Math 350, Exam 2 Review SOLUTIONS

1. What's a Voronoi diagram?

   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).

2. Data clustering is unsupervised learning. Given data and the number of centers k, we try to determine a membership function m so that $m(\mathbf{x}) = j$, where j is the label of the j^{th} cluster.

3. How is the rank computed when we construct either the reduced SVD or the pseudoinverse?

   SOLUTION: The theoretical rank (from the SVD) is the number of non-zero singular values. Numerically, we look at the normalized eigenvalues of the covariance (square of the singular values):
   $$\hat{\lambda}_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}$$
   And rank k is determined so that some fixed percent of the overall variance is retained. For example, if the variance level is 0.90, we choose k so that
   $$\sum_{j=1}^{k} \hat{\lambda}_j \geq 0.90 \quad \text{but} \quad \sum_{j=1}^{k-1} \hat{\lambda}_j < 0.90$$

4. How do you change an affine equation into a linear equation?

   To change the matrix-vector equation:
   $$A\mathbf{x} + \mathbf{b} = \mathbf{y}$$
   into an equivalent linear equation, $\hat{A}\hat{\mathbf{x}} = \mathbf{y}$, the values of $\mathbf{b}$ are appended as a last column to A, and ones are appended as the last row of $\mathbf{x}$ (we could have a matrix-matrix equation $AX + b = Y$).

5. Recall that if we have a matrix B so that $AB = I$ and $BA = I$, then matrix B is called the inverse of matrix A.

   Does the pseudo-inverse of the matrix A, $A^{\dagger}$, satisfy the same properties? Explain (using the SVD):

   If $A = U\Sigma V^T$ is the (reduced) SVD, then $A^{\dagger} = V\Sigma^{-1}U^T$, and
   $$AA^{\dagger} = U\Sigma V^T V\Sigma^{-1}U^T = UU^T$$
   $$A^{\dagger}A = V\Sigma^{-1}U^T U\Sigma V^T = VV^T$$

So in the first case, $AA^\dagger$ is the projection matrix to the columnspace of U (which is the column space of A). In the second case, $A^\dagger A$ is the projection matrix to the columnspace of V (which is the row space of A).

If matrix A was invertible, it would be square with full rank, so in that particular case, $UU^T = VV^T = I$.

6. What is Hebb's rule (the biological version)? (See the notes)

7. Widrow-Hoff update rule for a linear network, $\mathbf{y} = W\mathbf{x} + \mathbf{b}$ is given by the following, where $\mathbf{x}$ is a point chosen at random:

$$W_{\text{new}} = W_{\text{old}} + \alpha(\mathbf{t} - \mathbf{y})\mathbf{x}^T$$

$$b_{\text{new}} = b_{\text{old}} + \alpha(\mathbf{t} - \mathbf{y})$$

8. In pattern classification, suppose I have data in the plane that I want to divide into 5 classes. Would I want to build a pattern classification function f so that the range is the following set:

$$\{1, 2, 3, 4, 5\}$$

Why or why not? If not, what might be a better range?

SOLUTION: Using this classification implies that there is a metric with meaning- That class 1 is closer to class 2 than class 5, for example. Unless that is what you want, you should try to use class labels without so much ordering. With 5 classes, you might consider the 5 classes labels:

$$\begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix}, \quad \begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}$$

which are the 5 rows of I_5. In class, when we had an even number of classes, we decided we could use ±1 in each entry, like

$$\begin{bmatrix}1\\1\end{bmatrix}, \quad \begin{bmatrix}-1\\1\end{bmatrix}, \quad \begin{bmatrix}1\\-1\end{bmatrix}, \quad \begin{bmatrix}-1\\-1\end{bmatrix},$$

9. Given the function $f(x, y)$, 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: Given a function $z = f(x, y)$, at a point (a, b) we measure the rate of change in the direction of unit vector $\mathbf{u}$ as:

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

where θ is the (acute) angle between ∇f and $\mathbf{u}$. This simplifies, since we have a unit vector:

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

The "most negative" this quantity can be is $-\|\nabla f\|$, when $\cos(\theta) = 180$, or when we move in the negative direction of the gradient.

10. Illustrate the technique of gradient descent using

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

(a) Find the minimum.

Find the critical points first. The gradient is $\langle 2x - y, 2y - x\rangle$, so that setting them to zero gives

$$\begin{aligned} 2x &= y \\ 2y &= x \end{aligned} \qquad \Rightarrow \qquad x = 0, y = 0$$

The Hessian matrix is

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \qquad \lambda = 1, 3$$

Since both are positive, we have a local minimum. It will be the global minimum since the Hessian does not change with $\mathbf{x}$.

