Review Topics, Exam 2

1. Linear Algebra (4 Sessions)

   Four Fundamental Subspaces; Express $\mathbf{x}$ in terms of its coordinates. Be able to compute the coordinates of $\mathbf{x}$ with respect to a given basis (orthogonal and non-orthogonal); Eigenvalues/Eigenvectors (definition, how to compute in general); The Spectral Theorem; The Singular Value Decomposition (SVD). (Definition, how to compute the matrices, relationship of the matrices to the four fundamental subspaces); Generalized Inverse (Moore-Penrose Inverse). Be able to construct it from the SVD.

2. Statistics (2 Sessions)

   Definitions of the sample mean, variance, standard deviation, correlation coefficient, covariance matrix. Be able to compute these. Look at the relationships between mean and variance of data vs data projected to a vector. Look at the relationships between mean, covariance and correlation coefficient of a data set $\mathbf{x}$ and $\mathbf{y} = m\mathbf{x} + b$.

3. The Best Basis and Eigenfaces (2 Sessions)

   The best basis produces the basis that does what the best (what is being optimized)? Be able to compute the best set k basis vectors given a set of data, and be able to plot the best 2-dimensional represention. Give the $k-$dimensional reconstruction of the data and be able to compute the error.

4. Clustering (4 Sessions)

   Algorithms: $K-$means, Neural Gas, DBSCAN. 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).

5. Optimization (3 Sessions)

   Algorithms: Bisection, Newton's Method (multivariate Newton's method), Gradient Descent, Stochastic Gradient Descent. Be familiar enough to write down the algorithm (pseudo-code or in words and formulas). Be able to apply any of the algorithms to a particular function. Understand the role of the parameter in Gradient Descent and Stochastic Gradient Descent, and discuss ways of dealing with it.

I. Code Snippets

Collect the code snippet that would allow you to do each of the following tasks. The first one is done for you so you understand what I mean.

1. Let U be a matrix with orthonormal columns, and X be some data in the same space as the columns of U. If we have p data points in $\mathbb{R}^n$, how do I compute the coordinates of the data in X with respect to the first three columns of U (and give any relevant dimensions).

   SOLUTION: The coordinates are given by the expression below, where U has at least 3 columns, and X is $n \times p$, so that the coordinates matrix is $3 \times p$. (You can choose either Matlab or Python; you don't need to know both).

Matlab	Python
`Coords = U(:,1:3)'*X;`	`Coords = U[:,:3].T @ X`

2. Continuing with the first snippet, plot the first two coordinates of matrix `Coords` in the plane (single points without line segments connecting them).

3. Plot $x^2 - 3x$ using 500 points evenly spaced between $x = -1$ and $x = 4$.

4. Compute the variance of data in a vector $\mathbf{x}$, possibly varying in length (without using the built-in function `var`).

5. Compute the covariance of data in vectors $\mathbf{x}$ and $\mathbf{y}$, of the same, but possibly varying, length (without using the built-in function `cov`).

6. Find the rank 4 representation of the data in matrix X using the columns of U. (Your answer should be vectors back in $\mathbb{R}^n$, and not in $\mathbb{R}^4$).

7. Take matrix X, and mean subtract it, where the mean should be a column vector.

8. Take a matrix X and mean subtract it, where the mean is a row vector.

9. Find the SVD of a matrix X (assume mean subtracted).

10. If U, S, V came out of the SVD for matrix A, and we've determined the rank to be 3, how would we construct the pseudo-inverse of A?

11. Be able to load a matrix X from a Matlab *.mat file (in Matlab or Python; in Python there are a few lines of code to use).

II. Questions

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

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

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

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

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

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

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:

   (b) Compute the pseudo-inverse of A, by hand:

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.

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

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 $\mathbf{a}$.

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

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.

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.

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

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

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

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

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

   (g) Let

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

   Then the rank of AA^T is 2.

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.

   (b) True or false? The eigenvectors form an orthogonal 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 $\mathbf{x}^T \mathbf{v}_i$.

   (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

$$\|\mathbf{x}\|_2 = \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_n^2$$

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

   (e) Show that $A\mathbf{v}_i \perp A\mathbf{v}_j$

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

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?

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

(c) What is the rank of A?

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

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

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

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.

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

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.

16. 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.

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

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).

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

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".

19. Give a summary of the DBSCAN algorithm.

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

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

22. Illustrate the technique of gradient descent using

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

(a) Find the critical point.

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

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

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

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).

(b) Write down the full gradient.

(c) Describe how we implement stochastic gradient in this problem.