Comment on the code:

In addition to the code snippets, please be sure that you're able to run the scripts and/or built-in commands for the algorithms we've discussed in this section (includes interpreting the output from the algorithms).

The algorithms are:

K-means, Neural Gas, DBSCAN, Bisection method, Newton's Method, Multivariate Newton's Method, Gradient Descent and Stochastic Gradient Descent.

II. Questions

1. Suppose the matrix A and the RREF of A are given as below:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \qquad \mathrm{rref}(A) = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix}$$

   (a) Is A full rank? If so, what is its rank?

   SOLUTION: Yes, A is full rank with a rank of 2.

   (b) What is the dimension of the four fundamental subspaces?

   SOLUTION: Using the rank, the dimensions of the row space and column space are both 2. The dimension of the null space is 1. The dimension of the null space of A^T is 0.

   (c) Find a basis for each of the fundamental subspaces (without the SVD).

   SOLUTION:

   For the row space, use the rows from the row reduced matrix:

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} \right\}$$

   For the null space, solve $A\mathbf{x} = 0$. In this case, we already have the RREF(A), so we just need to read off the answer (to check yourself, the vector should be orthogonal to the row space).

$$\left\{ \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \right\}$$

   For the column space, we would typically use the columns from the matrix A. In this case, notice that the column space takes up all of $\mathbb{R}^2$, so you could use any basis for $\mathbb{R}^2$, including the standard basis. However, some of you may mistake that for a rule, so I'll use the columns that we would typically use- The pivot columns from the original matrix.

$$\left\{ \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \end{bmatrix} \right\}$$

   The null space of A^T will then contain only the zero vector: $\{0\}$. This isn't really a basis, though, since the zero vector is not linearly independent (and having dimension zero means that there isn't a basis vector).

2. In words, describe how we determine the rank of a matrix (using the SVD).

   SOLUTION: Theoretically, the rank is determined by the number of non-zero singular values. In practice, we choose the number of singular values that are greater than some small tolerance level (like `1e-10` or something similar).

3. Below is a matrix A. To save you some time, I've also included A^TA, some eigenvalues of A^TA and the matrix AA^T.

$$A = \begin{bmatrix} 2 & -2 \\ 0 & 0 \\ 2 & -2 \end{bmatrix} \qquad A^TA = \begin{bmatrix} 8 & -8 \\ -8 & 8 \end{bmatrix}, (\lambda_{1,2} = 0, 16), \quad AA^T = \begin{bmatrix} 8 & 0 & 8 \\ 0 & 0 & 0 \\ 8 & 0 & 8 \end{bmatrix}$$

(a) Find the full SVD of A, by hand:
SOLUTION: Since we have the eigenvalues, we just need the eigenvectors:
For $\lambda = 16$,

$$(A^TA - \lambda I) = \begin{bmatrix} -8 & -8 \\ -8 & -8 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \Rightarrow \mathbf{v} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Since we're in $\mathbb{R}^2$ and the eigenvectors are orthogonal, the one for $\lambda = 0$ is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Furthermore, we know that $A\mathbf{v}$ is a non-normalized eigenvector for $\lambda = 16$: so

$$\mathbf{u} = A\mathbf{v} = \begin{bmatrix} 2 & -2 \\ 0 & 0 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -4 \\ 0 \\ -4 \end{bmatrix}$$

so we'll take that as $(1\sqrt{2})(-1, 0, -1)$. Lastly, just need the eigenvectors for $\lambda = 0$ in AA^T. Reducing the matrix:

$$\begin{bmatrix} 8 & 0 & 8 \\ 0 & 0 & 0 \\ 8 & 0 & 8 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \Rightarrow \quad \begin{aligned} x_1 &= -x_3 \\ x_2 &= x_2 \\ x_3 &= x_3 \end{aligned} \quad \Rightarrow \quad \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

We have all the information we need now:

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

Note: I recommend finding the $\mathbf{u}$'s by taking $A\mathbf{v}$'s because, while any scalar multiple of an eigenvector is an eigenvector, the signs of $\mathbf{u}$'s and $\mathbf{v}$'s need to match up. For example, if I had taken $\mathbf{u} = (1/\sqrt{2}, 0, 1/\sqrt{2})$, then $U\Sigma V^T$ would be $-A$ instead of A.

