w is the unique sparsest solution. So, if we have $\|w\|_0 < \frac{3}{2}$, then w is unique. This can only happen if $\|w\|_0 \in \{0, 1\}$. If $\|w\|_0 = 0$, then $w = 0$ and $y = 0$. Thus, setting $y = 0$ trivially makes w the unique sparsest solution. More interestingly, if $\|w\|_0 = 1$, then y is a multiple of one the columns of D. Thus, a sufficient condition to make w the unique sparsest solution is that

$$y = \alpha D_i$$

where $\alpha \in \mathbb{R}$ and D_i is the i^{th} column of D. I.e., y has to be a scalar multiple of any of the columns of D.