Data Mining Quiz April 15, 2016 *Solutions*

1. What is the input to and output of the Gram-Schmidt algorithm?

   The input to the Gram-Schmidt algorithm is a set of vectors. The output is an orthonormal basis for the space spanned by the input vectors

2. Does compressive sampling lead to an *underdetermined* or *overdetermined* system of equations?

   Compressive sampling is underdetermined. Solutions are not well defined which is why we need to introduce regularization terms such as a sparsity penalty.

3. How is *locally-linear embedding* similar to principal components analysis and how does it differ?

   LLE and PCA are both unsupervised techniques for dimensionality reduction. PCA however is only able to find structures based on linear equations while LLE is able to discover how a dataset may be embedded as a non-linear manifold.

4. Given an n-dimensional orthogonal basis $E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ and a vector $\mathbf{v}$ provide an explicit formula for the expansion coefficients $\mathbf{v} = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots$.

$$c_i = \frac{\langle \mathbf{v}, \mathbf{e_i} \rangle}{\langle \mathbf{e_i}, \mathbf{e_i} \rangle} \tag{1}$$

5. Fig. 1 shows the values of an element of $\mathbb{R}^{20}$ in a stem plot. Answer the following. Would you conclude that this vector is sparse? If we take the first difference of the given vector will the result become more or less sparse?

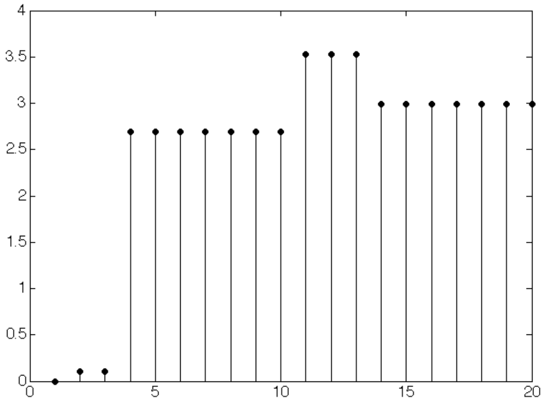

Figure 1: Vector in $\mathbb{R}^{20}$

Generally we should not conclude that this vector is sparse since the majority of elements are non-zero. However, the first difference of the vector is sparse since neighboring elements are only different at 4 places.

6. Let the Euclidean norm of a vector $\mathbf{x}$ in n dimensional space be denoted $||\mathbf{x}|| \equiv \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$. In this case the norm is induced by an inner product, that is, $||\mathbf{x}||^2 = \langle \mathbf{x}, \mathbf{x} \rangle$ Refer to Fig. 2 and answer the following questions.

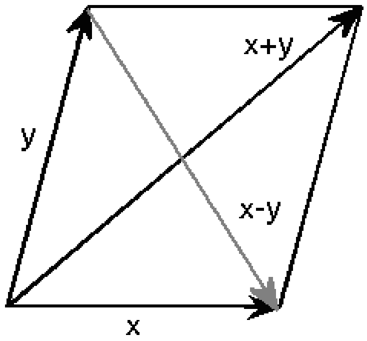

Figure 2: Parallelogram

- What is $\langle \mathbf{x}, \mathbf{y} \rangle$ in terms of components in 2 dimensions?
  $\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_N y_N$
- Express the sum of the squares of the lengths of the four sides of the parallelogram in terms of the norms of $\mathbf{x}$ and $\mathbf{y}$.
  $l = 2||\mathbf{x}||^2 + 2||\mathbf{y}||^2$
- Use elementary properties of norms and inner products to prove that the quantity calculated in the previous problem is equal to $||\mathbf{x} + \mathbf{y}||^2 + ||\mathbf{x} - \mathbf{y}||^2$
  We want to demonstrate the equality $2||\mathbf{x}||^2 + 2||\mathbf{y}||^2 = ||\mathbf{x} + \mathbf{y}||^2 + ||\mathbf{x} - \mathbf{y}||^2$
  First write,
  $||\mathbf{x} + \mathbf{y}||^2 + ||\mathbf{x} - \mathbf{y}||^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle + \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle$
  Then use bilinearity of the inner product.
  $\text{LHS} = 2\langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + 2\langle \mathbf{y}, \mathbf{y} \rangle - 2\langle \mathbf{x}, \mathbf{y} \rangle$
- (*Polarization Identity*) Use similar reasoning to show that $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4}\left(||\mathbf{x} + \mathbf{y}||^2 - ||\mathbf{x} - \mathbf{y}||^2\right)$
  Do the same as above and solve for $\langle \mathbf{x}, \mathbf{y} \rangle$.

7. Do the *rows* of the matrix $\mathbf{B}$ given below form an orthogonal basis for $\mathbb{R}^3$? Show why or why not. Answer the same question for the *columns*.

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \tag{2}$$

No for both. For the second and third rows are not orthogonal. The first and third columns are not orthogonal.

8. Compute $||\mathbf{x}||_0$ and $||\mathbf{x}||_1$ for the vector $\mathbf{x}$.

$$\mathbf{x} = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & -1 & 0 & 1 & 0 & 2 \end{bmatrix}^T \tag{3}$$

$||\mathbf{x}||_0 = 5$ (five nonzero entries) and $||\mathbf{x}||_1 = 6$, the sum of the absolute values of the entries.