(b) Compute the pseudo-inverse of A, by hand:
SOLUTION: You can leave it factored, or multiply it out. Using the **reduced** SVD,

$$V\Sigma^{-1}U^T = \left(\frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right) \left(\frac{1}{4} \right) \left(\frac{1}{\sqrt{2}} [-1, 0, -1] \right) = \frac{1}{8} \begin{bmatrix} 1 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix}$$

Remarks: You might notice that $A^\dagger A$ is not the identity. What does it represent? (Hint: The "solution" to $A\mathbf{x} = \mathbf{b}$ is given by $\hat{\mathbf{x}} = A^\dagger \mathbf{b}$, and notice that $A\hat{\mathbf{x}} = AA^\dagger \mathbf{b}$ and may not equal $\mathbf{b}$).

To specifically answer the question, $A^\dagger A$ is a projection to the row space of A, and $AA^\dagger$ is a projection to the column space of A (these are easier to see if you write $A, A^\dagger$ in terms of U, Σ, V).

4. (Calculator, or by hand) Use two steps of the bisection algorithm on $f(x) = x^2 - 2$ on the interval $[0, 1]$. Be sure you follow the steps.

SOLUTION: **TYPO:** Interval should be $[0, 2]$ since the solution is $\sqrt{2} > 1$.

- Check that $f(0)f(1) < 0$: $f(0) = -2 < 0$, $f(2) = 2 > 0$.
- Let $c = (0 + 2)/2 = 1$. Then $f(1) = -1 < 0$, and the new interval is $[1, 2]$.

- Let $c = (1+2)/2 = 3/2$. Then $f(3/2) = 0.25 > 0$, and the new interval is $[1, 3/2]$.

5. (Calculator, or by hand) Use two steps of Newton's Method on $f(x) = x^2 - 2$ with $x_0 = 1$.

   SOLUTION: Not necessary, but we might simplify the expression first to make the computations easier:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} = x_i - \frac{x_i^2 - 2}{2x_i} = \frac{1}{2}x_i + \frac{1}{x_i}$$

   - $x_1 = \frac{1}{2} + 1 = \frac{3}{2}$
   - $x_2 = \frac{1}{2}\frac{3}{2} + \frac{2}{3} = \frac{17}{12}$

   Side remark: Note that $17/12 \approx 1.4167$ and $\sqrt{2} \approx 1.4142$, so we're already very close.

6. Given vector $\mathbf{a} = [1, 1, 1]^T$, find the projection matrix P so that $P\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto $\mathbf{a}$.

   SOLUTION: Let $U = \frac{1}{\sqrt{3}}\mathbf{a}$. Then the projection is UU^T, or

$$\frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

7. Show that Null$(A) \perp$ Row(A).

   SOLUTION: This is really all based on how we compute $A\mathbf{x}$. In performing this multiplication, the end result is the dot product between each row of A and the vector $\mathbf{x}$. Therefore, if $A\mathbf{x} = \mathbf{0}$ (so that $\mathbf{x} \in$ Null(A)), then $\mathbf{x}$ must be orthogonal to every row of A, and so it will be orthogonal to any linear combination of the rows of A. This argument also goes in reverse: If $\mathbf{x}$ is orthogonal to every row of A, then $A\mathbf{x} = \mathbf{0}$, so that $\mathbf{x}$ must be in the null space.

8. Let

$$U = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

   Find $[\mathbf{x}]_U$. Find the projection of $\mathbf{x}$ into the subspace spanned by the columns of U. Find the distance between $\mathbf{x}$ and the subspace spanned by the columns of U.

