Review Questions (Exam 2)

1. 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,

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

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

2. Define the "confusion matrix", and how it is used.

   SOLUTION: A confusion matrix is used to assess how well a classification has performed. Normally, the actual classes are listed along the top (columns), and the predicted classes are down the rows. For example, $C(i, j)$ would be the percent something in class j was classified as being in class i.

   The confusion matrix not only tells us how well we did (diagonal elements), but also tells us what sorts of errors occurred when the classification was incorrect.

3. 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:

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

4. 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.

5. 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,

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

Therefore, we have a minimum and not a maximum.

6. 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.

7. 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:

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

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

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

8. Show that, for the line of best fit, the normal equations produce the same equations as minimizing an appropriate error function. To be more specific, set the data as $(x_1, t_1), \ldots, (x_p, t_p)$ and define the error function first. Minimize the error function to find the system of equations in m, b. Show this system is the same you get using the normal equations.

   SOLUTION: Done as a homework problem. The model equations are

$$\begin{array}{rl} mx_1 + b & = t_1 \\ mx_2 + b & = t_2 \\ \vdots & \vdots \\ mx_p + b & = t_p \end{array} \quad \Rightarrow \quad \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_p & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{bmatrix} \quad \Rightarrow \quad A\mathbf{c} = \mathbf{t}$$

   Using the normal equations, we get the following for A^TA and $A^T\mathbf{t}$:

$$A^TA = \begin{bmatrix} x_1 & x_2 & \ldots x_p & \\ 1 & 1 & \ldots & 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_p & 1 \end{bmatrix} = \begin{bmatrix} \sum x_k^2 & \sum x_k \\ \sum x_k & p \end{bmatrix}$$

$$A^T\mathbf{t} = \begin{bmatrix} x_1 & x_2 & \ldots x_p & \\ 1 & 1 & \ldots & 1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{bmatrix} = \begin{bmatrix} \sum x_k t_k \\ \sum t_k \end{bmatrix}$$

   where each sum ranges from $k = 1$ to $k = p$.

   On the other hand, as a Calculus question, we are minimizing the sum of the squared error:

$$E(m, b) = \sum_{k=1}^{p} (t_i - (mx_i + b))^2$$

   Set the partial derivatives to zero:

$$\frac{\partial E}{\partial m} = \sum_{k=1}^{p} 2(t_i - (mx_i + b))(-x_i) = 0 \quad \Rightarrow \quad -\sum_{k=1}^{p} x_i t_i + m \sum_{k=1}^{p} x_i^2 + b \sum_{k=1}^{p} x_i = 0$$

And

$$\frac{\partial E}{\partial b} = \sum_{k=1}^{p} 2(t_i - (mx_i + b))(-1) = 0 \quad \Rightarrow \quad -\sum_{k=1}^{p} t_i + m\sum_{k=1}^{p} x_i + b\sum_{k=1}^{p} 1 = 0$$

Putting the two equations together, we have the system:

$$\begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & p \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \sum x_i t_i \\ \sum t_i \end{bmatrix}$$

which is the same as we had before.

9. Given data:

x	-1	0	1
y	2	1	1

   (a) Give the matrix equation for the *line of best fit*.
   SOLUTION:

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

   (b) Compute the normal equations.
   SOLUTION:

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$$

   (c) Solve the normal equations for the slope and intercept.
   SOLUTION: This is easy enough to solve directly- $m = -1/2$ and $b = 4/3$.

10. Use the data in Exercise (9) to find the parabola of best fit: $y = ax^2 + bx + c$. (NOTE: Will you only get a least squares solution, or an actual solution to the appropriate matrix equation?)

    SOLUTION: This will be a unique solution (invertible matrix), which we can find by row reduction:

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \quad \Rightarrow \quad \left[\begin{array}{ccc|c} 1 & -1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1/2 \\ 0 & 1 & 0 & -1/2 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

11. What is Hebb's rule (the biological version- you can paraphrase)?

    In words, Hebb's rule said:

    > When an axon of cell A is near enough to excite a cell B and repeatedly takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

In notation, we said that

$$\Delta W_{ij} = \alpha y_i x_j$$

12. What is the Widrow-Hoff learning rule? How is it related to Hebb's rule?

$$\Delta W_{ij} = \alpha(t_i - y_i)x_j$$

It is an improvement over Hebb's rule to incorporate the target information.

