Review Questions, part 1

1. (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.

   TYPO: Make the interval $[0, 2]$ (Sorry about that!)

   SOLUTION:

   - $fa = f(0) = -2$ and $fb = f(2) = 4 - 2 = 2$. We see that fa and fb are opposite in sign.
   - Let $c = (0 + 2)/2 = 1$. Now $fc = -1$, so the new interval is $[1, 2]$
   - Reset $a = 1$ and $fa = fc$, and we repeat:
   - Let $c = (1 + 2)/2 = 3/2$, and $fc = 0.25$, so the new interval is $[1, 1.5]$
   - We output the middle point now that we've performed two steps: $x = 1.25$

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

   SOLUTION:

   - $f(1) = -1$ and $f'(1) = 2$.
   - New $x = 1 - (-1/2) = 3/2$
   - $f(3/2) = 0.25, f'(3/2) = 3$.
   - New $x = 1.5 - 0.0833 \approx 1.416$ (We're very close to $\sqrt{2}$ already!

3. Given that the SVD of a data matrix X was given in Matlab as:

```
>> [U,S,V]=svd(X)
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
```

   SOLUTIONS:

   (a) What was the dimensions of the matrix X?
   The matrix would be the same size as S, or 5×3.
   What is its rank?
   Rank is 2.

(b) The data in X actually lies on a plane. How do I know that, and what is a basis for the plane (will you only have one possible answer?)

A "plane" through the origin is a two dimensional subspace, and the rank of our data matrix is 2. NOTE: We don't know if the original data is in $\mathbb{R}^5$ or $\mathbb{R}^3$- If the data is in $\mathbb{R}^5$, the plane is spanned by the first two columns of U. If the data is in $\mathbb{R}^3$, then the first two columns of V will span the plane.

(c) Assuming the "points" in X have been mean subtracted, what is the best basis?

See the answer above (the best basis is given by that plane)

(d) Out of all possible non-zero vectors in $\mathbb{R}^3$, which vector will maximize the variance of the data in X, projected to that vector?

The first vector in V.

(e) (Continuing from the last question) In that case, what is the variance of the projected data?

The largest eigenvalue of the covariance matrix. In this case, $\sigma_1^2/(p-1)$, so that would be $5.72^2/4 \approx 8.719$

4. Given a basis vector ϕ and a set of data $\mathbf{x}_1, \ldots, \mathbf{x}_p$, how would you go about computing the "error" is using ϕ as a basis for the data (you may assume the data is mean-subtracted). Be specific about what computations you have to make.

SOLUTION: We sum the magnitude (squared) of the sum of the error vectors. Recall that the error vector is the difference between the original $\mathbf{x}$ and the projected vector, so that we get

$$\sum_{n=1}^{p} \|\mathbf{x}_i - \phi\phi^T \mathbf{x}_i\|^2$$

5. Suppose that $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_k$ (assume these are fixed), and that $p_1, p_2, \ldots, p_k$ are k numbers so that $p_i \geq 0$ for each i, and the sum of the k numbers is 1. How would we choose the number $p_1, \ldots, p_k$ so that the sum below is maximum?

$$\lambda_1 p_1 + \lambda_2 p_2 + \cdots + \lambda_k p_k$$

What is the value of the maximum? Similarly, how would the p_i be chosen if you wanted to minimize the sum above? What is the value of that minimum?

SOLUTION: These came from our discussion of best basis- The max value is λ_1, when $p_1 = 1$ and all other p's are 0. The minimum is λ_k, and that happens when $p_k = 1$ and the rest of the p's are 0.

6. How should we determine what number of basis vectors to use (that is, what is the value of k) when computing a best basis?

SOLUTION: Consider the rank of the matrix. If there is a clear drop between $\lambda_k > 0$ and the remaining singular values of the matrix are 0, then the rank is k. Otherwise, we could use a "tolerance", and take k so that

$$\sum_{n=1}^{k} \frac{\lambda_n}{\sum \lambda_i} > tol$$

(but summing the first $k-1$ normalized eigenvalues of the covariance matrix is not above the tolerance).

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

Given a set of points called centers, the Voronoi diagram is formed by partioning the plane into regions about each center. The region (or cluster) for a given center is the set of points in the plane that are closer to that center than any other (therefore, the exact diagram also depends on the metric being used).

8. 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.

   $\alpha(t+1) = \alpha_{\text{initial}} \left(\frac{\alpha_{\text{final}}}{\alpha_{\text{initial}}} \right)^{t/\text{MaxIters}}$

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

   SOLUTION: Given a "hump" function (or normal curve, or Gaussian function), ϵ is the "learning rate". which controls the height of the hump, and λ is the how wide-spread the change will be- If λ is large, we move many centers towards the data point chosen, if λ is small, we may even isolate a single center.

10. 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: This could have been phrased better- See Problem 10 in the second set of questions.

11. 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: **TYPO: We should use k-means (that was left out)**

    LUTION: 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.

12. 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:

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

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

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

13. 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: 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.

14. 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.

    NOTE: The review questions are a draft, and will be finished this weekend. This is an "early preview".