   SOLUTION: We have three parts-

   - **TYPO:** Did you notice that $\mathbf{x}$ is not contained in the span of the first two columns? Therefore, the coordinates are not defined. HOWEVER, we can find the coordinates of the **projection** of $\mathbf{x}$ to the column space spanned by the first two columns of U, and in that case:

     The coordinates of $\mathbf{x}$ with respect to U is given by $U^T\mathbf{x} = (1/\sqrt{2})(4, 2)$
   - To find the projection, we could either compute $UU^T\mathbf{x}$, or since we already have $U^T\mathbf{x}$, use the previous problem to get:

$$\text{Proj}(\mathbf{x}) = 2\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}$$

     And notice that that is the projection of $\mathbf{x}$ to the $xy-$plane.

- The distance between $\mathbf{x}$ and the projection of $\mathbf{x}$ is similar to our error function in the best basis (that's why we compute it here). It's easy in this case:

$$\|\mathbf{x} - \text{Proj}(\mathbf{x})\| = \| \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \| = 2$$

9. True or False, and give a short reason:

   (a) If the rank of A is 3, the dimension of the row space is 3.

   SOLUTION: TRUE. The rank is the dimension of the column space, which is the number of pivot columns of a matrix. That is the same as the number of pivot rows of A, which is the dimension of the row space.

   (b) If the correlation coefficient between two sets of data is 1, then the data sets are the same.

   SOLUTION: FALSE, however there is a "positive" linear relationship between the two (positive meaning positive slope).

   (c) If the correlation coefficient between two sets of data is 0, then there is no functional relationship between the two sets of data.

   SOLUTION: FALSE. To make the statement true, there is **linear** relationship between the two sets of data.

   (d) If U is a 4×2 matrix, then $U^T U = I$.

   SOLUTION: False in general, but if U has orthonormal columns, it is true (I meant it that way).

   (e) If U is a 4×2 matrix, then $UU^T = I$.

   SOLUTION: False in general and if U has o.n. columns. In the later case, UU^T is a projection matrix to the column space of U.

   (f) If A is not invertible, then $\lambda = 0$ is an eigenvalue of A.

   SOLUTION: Assuming A is square, then yes (otherwise, false). That's because $\det(A) = 0$ means that $\det(A - 0I) = 0$.

   (g) Let

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 2 & 0 \end{bmatrix}$$

   Then the rank of AA^T is 2.

   SOLUTION: True. The reason is that the rank of A is 2 (the columns are not multiples of each other), and the rank of A is the same as the rank of AA^T. (Think about the number of singular values in A versus the number of non-zero eigenvalues of AA^T).

10. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be the normalized eigenvectors of $A^T A$, where A is $m \times n$.

   (a) Show that if λ_i is a non-zero eigenvalue of $A^T A$, then it is also a non-zero eigenvalue of AA^T.

   SOLUTION: Let $\mathbf{v}$ be an eigenvector associated with λ_i. Then:

$$A^T A\mathbf{v} = \lambda_i \mathbf{v} \quad \Rightarrow \quad AA^T A\mathbf{v} = \lambda_i A\mathbf{v} \quad \Rightarrow \quad AA^T \mathbf{u} = \lambda_i \mathbf{u}$$

   (b) True or false? The eigenvectors form an orthogonal basis of $\mathbb{R}^n$.

   SOLLUTION: True by the Spectral Theorem; they can be selected so that they form an orthonormal basis of $\mathbb{R}^n$.

(c) Show that, if $\mathbf{x} \in \mathbb{R}^n$, then the i^{th} coordinate of $\mathbf{x}$ (with respect to the eigenvector basis) is $\mathbf{x}^T\mathbf{v}_i$.
SOLUTION: First, let

$$\mathbf{x} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$$

and now dot both sides of the equation with $\mathbf{v}_i$. On the right, most terms will be zero because the vectors $\mathbf{v}$ are orthogonal to each other. This gives us:

$$\mathbf{x} \cdot \mathbf{v}_i = 0 + 0 + \cdots + c_i\mathbf{v}_i \cdot \mathbf{v}_i + 0 + \cdots + 0$$

Therefore, the i^{th} coordinate of $\mathbf{x}$ is given by

$$c_i = \frac{\mathbf{x} \cdot \mathbf{v}_i}{\mathbf{v}_i \cdot \mathbf{v}_i}$$

(d) Let $\alpha_1, \ldots, \alpha_n$ be the coordinates of $\mathbf{x}$ with respect to $\mathbf{v}_1, \ldots, \mathbf{v}_n$.
Show that (**TYPO:** The subscript 2 should be an exponent, corrected below)

$$\|\mathbf{x}\|^2 = \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_n^2$$