13. Let $W = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. If $\mathbf{x} = [-1, 0, 1]^T$ and $\mathbf{t} = [2, 3]^T$, use Widrow-Hoff to update $W, \mathbf{b}$ one time using a learning rate of 1 (This is too big of a learning rate to actually use, but it will make your computations easier).

SOLUTION: First we need $\mathbf{y} = W\mathbf{x} + \mathbf{b} = [1, 1]^T$. Now,

$$W = W + \Delta W = W + \alpha(\mathbf{t} - \mathbf{y})\mathbf{x}^T$$

$$W = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} [-1, 0, 1] = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 2 \\ -3 & 1 & 2 \end{bmatrix}$$

And

$$\mathbf{b} = \mathbf{b} + \alpha(\mathbf{t} - \mathbf{y}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

14. Let $\mathbf{x} = [1, 2, 1]^T$. Find the matrix $\mathbf{x}\mathbf{x}^T$, its eigenvalues, and eigenvectors. (Also, think about what happens in the general case, where a matrix is defined by $\mathbf{x}\mathbf{x}^T$).

SOLUTIONS:

$$\mathbf{x}\mathbf{x}^T = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

In computing the determinant, it should simplify to $\lambda^2(\lambda - 6) = 0$, so $\lambda = 0$ (double root), and $\lambda = 6$.

For $\lambda = 0$, we get the null space of A. After row reduction, we have:

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \Rightarrow \quad E_0 = \text{span}\left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\}$$

If $\lambda = 6$, after row reduction we get:

$$\left[\begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \Rightarrow \quad E_6 = \text{span}\left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \right\}$$

(You might notice that all eigenvectors are orthogonal, as predicted by the Spectral Theorem)

15. Suppose $\mathbf{x}$ is a vector containing n real numbers, and we understand that $m\mathbf{x} + b$ is Matlab-style notation (so we can add a vector to a scalar, done component-wise).

(a) Find the mean of $\mathbf{y} = m\mathbf{x} + b$ in terms of the mean of $\mathbf{x}$.
SOLUTION:
In our notation, the mean of the data in the vector $\mathbf{x}$ is: $\bar{x} = \frac{1}{p}\sum_{k=1}^{p} x_k$
The mean of the data in vector $\mathbf{y}$ would then be:

$$\bar{y} = \frac{1}{p}\sum_{k=1}^{p} y_k = \frac{1}{p}\sum_{k=1}^{p}(mx_i + b) = m\frac{1}{p}\sum_{k=1}^{p} x_k + \frac{1}{p}bp = m\bar{x} + b$$

(b) Show that, for fixed constants a, b, $\mathrm{Cov}(\mathbf{x} + a, \mathbf{y} + b) = \mathrm{Cov}(\mathbf{x}, \mathbf{y})$
SOLUTION: Using our previous notation and the previous question,

$$\overline{\mathbf{x} + a} = \bar{x} + a \qquad \overline{\mathbf{y} + b} = \bar{y} + b$$

Therefore,

$$(x_i + a) - \overline{\mathbf{x} + a} = x_i - \bar{x}$$

and similarly for $\mathbf{y}$,

$$(y_i + b) - \overline{\mathbf{y} + b} = y_i - \bar{y}$$

The covariance is the sum of this product:

$$\frac{1}{p-1}\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y}) = \mathrm{Cov}(\mathbf{x}, \mathbf{y})$$

Therefore, the covariances are the same.

(c) If $\mathbf{y} = m\mathbf{x} + b$, then find the covariance and correlation coefficient between $\mathbf{x}$ and $\mathbf{y}$.
SOLUTION: Following a similar approach, note that

$$\overline{\mathbf{y}} = \overline{m\mathbf{x} + b} = m\bar{x} + b$$

so that

$$y_i - \bar{y} = (mx_i + b) - (m\bar{x} + b) = m(x_i - \bar{x})$$

Now the covariance between $\mathbf{x}$ and $\mathbf{y}$ is given by

$$s_{xy} = \frac{1}{p-1}\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{p-1}\sum_{i=1}^{p}(x_i - \bar{x})(m(x_i - \bar{x})) =$$

$$m\frac{1}{p-1}\sum_{i=1}^{p}(x_i - \bar{x})^2 = ms_x^2$$

For the correlation coefficient, we'll need the variance of y:

$$s_y^2 = \frac{1}{p-1}\sum_{i=1}^{p}(y_i - \bar{y})^2 = \frac{1}{p-1}\sum_{i=1}^{p} m^2(x_i - \bar{x})^2 = m^2 s_x^2$$

so that $s_y = \sqrt{m^2}s_x = |m|s_x$.

