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. 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 $\theta = \pi$, or when we move in the negative direction of the gradient.

5. 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

$$\begin{array}{ll} 2x & = y \\ 2y & = x \end{array} \quad \Rightarrow \quad x=0, y=0$$

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 one steps of gradient descent (use your calculator).

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

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

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

(c) Same, but use line search to find the optimal value of the step size.

SOLUTION: From our previous computation, the line is given by:

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

We could compute f on this line directly, or we can use the chain rule:

$$\frac{df}{d\alpha} = f_x\frac{dx}{d\alpha} + f_y\frac{dy}{d\alpha}$$

Computing these:

$$f_x(1-2\alpha,\alpha) = 2-5\alpha, \qquad f_y(1-3\alpha,\alpha) = 4\alpha - 1 \qquad \frac{dx}{d\alpha} = -2 \qquad \frac{dy}{d\alpha} = 1$$

Therefore,

$$f_x\frac{dx}{d\alpha} + f_y\frac{dy}{d\alpha} = -5 + 14\alpha = 0 \quad \Rightarrow \quad \alpha = \frac{5}{14}$$

(d) One step of multivariate Newton:

$$\mathbf{x}_i - (Hf(\mathbf{x}_i)^{-1}\nabla f(\mathbf{x}_i)$$

We have all the pieces:

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad Hf = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \Rightarrow Hf^{-1} = \frac{1}{5}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \nabla f(1,0) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Put these into the formula to get $(2/5, 0)$.

6. 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

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} + (t-1)\begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

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

SOLUTION:

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

8. 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:

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

9. 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 $(C - 4I)\boldsymbol{\phi} = 0$:

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

10. 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.

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

$$\sum_{i=1}^{p} x_i = mp \quad \Rightarrow \quad m = \frac{1}{p} \sum_{i=1}^{p} x_i$$

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

11. 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).

12. 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,

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

13. 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).

14. 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".

15. 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.

16. 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.

17. In the DBSCAN algorithm, points are classified into three different groups- What were the groups, and how were the groups defined?

    SOLUTION: Core points (points that have at least `MinPts` points in its $\epsilon-$neighborhood), boundary points (points within ϵ of a core point, but with less than MinPts points in its neighborhood), and noise (not a core or boundary point).

18. What are the training parameters that must be set before using the DBSCAN algorithm?

    SOLUTION: MinPts and ϵ.

19. Define what it means for q to be (directly) density-reachable from p (in the DBSCAN context).

    SOLUTION: Point q is directly density reachable from point p if two things are true: p is a core point, and q is within ϵ of p.

20. What are the "inputs" to Kohonen's SOM? (That is, what information needs to be provided to the algorithm)?

    SOLUTION: Kohonen's SOM primarily needs the structure we've placed on the centers (like in a two dimensional array). We'll need to be able to compute the distance between centers using this construction. The training parameters were ϵ (the learning rate) and λ (the spread of the Gaussian).

21. Similarly, what are the inputs to Neural Gas?

    SOLUTION: Like the SOM, we need ϵ and λ, but in place of the center topology, we'll need to know how long an edge can age before it is removed.

22. Let $f(x, y) = 3x^2 + xy - y^2 + 3x - 5y$.

    (a) From the point $(1, 1)$, in which direction is f increasing the fastest?

    That would be in the direction of the gradient, which in this case is $(10, -6)$ (I'm using Lay's notation so that this would be equivalent to a column vector).

    (b) Find the critical point of f.

    Setting the gradient to 0, we should find that $x = -1/13$ and $y = -33/13$.

    (c) Compute the Hessian of f and determine if f has a local max or min at the critical point (recall that we compute eigenvalues, but we only need the signs of the eigenvalues).

    The Hessian is $\begin{bmatrix} 6 & 1 \\ 1 & -2 \end{bmatrix}$, so the characteristic equation is

$$\lambda^2 - 4\lambda - 13 = 0 \quad \Rightarrow \quad \lambda^2 - 4\lambda + 4 = 17 \quad \Rightarrow \quad \lambda = -2 \pm \sqrt{17}$$

    So the eigenvalues are mixed in sign- we have a saddle point.

23. Let $f(x, y) = 3xy + x^2$.

(a) Linearize f about the point $(1, 1)$.

$$f(1,1) + \nabla f(1,1)(x-1, y-1) = 4 + (5,3)(x-1, y-1) = 4 + 5(x-1) + 3(y-1)$$

(b) Compute the Hessian of f.

$$\begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix}$$

(c) Show that Newton's Method, starting at $(1, 1)$, converges in one step.

$$(1,1) - (Hf^{-1}(1,1))(\nabla f(1,1)) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{9}\begin{bmatrix} 0 & -3 \\ -3 & 2 \end{bmatrix}\begin{bmatrix} 5 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$