I'll allow you to show it just using just two vectors, $\mathbf{v}_1, \mathbf{v}_2$.
SOLUTION: Using the pair,

$$\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x} = (\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2) \cdot (\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2) = \alpha_1^2\|\mathbf{v}_1\|^2 + \alpha_2^2\|\mathbf{v}_2\|^2 + 2\alpha_1\alpha_2\mathbf{v}_1 \cdot \mathbf{v}_2$$

By orthogonality of the vectors, and assuming that $\|\mathbf{v}_i\|^2 = 1$, we get the anser:

$$\|\mathbf{x}\|^2 = \alpha_1^2 + \alpha_2^2$$

(e) Show that $A\mathbf{v}_i \perp A\mathbf{v}_j$
SOLUTION: We use the fact that the $\mathbf{v}$'s are eigenvectors of A^TA. Now:

$$(A\mathbf{v}_i)^T(A\mathbf{v}_j) = \mathbf{v}_i^T A^T A\mathbf{v}_j = \lambda_j \mathbf{v}_i^T\mathbf{v}_j = 0$$

by the orthogonality of $\mathbf{v}$'s.

(f) Show that $A\mathbf{v}_i$ is an eigenvector of AA^T.
This is shown in part(a).

11. (SVD) Given that the SVD of a matrix was given in Matlab as:

```
>> [U,S,V]=svd(A)
U =
   -0.4346   -0.3010    0.7745    0.3326   -0.1000
   -0.1933   -0.3934    0.1103   -0.8886   -0.0777
    0.5484    0.5071    0.6045   -0.2605   -0.0944
    0.6715   -0.6841    0.0061    0.1770   -0.2231
    0.1488   -0.1720    0.1502   -0.0217    0.9619
S =
    5.72         0           0
         0       2.89         0
         0          0         0
         0          0         0
         0          0         0
V =
    0.2321   -0.9483    0.2166
   -0.2770    0.1490    0.9493
    0.9324    0.2803    0.2281
```

(a) Which columns form a basis for the null space of A? For the column space of A? For the row space of A?

SOLUTION: First, the size of A and the size of S (or Σ) are the same in the full SVD. Therefore, we know that the row space and null space are vectors in $\mathbb{R}^3$ (and the column space and null space of A^T are in $\mathbb{R}^5$). Since we have two non-zero singular values, the rank of A is 2, and therefore, the third column of V gives a basis for the null space of A.

The first 2 columns of U form a basis for the column space of A, and the first two columns of V form a basis for the row space.

(b) How do we "normalize" the singular values? In this case, what are they (numerically)?

SOLUTION: We normalized singular values (or eigenvalues) by dividing by their sum. In this case, 8.61. This gives:

$$\frac{5.72}{8.61} \approx 0.66 \qquad \frac{2.89}{8.61} \approx 0.34$$

(c) What is the rank of A?

SOLUTION: It is the number of non-zero singular values of A, which we determined in part (a) to be 2.

(d) How would you compute the pseudo-inverse of A (do not actually do it):

SOLUTION: In Matlab notation, `V(:,1:2) * (diag(1./diag(S))) * U(:,1:2)'`

(e) Let B be formed using the first two columns of U. Would the matrix B^TB have any special meaning? Would BB^T?

SOLUTION: B^TB would be a 2×2 identity matrix. BB^T would be a projection matrix (to the column space of U).

12. Define a "voronoi cell" and its relation to data clustering.

SOLUTION: A voronoi cell is defined by its centers, $\mathbf{c}_1, \ldots, \mathbf{c}_k$. Then the j^{th} voronoi cell is the set of $\mathbf{x}$ that is closer to $\mathbf{c}_j$ than any other center- Or,

$$V_j = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{c}_j\| \leq \mathbf{c}_i, \text{ for } i = 1, 2, \ldots, k\}$$

Points that lie along the boundary may be left unclassified, or randomly assigned to bordering cells.