You could also use the second derivatives test, where $D = f_{xx}f_{yy} - f_{xy}^2 = 3 > 0$ and $f_{xx} > 0$, so we have a local minimum.

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

SOLUTION: The update algorithm is $\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha \nabla f(\mathbf{x}_i)$.

- First step, with $\nabla f = [2x - y, -y + 2x]^T$:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.1 \end{bmatrix}$$

- Second step:

$$\mathbf{x}_2 = \begin{bmatrix} 0.8 \\ 0.1 \end{bmatrix} - 0.1 \begin{bmatrix} 1.5 \\ -0.6 \end{bmatrix} = \begin{bmatrix} 0.65 \\ 0.16 \end{bmatrix}$$

(Although the y coordinate is going away from the origin, it will eventually go back to zero).

11. If

$$f(t) = \begin{bmatrix} 3t - 1 \\ t^2 \end{bmatrix}$$

find the tangent line to f at $t = 1$.

SOLUTION: The tangent line will be $f(1) + f'(1)(t-1)$, or

$$\left[\begin{array}{c} 2 \\ 1 \end{array}\right] + (t-1)\left[\begin{array}{c} 3 \\ 2 \end{array}\right]$$

NOTE: You could verify this by translating the function into the form $y = f(x)$.

12. If $f(x,y) = x^2 + y^2 - 3xy + 2$, find the linearization of f at $(1,0)$.

    SOLUTION:

$$L(x,y) = f(1,0) + \nabla f(1,0)\left[\begin{array}{c} x-1 \\ y-0 \end{array}\right] = 3 + [2 \quad -3]\left[\begin{array}{c} x-1 \\ y \end{array}\right] = 3 + 2(x-1) - 3y$$

13. Given just one data point:

$$X = \left[\begin{array}{c} 2 \\ -1 \end{array}\right] \qquad T = [1]$$

Initializing W and $\mathbf{b}$ as an appropriately sized arrays of ones, perform three iterations of Widrow-Hoff using $\alpha = 0.1$ (by hand, you may use a calculator- On the exam, the numbers should work out somewhat nicely without). You should verify that the the weights and biases are getting better.

SOLUTION: $W = [1 \quad 1]$, $b = 1$, so $y = (2-1) + 1 = 2$. Therefore,

$$W = W + \alpha(t-y)\mathbf{x}^T = [1 \quad 1] + 0.1(1-2)[2, \quad -1] = [0.8, \quad 1.1]$$

and $b = 1 + 0.1(1-2) = 0.9$.

Now the new value of $y = 1.4$, so that

$$W = [0.8 \quad 1.1] + 0.1 \cdot (1 - 1.4)[2 \quad -1] = [0.72 \quad 1.14]$$

$$b = 0.9 + 0.1 \cdot (1 - 1.4) = 0.86$$

And the new value of $y = 1.16$. One last update:

$$W = [0.72 \quad 1.14] + 0.1 \cdot (1 - 1.16)[2 \quad -1] = [0.688 \quad 1.156]$$

$$b = 0.86 + 0.1 \cdot (1 - 1.16) = 0.844$$

And the new value of y will be 1.06, so we are coming close to the desired value.

14. How did we define the notion of "best" in the best basis? To help, suppose we have an arbitrary orthonormal basis $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n\}$ and data $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p\}$.

    SOLUTION: Probably easiest to define $\mathbf{x}_{\text{err}}^{(j)}$ first. For the j^{th} data point, we write $\mathbf{x}^{(j)}$ in terms of our given orthonormal basis. The error vector for this point is the vector formed by using basis vectors $k+1$ to n:

$$\mathbf{x}_{\text{err}}^{(j)} = \sum_{i=k+1}^{n} \alpha_i^{(j)} \boldsymbol{\phi}_i$$

Then the overall error is the average:

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

15. If C is the covariance matrix given below, find the maximum and minimum of $F(\phi)$, and give the ϕ for which the maximum occurs (we may assume ϕ is not the zero vector, and that ϕ is a vector with 2 elements).

$$F(\phi) = \frac{\phi^T C \phi}{\phi^T \phi} \qquad \text{for } C = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

(Hint: You may find it easily using our theorems)

SOLUTION: The maximum and minimum values are given by the eigenvalues (there are only two) to C, and the vectors ϕ are the corresponding eigenvectors. Therefore, we just need to find the evecs and evals of C:

$$\det(C - \lambda I) = (3-\lambda)^2 - 1 = \lambda^2 - 6\lambda + 8 = 0$$

so $\lambda = 2, 4$. For $\lambda = 2$, solve $(C - 2I)\boldsymbol{\phi} = 0$:

$$\left[\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 1 & 0 \end{array}\right] \quad \Rightarrow \quad \boldsymbol{\phi} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

This $\boldsymbol{\phi}$ is the minimizer for F. For the maximizer, we consider $\lambda = 4$, and solve $(C - 4I)\boldsymbol{\phi} = 0$:

$$\left[\begin{array}{cc|c} -1 & 1 & 0 \\ 1 & -1 & 0 \end{array}\right] \quad \Rightarrow \quad \boldsymbol{\phi} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

16. Questions from the best basis material:

(a) If $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ are eigenvectors of a symmetric matrix, do they have any nice properties?

