Review Topics, Exam 2

For the exam, no calculators will be allowed, but on this review, you may use a calculator for basic arithmetic operations. You may prepare a one-page (8.5 × 11 in) page of notes, written on one side.

1. The Best Basis and Eigenfaces

   - Be able to define what is meant by "best". That is, given an arbitrary orthonormal bases $\phi_1, \ldots, \phi_n$, the error vector for one data point using k vectors is given by:

     $$\mathbf{x} = (c_1\phi_1 + \cdots c_k\phi_k) + (c_{k+1}\phi_{k+1} + \cdots c_n\phi_n) = \Phi\Phi^T\mathbf{x} + \mathbf{x}_{\text{err}}$$

     The overall error (over p points) is then the sum of the magnitude of each error squared:

     $$\text{Err} = \sum_{i=1}^{p} \|\mathbf{x}_{\text{err}}^{(i)}\|^2$$

     In words, this is the sum of the magnitude squared of the reconstruction error.
   - Define the covariance matrix, and how to compute it by hand.
   - Show that the projection of the mean is the mean of the projection.
   - Show that the variance of the (mean-subtracted) data projected to a vector $\mathbf{u}$ is given by $\mathbf{u}^T C\mathbf{u}$.
   - Given that the best one-dimensional basis ϕ is found by maximizing $\phi^T C\phi$, we can view this as finding vector ϕ that maximizes the variance of the projection of the data (why?)
   - Give an argument that the best k eigenvectors of the covariance matrix C can be obtained by using the first k eigenvectors in the SVD (either in U or V).
   - (Matlab) Be able to compute the best set of k basis vectors given a set of data, and be able to plot the best 2-dimensional represention.
   - (Matlab) Give the $k-$dimensional reconstruction of the data and be able to compute the error (in Matlab).

2. Clustering

   - What is the clustering problem? Is it an example of "supervised" or "unsupervised" problem?
   - What are Voronoi cells?
   - What is the "Euclidean distance matrix" (EDM)?
   - For each algorithm: $K-$means, SOM, Neural Gas, DBSCAN.
     - State what inputs you have to provide to the algorithm, then what comes out of the algorithm.
     - Understand the update rule on the cluster centers (except for DBSCAN, since it doesn't have centers).
     - For DBSCAN, how does the algorithm produce a single "cluster"? (I'm looking for a short verbal description).
     - For SOM and Neural Gas, we also have connections between cluster centers. How are these constructed?
     - For DBSCAN, be able to define core, border and noise points.
     - For DBSCAN, understand the difference between points being "directly density-reachable" versus "density-reachable".
     - (Matlab) Be able to run these algorithms on a given data set and tweak the parameters to get a reasonably good output.

- Understand the differences in what the algorithms produce, and when you might use one over another- For example, why would we not run DBSCAN on the obstacle course data? (Because it would just cluster every point in a single cluster and would not tell us a path to take to avoid the obstacles).

3. Chapter on optimizqation
   - Be able to "linearize" a function. The most general case is for a function G mapping $\mathbb{R}^n \to \mathbb{R}^m$.
   - Root-finding algorithms:
     - Why do we include root-finding algorithms in optimization?
     - For bisection and Newton's method (one dimensional case):
       * What do you have to provide the algorithm, and what comes out of it?
       * Explain the method geometrically and be able to summarize the algorithm.
       * What are some "pros" and "cons" of using the algorithm?
     - Using the one dimensional case as a reference, be able to write down the multivariate version of Newton's method (there are two- one for a multivariate function G and one for the gradient of a one-dimensional function f).
     - Hybrid method: Why might one use both bisection and Newton's method?
   - Optimization algorithms: Gradient descent and stochastic gradient descent (SGD).

     Be able to explain the methods and the algorithms. Understand the role of the parameters in both algorithms and discuss ways of dealing with it.

Review Questions (written part of exam)

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

   (a) What was the dimensions of the matrix X? What is its rank?

(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?)

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

(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?

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

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.

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?

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?

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

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.

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

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.

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.

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

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).

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

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".

14. Give a summary of the DBSCAN algorithm.

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