13. What is the basic update rule we use for all our parameters? Hint: We want to go from α_{initial} to α_{final} in some number (MaxIters) of steps.

SOLUTION: This actually goes back a ways to the $n-$armed bandit. We said that at step i:

$$\alpha_i = \alpha_{\text{init}} \left(\frac{\alpha_{\text{final}}}{\alpha_{\text{init}}} \right)^{\frac{i}{\text{MaxIters}}}$$

14. Explain the roles that ϵ and λ play in the Neural Gas algorithm.

SOLUTION: We said that ϵ was the maximum amount of "attracting" force, and λ controlled the spread of the attracting force. Thus, at the beginning of training, ϵ and λ are relatively large, and decrease as training progresses.

15. Show that, for all numbers μ, the value that minimizes the (squared) distortion error for a single cluster is the (arithmetic) mean. You may assume your data is one dimensional, and that you have only one cluster.

SOLUTION: If our one dimensional data is given as $x_1, x_2, \ldots, x_p$, then the sum of squares distortion error is

$$E(\mu) = \sum_{k=1}^{p} (x_k - \mu)^2$$

To minimize E, differentiate and set the derivative to zero (find the critical points):

$$\frac{dE}{d\mu} = \sum_{k=1}^{p} 2(x_k - \mu)(-1) = 0 \quad \Rightarrow \quad \sum_{k=1}^{p} x_k - \mu \sum_{k=1}^{p} 1 = 0 \quad \Rightarrow \quad \sum_{k=1}^{p} x_k = \mu p$$

Therefore, the critical point is when

$$\mu = \frac{1}{p} \sum_{k=1}^{p} x_k$$

which is the arithmetic average. Further, if we take the second derivative,

$$\frac{d^2 E}{d\mu^2} = 2p > 0$$

Therefore, we have a minimum and not a maximum.

16. **TYPO: Should have specified this is k-means.** Here are 5 points in the matrix X. Initialize the two centers as the first two columns of X, then perform 1 update, and show there is a decrease in the distortion error.

$$X = \begin{bmatrix} -1 & 1 & 1 & -2 & -1 \\ 1 & 0 & 2 & 1 & -1 \end{bmatrix}$$

SOLUTION: As a computational note, it is easier to find the squared distances, and the order will remain the same. The EDM of squared distances is

$$\begin{bmatrix} 0 & 5 \\ 5 & 0 \\ 5 & 4 \\ 1 & 10 \\ 4 & 5 \end{bmatrix} \quad \Rightarrow \quad \begin{array}{l} \text{Cluster 1: } 1,4,5 \\ \text{Cluster 2: } 2,3 \end{array} \quad \Rightarrow \quad C = \begin{bmatrix} -4/3 & 1 \\ 1/2 & 1 \end{bmatrix}$$

It takes a little while to compute, but the new EDM shows that the classifications do not change, and the new distortion errors (squared) are approximately:

$$0.55, 1.0, 1.0, 0.88, 1.88$$

We can see that the overall distortion error has decreased.

17. Given the data vector $\mathbf{x}$ below and the three centers in C, update the set of centers using Neural Gas, with $\epsilon = \lambda = 1$ (not realistic, but since we're doing it by hand, we'll use easy numbers).

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad C = \begin{bmatrix} -1 & 1 & 2 \\ 1 & 0 & 3 \end{bmatrix}$$

SOLUTION: First we need the distances between $\mathbf{x}$ and the three centers. In order, we have: $\sqrt{5}, 2, \sqrt{2}$, therefore, the third center is closest, and in the notation employed by our text, we have

$$s_3 = 0 \quad s_2 = 1 \quad s_1 = 2$$

Now update the centers by the index:

$$\mathbf{c}_3 = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + 1 \begin{bmatrix} 1-2 \\ 2-3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\mathbf{c}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \mathrm{e}^{-1} \begin{bmatrix} 1-1 \\ 2-0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2/e \end{bmatrix}$$

$$\mathbf{c}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \mathrm{e}^{-2} \begin{bmatrix} 1 - -1 \\ 2-1 \end{bmatrix} = \begin{bmatrix} -1 + 2/e^2 \\ 1 + 1/e^2 \end{bmatrix}$$

18. In the DBSCAN algorithm, is there a difference between *indirectly density-reachable* and *density-reachable*?

    These topics are important because they tell us how DBSCAN creates clusters: "A cluster is the set of all points that are density-reachable from a (arbitrary) core point p".