Now for the correlation coefficient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{ms_x^2}{s_x s_y} = \frac{ms_x^2}{s_x|m|s_x} = \frac{m}{|m|} = \text{signum}(m) = \begin{cases} 1 & \text{if } m > 0 \\ -1 & \text{if } m < 0 \end{cases}$$

16. Show that the affine mapping: $f(\mathbf{x}) = W\mathbf{x} + \mathbf{b}$ can be written as a linear mapping $\hat{W}\hat{\mathbf{x}}$ for an appropriate $\hat{W}$ and $\hat{\mathbf{x}}$

    SOLUTION: Define

    $$\hat{W} = [\, W \,|\, b \,] \quad \text{and} \quad \hat{x} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

    Then $\hat{W}\hat{\mathbf{x}} = W\mathbf{x} + \mathbf{b}$.

17. What does "training" mean in terms of our mathematical model?

    SOLUTION: To train a network means to find values of the parameters of the model so that we minimize some error function (usually sum of squares error). In terms of the linear network, the parameters were the weights W and bias $\mathbf{b}$.

18. If we use all the data we have at once, what kind of training are we doing? If we learn one data point at a time, what kind of training are we doing?

    SOLUTION: This is batch training vs online training.

19. Suppose I have some data in $\mathbb{R}^3$ that belongs to 4 different classes. Do I want my targets to be the real numbers $1, 2, 3, 4$, or are there better ways to build the target values?

    SOLUTION: In our examples, if we had k classes, then we used targets in $\mathbb{R}^k$. For example, class 1 corresponded to $\vec{e}_1$ (the first column of the $k \times k$ identity matrix, or the first standard basis vector of $\mathbb{R}^k$), the second class corresponded to $\vec{e}_2$, and so on.

20. Given the function $z = f(x, y)$, show that the direction in which f decreases the fastest from a point (a, b) is given by the negative gradient (evaluated at (a, b)).

    SOLUTION: Evaluating the direction derivative in an arbitrary direction $\mathbf{u}$ (where $\mathbf{u}$ is a unit vector), we get:

    $$D_u f(a, b) = \nabla f(a, b) \cdot \mathbf{u} = \|\nabla f(a, b)\| \cos(\theta)$$

    where θ is the angle between $\mathbf{u}$ and $\nabla f(a, b)$. This is a minimum when $\cos(\theta) = -1$, or when $\theta = \pi$, which occurs if $\mathbf{u} = -\nabla f(a, b)/\|\nabla f(a, b)\|$ (normalized to have unit length). In this case, we see that the value of the directional derivative is then $-\|\nabla f(a, b)\|$.

21. Illustrate the technique of gradient descent using

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

(a) Find the minimum.

SOLUTION: Set the partial derivatives to zero for the critical points:

$$\begin{array}{rl} f_x = 2x - 3y & = 0 \\ f_y = 2y - 3x & = 0 \end{array} \quad \Rightarrow \quad (x, y) = (0, 0)$$

NOTE: Continuing with this, $f_{xx}f_{yy} - f_{xy}^2 = 4 - 9 < 0$ and $f_{xx} > 0$, so the origin is a minimum.

(b) Use the initial point $(1, 0)$ and $\alpha = 0.1$ to perform two steps of gradient descent (use your calculator).

SOLUTION: Step 1 (Notice the negative sign! If we were trying to maximize the error, that would be addition instead):

$$\mathbf{x} = \mathbf{x} - \alpha\nabla f = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1}{10}\begin{bmatrix} 2 \\ -3 \end{bmatrix} = \begin{bmatrix} 4/5 \\ 3/10 \end{bmatrix}$$

After the second step, we get $x = 0.73$ and $y = 0.48$.

22. Suppose we have a subspace W spanned by an orthonormal set of non-zero vectors, $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, each is in $\mathbb{R}^{1000}$. If a vector $\mathbf{x}$ is in W, then there is a low dimensional (three dimensional in fact) representation of $\mathbf{x}$. What is it?

