Math 350 Exam 2 Review Questions

1. What is a Voronoi diagram?

2. Is data clustering an example of supervised or unsupervised learning? Explain (and give an explanation of the overall problem).

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

4. How do you change an affine equation into a linear equation? That is, change the matrix-vector equation:
$$A\mathbf{x} + \mathbf{b} = \mathbf{y}$$
into an equivalent linear equation, $\hat{A}\hat{\mathbf{x}} = \mathbf{y}$:

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):

6. What is Hebb's rule (the biological version)? You can paraphrase it:

7. What is the Widrow-Hoff update rule? You may write it either in matrix form or in scalar form.

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?

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)).

10. Illustrate the technique of gradient descent using
$$f(x, y) = x^2 + y^2 - xy + 2$$
    (a) Find the minimum.
    (b) Use the initial point $(1, 0)$ and $\alpha = 0.1$ to perform two steps of gradient descent (use your calculator).

11. If
$$f(t) = \begin{bmatrix} 3t - 1 \\ t^2 \end{bmatrix}$$
find the tangent line to f at $t = 1$.

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

13. Given just one data point:

$$X = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \qquad T = [1]$$

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). You should verify that the the weights and biases are getting better.

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 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p\}$.

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)

16. A few questions stemming from the best basis:

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

    (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

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

    (c) Given the previous statement, show that

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

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

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

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

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

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

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

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 needed?

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.

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

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

22. Give the cluster update rule for Neural Gas.

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

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:

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

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.