    SOLUTION: No, although there is a difference between *directly* density-reachable: Point q is directly density-reachable from p if p is a core point and q is within the ϵ neighborhood of p. If we drop the word "directly", then we can have a chain of intermediate points taking us from one point to the other.

19. Give a summary of the DBSCAN algorithm.

    SOLUTION: We just start with a random point in the set:

    - Has it been classified? If not, then check if it is a core point.
    - If the point is a core point, we now have a new cluster- Collect all the points density-reachable to the point.

      Otherwise, classify the point as noise.

20. Describe a situation where DBSCAN would work very well.

    SOLUTION: We saw a couple of great examples in the text and homework- One of them was the interlocking rectangles with some scattered point (pg 97). In the homework, we had that data set with interlocking horsehoes. These sets have complex shapes.

21. Describe a situation where DBSCAN would work very poorly.

    SOLUTION: There's an example in the text (p. 97), but also when the data has a single overall shape- like the clustering we found using Neural Gas. In fact, it wouldn't work at all if you required some kind of cluster centers to be produced from the algorithm (like if you were reducing the number of points in your set).

22. Illustrate the technique of gradient descent using

$$f(x, y) = x^2 + y^2 - 3xy + 2$$

    (a) Find the critical point.

    SOLUTION: Set the gradient equal to zero. In this case, we have only $(0, 0)$.

    (b) Use the initial point $(1, 0)$ and find the optimal step size, then compute the next point.

    SOLUTION: The gradient at $(1, 0)$ is $\langle 2, -3 \rangle$, so the line we're using is

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} - h \begin{bmatrix} 2 \\ -3 \end{bmatrix} = \begin{bmatrix} 1 - 2h \\ 3h \end{bmatrix}$$

    Substituting $x = 1 - 2h$, $y = 3h$ into $f(x, y)$ (this means we're restricting f to the line), we get

$$\phi(h) = 13h^2 - 13h + 3 \quad \Rightarrow \quad \phi'(h) = 0 \quad \Rightarrow \quad h \approx 0.21$$

    That puts our new point at:

$$\begin{bmatrix} 1 - 2h \\ 3h \end{bmatrix} \approx \begin{bmatrix} 0.58 \\ 0.62 \end{bmatrix}$$

    (c) Classify the critical point by considering the eigenvalues of the Hessian (compute the Hessian and the eigenvalues).

    The Hessian in Matlab is `[2,-3;-3,2]` , and it has eigenvalues $5, -1$. Because they are mixed in sign, the point $(1, 0)$ is a saddle point.

23. What's the difference between *gradient descent* and *stochastic gradient descent*? (Be specific).

    We'll assume that the error function has a large sum (like sum of squares error), and typically that sum is taking place over the points you have. Then the gradient will involve the full sum, but in stochastic gradient descent, we will be **estimating** the gradient using a single point.

24. Consider the table of data below, where we want to find a line of best fit.

x	-1	1	2	3
y	0	1	3	2

(a) Write down the full error function (that depends on m, b).
SOLUTION:

$$E(m,b) = (0 - (m(-1) + b))^2 + (1 - (m + b))^2 + (3 - (2m + b))^2 + (2 - (3m + b))^2$$

(b) Write down the full gradient.
SOLUTION:

$$\frac{\partial E}{\partial m} = 30m + 10b - 26, \qquad \frac{\partial E}{\partial b} = 10m + 8b - 12$$

(c) Describe how we implement stochastic gradient in this problem.
SOLUTION: We would only use a single point as an estimate. For example, using point 2, the error estimate would be $(1 - (m + b))^2$, so then the gradient is $E_m = 2(1 - m - b)(-1)$ and $E_b = 2(1 - m - b)(-1)$.