SOLUTION: Short answer- The low dimensional representation is the set of coordinates for $\mathbf{x}$. That is, if $\mathbf{x} \in W$, then

$$\mathbf{x} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 \quad \rightarrow (c_1, c_2, c_3)$$

23. Let the matrix A be defined below.

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}$$

(a) Find the psuedoinverse of A

SOLUTION: The pseudoinverse is from solving the normal equations (if A is full rank, which it is in this case, the rank is 2 and cannot be any bigger). Then

$$A^\dagger = (A^TA)^{-1}A^T \qquad \text{where} \qquad A^TA = \begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix}$$

Now, we do the indicated operations:

$$(A^TA)^{-1}A^T = \frac{1}{6}\begin{bmatrix} 3 & -6 \\ -6 & 14 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} -3 & 0 & 3 \\ 8 & 2 & -4 \end{bmatrix}$$

(b) Using the A from the previous exercise, consider the vector $[-1, 0, 1]^T$. Is the vector in the column space of A? If so, provide its coordinates with respect to the columns of A (for the basis).

SOLUTION: We can check directly by row reducing:

$$\left[\begin{array}{cc|c} 1 & 1 & -1 \\ 2 & 1 & 0 \\ 3 & 1 & 1 \end{array}\right] \Rightarrow \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{array}\right] \Rightarrow \left[\begin{array}{c} -1 \\ 0 \\ 1 \end{array}\right] = \left[\begin{array}{c} 1 \\ 2 \\ 3 \end{array}\right] - 2\left[\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right]$$

Therefore, the coordinates are $(1, -2)$.

(c) What happens if we try to project $[1, -2, 1]^T$ into the column space of A? Explain in terms of fundamental subspaces.

SOLUTION: If you don't see that the vector is orthogonal to the columns of A, you can row reduce:

$$\left[\begin{array}{cc|c} 1 & 1 & 1 \\ 2 & 1 & -2 \\ 3 & 1 & 1 \end{array}\right] \Rightarrow \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]$$

From this, we can at least conclude that the new vector is linearly independent of the columns of A. Checking, we see that it is orthogonal, so that $[1, -2, 1]^T$ is in null space of A^T, so projecting it into the column space would give the zero vector.

24. Consider the underdetermined "system of equations": $x+3y+4z = 1$. In matrix-vector form $A\mathbf{x} = \mathbf{b}$, write the matrix A first.

(a) What is the dimension of each of the four fundamental subspaces?

SOLUTION: First, we note that the domain is $\mathbb{R}^3$ and the codomain is $\mathbb{R}^1$. The dimension of the row space is 1, since A only has one row, and so that is also the rank of A and the dimension of the column space. That leaves the null space of A to have dimension 2 and the null space of A^T is simply $\vec{0}$ (or 0 dimensional).

(b) Find bases for the four fundamental subspaces.

SOLUTION:

A basis for the row space is the row, $[1, 3, 4]^T$ (written as a column vector).

The null space is spanned by $[-3, 1, 0]^T$ and $[-4, 0, 1]^T$.

The column space is spanned by the number 1.

The null space of A^T is only the zero vector.

(c) Find a solution with at least 2 zeros (the slash command in Matlab looks for answers with the most zeros).

SOLUTION: You can find them by inspection. For example, $(1, 0, 0)$ is a solution, as is $(0, 0, 1/4)$.

(d) Find a 3×3 matrix P so that given a vector $\mathbf{x}$, $P\mathbf{x}$ is the projection of $\mathbf{x}$ into the row space of A.

SOLUTION: This will be a projection matrix into the space spanned by a single vector. In the notes, we had:

$$P = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}} = \frac{1}{26}\begin{bmatrix} 1 & 3 & 4 \\ 3 & 9 & 12 \\ 4 & 12 & 16 \end{bmatrix}$$

25. (Eigenvalues) Find the eigenvalues and eigenvectors for $A = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}$

SOLUTION: First we compute the characteristic equation, $|A - \lambda I| = 0$, or $\lambda^2 - \lambda - 2 = 0$, from which $\lambda = 2$ or $\lambda = -1$.

For $\lambda = 2$, we find the null space of $A - 2I$. Below, we'll use v_1 as free (but you could use either).

$$\left[\begin{array}{cc|c} -2 & 1 & 0 \\ 2 & -1 & 0 \end{array}\right] \quad \Rightarrow \quad \begin{array}{rl} v_1 & = v_1 \\ v_2 & = 2v_1 \end{array} \quad \Rightarrow \quad E_2 = \text{span}\left\{\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right\}$$

Similarly, you should find that E_{-1} is the span of $[-1, 1]^T$.