SOLUTION: They form an orthonormal (o.n.) basis for $\mathbb{R}^n$.

(b) Given that the previous set of eigenvectors forms an o.n. basis for $\mathbb{R}^n$, and $\boldsymbol{\phi}$ is a vector in $\mathbb{R}^n$, then justify the fact that we can write

$$\boldsymbol{\phi} = V\mathbf{a}$$

SOLUTION: This is the matrix form of:

$$\boldsymbol{\phi} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n \qquad \text{where} \qquad V = [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_n]$$

Additionally, we might note that $a_i = \boldsymbol{\phi} \cdot \mathbf{v}_i$.

(c) Given the previous statement, show that

$$\|\boldsymbol{\phi}\|^2 = \mathbf{a}^T\mathbf{a}$$

SOLUTION:

$$\|\boldsymbol{\phi}\|^2 = \boldsymbol{\phi}^T\boldsymbol{\phi} = \mathbf{a}^T V^T V\mathbf{a} = \mathbf{a}^T\mathbf{a}$$

since $V^TV = I$.

(d) Recall that the eigenvectors of C are V so that $C = V\Lambda V^T$. Show that this implies that

$$\frac{\boldsymbol{\phi}^T C\boldsymbol{\phi}}{\boldsymbol{\phi}^T\boldsymbol{\phi}} = \frac{\mathbf{a}^T\Lambda\mathbf{a}}{\mathbf{a}^T\mathbf{a}}$$

SOLUTION: Substitute $\boldsymbol{\pi} = V\mathbf{a}$ from the previous problem along with $C = V\Lambda V^T$ now, and do some matrix algebra:

$$\frac{\boldsymbol{\phi}^T C\boldsymbol{\phi}}{\boldsymbol{\phi}^T\boldsymbol{\phi}} = \frac{(V\mathbf{a})^T V\Lambda V^T(V\mathbf{a})}{\mathbf{a}^T\mathbf{a}} = \frac{\mathbf{a}^T V^T V\Lambda V^T V\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = \frac{\mathbf{a}^T\Lambda\mathbf{a}}{\mathbf{a}^T\mathbf{a}}$$

(e) **TYPO:** a_i in the numerators should be a_i^2.

Show that, assuming $\lambda_1 \geq \lambda_2 \geq \cdot \geq \lambda_n$, and $a_i \geq 0$ for all i, then

$$\lambda_1\frac{a_1^2}{a_1^2 + a_2^2 + \cdots a_n^2} + \lambda_2\frac{a_2^2}{a_1^2 + a_2^2 + \cdots a_n^2} + \cdots \lambda_n\frac{a_n^2}{a_1^2 + a_2^2 + \cdots a_n^2} \leq \lambda_1$$

with equality if $a_1 = 1$ and all other $a_i = 0$.

SOLUTION: We showed this in class, where we defined p_i instead of this notation, but it works out the same. Let

$$p_i = \frac{a_i^2}{a_1^2 + \cdots a_n^2}$$

Therefore, $p_i \geq 0$ and $\sum p_i = 1$. Therefore,

$$\lambda_1 p_1 + \lambda_2 p_2 + \cdots \lambda_n p_n \leq \lambda_1(p_1 + \cdots + p_n) = \lambda_1$$

And a similar argument gives us the minimum λ_n.

17. Find the SVD of the “matrix”: $X = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$

SOLUTION:

$$XX^T = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \qquad X^TX = 6 \quad \Rightarrow \quad \lambda = 6, 0, 0$$

It's easy to get the eigenspace for $\lambda = 0$. If we row reduce the matrix, we simply get the following matrix. With two free variables, the eigenspace is two dimensional:

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \Rightarrow \quad \begin{array}{ll} x_1 & = -2x_2 - x_3 \\ x_2 & = x_2 \\ x_3 & = \qquad\quad x_3 \end{array} \quad \Rightarrow \quad \mathbf{u}_{2,3} = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

Technical Note: It is possible to re-do this basis so that it is orthogonal, but I didn't mean for you to do that extra work. You might just note that here... (It was called the Gram-Schmidt process in linear algebra, but I won't ask you to do that in-class.)