26. (Eigenvalues) Verify the 4 results of the Spectral Theorem for $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

SOLUTION: As we said in class, for now we only look at the first three items:

- The eigenvalues are all real. You should find $\lambda = 3$ and $\lambda = 1$.
- The algebraic and geometric multiplicities are equal: The eigenspaces are each 1-dimensional.
- The eigenspaces are the following, and the basis vectors are orthogonal:

$$E_1 = \text{span}\left\{\begin{bmatrix} -1 \\ 1 \end{bmatrix}\right\} \qquad E_3 = \text{span}\left\{\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right\}$$

27. (Eigenvalues) If λ_i is an eigenvalue of A^TA, then show that $\lambda_i \geq 0$ by showing that, if $\mathbf{v}_i$ is an eigenvector for λ_i, then $\|A\mathbf{v}_i\|^2 = \lambda_i$ (and lengths cannot be negative).

SOLUTION: Use the multiplicative form of the norm. That is,

$$\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x}$$

Therefore,

$$\|A\mathbf{v}_i\|^2 = (A\mathbf{v}_i)^TA\mathbf{v}_i = \mathbf{v}_i^TA^TA\mathbf{v}_i = \mathbf{v}_i^T(A^TA\mathbf{v}_i) = \lambda_i\mathbf{v}_i^T\mathbf{v}_i = \lambda_i\|\mathbf{v}_i\|^2$$

The left side of the equation is positive, so the right side must be as well. If we assume additionally that $\mathbf{v}_i$ is a unit vector (which is common), then we get the desired expression.

28. (Eigenvalues) If $\mathbf{v}_i$ and $\mathbf{v}_j$ are eigenvectors corresponding to distinct eigenvalues of A^TA, then show that $A\mathbf{v}_i \perp A\mathbf{v}_j$.

SOLUTION: Let's start backwards. We want to show that:

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

The last statement is true because A^TA is symmetric, and is from the Spectral Theorem (eigenvectors from distinct eigenvalues are orthogonal).

29. (Eigenvalues) Suppose that $\lambda_i, \mathbf{v}_i$ are eigenvalue/eigenvectors for a symmetric matrix A^TA (so the Spectral Theorem applies). Prove that, if $\boldsymbol{x} = \alpha_1 \boldsymbol{v}_1 + \ldots \alpha_n \boldsymbol{v}_n$, then

$$\|A\boldsymbol{x}\|^2 = \alpha_1^2 \lambda_1 + \ldots + \alpha_n^2 \lambda_n$$

SOLUTION: We could prove this directly, but we have already proven the Pythagorean Theorem, so let's use that. We will assume that the eigenvectors are unit vectors.

$$\boldsymbol{x} = \alpha_1 \boldsymbol{v}_1 + \ldots \alpha_n \boldsymbol{v}_n \quad \Rightarrow \quad A\boldsymbol{x} = \alpha_1 A\boldsymbol{v}_1 + \ldots \alpha_n A\boldsymbol{v}_n$$

From Exercise 28, we know these are all orthogonal, so the Pythagoream theorem applies:

$$\|A\boldsymbol{x}\|^2 = \|\alpha_1 A\boldsymbol{v}_1\|^2 + \ldots \|\alpha_n A\boldsymbol{v}_n\|^2 = \alpha_1^2 \|A\boldsymbol{v}_1\|^2 + \ldots \alpha_n^2 \|A\boldsymbol{v}_n\|^2$$

And from Exercise 27, we know that $\|A\mathbf{v}_i\|^2 = \lambda_i \|\mathbf{v}_i\|^2 = \lambda_i$, so that the equation above simplifies to what we want:

$$\|A\boldsymbol{x}\|^2 = \alpha_1^2 \lambda_1 + \ldots + \alpha_n^2 \lambda_n$$

30. (Eigenvalues) Prove that if λ_i is an eigenvalue of A^TA, then λ_i is also an eigenvalue of AA^T (Hint: Let $\mathbf{u}_i = A\mathbf{v}_i$, where $\mathbf{v}_i$ is an eigenvector associated with λ_i).

SOLUTION: We want to show that $AA^T\mathbf{w} = \lambda_i \mathbf{w}$ for some vector $\mathbf{w}$ (we're given a hint about $\mathbf{w}$). First, from the definition of an eigenvalue:

$$A^TA\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

Multiply both sides by the matrix A:

$$AA^TA\mathbf{v}_i = \lambda_i A\mathbf{v}_i \quad \Rightarrow \quad AA^T(A\mathbf{v}_i) = \lambda_i (A\mathbf{v}_i) \quad \Rightarrow \quad AA^T\mathbf{u}_i = \lambda_i \mathbf{u}_i$$