For the last eigenvector, you might note that $\mathbf{x}$ itself is an eigenvector, since:

$$(XX^T)X = X(X^TX) = 6X$$

So, the **reduced** SVD (since we don't want to use the null space) is:

$$\hat{U} = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \qquad \Sigma = \sqrt{6} \qquad V = 1$$

18. If I know the vector $\mathbf{v}_1$ and the singular value σ_1 from the SVD of a matrix A, can I compute $\mathbf{u}_1$ directly? Was σ_1 really needed?

    SOLUTION: Since $A = U\Sigma V^T$, then $AV = U\Sigma$, or columnwise for U,

    $$A\mathbf{v}_1 = \sigma_1 \mathbf{u}_1 \qquad \text{or} \qquad \frac{1}{\sigma_1} A\mathbf{v}_1 = \mathbf{u}_1$$

    We did not need σ_1. We could take $A\mathbf{v}_1$, then normalize it.

19. Given data in $\mathbb{R}$: $x_1, \ldots, x_p$, show that, if we define the function E below:

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

    then find the value of m that minimizes E.

    SOLUTION: Find dE/dm and set it to zero. Since we have quadratic terms, we will then find a minimum- there is no max.

    $$\frac{dE}{dm} = \frac{2}{p} \sum_{i=1}^{p} (x_i - m)(-1) = 0 \quad \Rightarrow \quad \sum_{i=1}^{p} x_i - \sum_{i=1}^{p} m = 0 \quad \Rightarrow \quad \sum_{i=1}^{p} x_i = mp \quad \Rightarrow \quad m = \frac{1}{p} \sum_{i=1}^{p} x_i$$

    so the best value of m is the average (and that's what makes k-means work!)

20. Give the algorithm for $k-$means clustering:

- Initialize by setting k and initializing the cluster centers.
- Repeat these steps:
  - Sort the centers by distance into k clusters.
  - Reset the centers as the mean of the data currently in the appropriate cluster (there are k of them).

21. Give the cluster update rule for Kohonen's self organizing map.

    SOLUTION: Be sure to define any variables used. We might start off this way:

    Initialize by setting the number of centers and the grid topology. This defines the grid metric between clusters i and w as $d_I(i,w)$. We also initialize the learning rate ϵ (we can optionally give initial and final values of that), and the spread λ (again, we might give initial and final values).

    With a given data point $\mathbf{x}$ and w is the index of the winning center,

    $$C_{\text{new}} = C_{\text{old}} + \epsilon \ \exp\left(\frac{-d_I^2(i,w)}{\lambda^2}\right)(\mathbf{x} - C_{\text{old}})$$

22. Give the cluster update rule for Neural Gas.

    SOLUTION: Be sure to define any variables used. We might start off this way:

    Initialize by setting the number of centers. The metric being used is "the number of centers closer to the winner than the current one", and that is $d_{ng}(i,w)$.

    We also initialize the learning rate ϵ (we can optionally give initial and final values of that), and the spread λ (again, we might give initial and final values).

    With a given data point $\mathbf{x}$ and w is the index of the winning center,

    $$C_{\text{new}} = C_{\text{old}} + \epsilon \ \exp\left(\frac{-d_{ng}^2(i,w)}{\lambda^2}\right)(\mathbf{x} - C_{\text{old}})$$

    We would also mark down that there is an edge between the winning cluster and the next closest cluster. Finally, we would remove edges that are too old (so we're also keeping time on each edge).

23. What is the main difference between SOM and Neural Gas?

    SOLUTION: SOM has a fixed, pre-determined topological structure for the centers. The neural gas algorithm tries to determine the topological structure that will make the clusteing "topology preserving".

24. Here is one data point. There are three centers in the matrix C which have a linear topology- That is, I gives the one-dimensional representation of each cluster center.

Perform one update of the centers using Kohonen's SOM update rule, assuming that $\epsilon = \lambda = 1$ (unrealistic, but easier to do by hand):

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad C = \begin{bmatrix} -1 & 1 & 2 \\ 1 & 0 & 3 \end{bmatrix} \qquad I = [1, 6, 3]$$

Also, for the distance in the plane, use the "taxicab" or "Manhattan" metric:

$$d(\mathbf{a}, \mathbf{b}) = |a_1 - b_1| + |a_2 - b_2|$$

SOLUTION: See me if you're having trouble with this. It's just to be sure you're comfortable with the update rule.

25. Same as the previous problem, but update using the Neural Gas algorithm (assume all the centers are connected and ignore the age). Use $\epsilon = \lambda = 1$ (unrealistic, but this is by hand). For the metric in the plane, again use the taxicab metric.

    SOLUTION: See me if you're having trouble with this. It's just to be sure you're comfortable with the update rule.