Math 350, Exam 2 Review SOLUTIONS

Overview

In this third of the course, we focused on linear learning algorithms to model data. To summarize:

1. Background: The SVD and the best basis (questions selected from Ch. 6- Can you fill in the exercises?)
2. How is the rank computed? (theoretically and computationally)

   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
   $$\sum_{j=1}^{k} \hat{\lambda}_j \geq 0.90 \quad \text{but} \quad \sum_{j=1}^{k-1} \hat{\lambda}_j < 0.90$$

3. The SVD and the pseudo-inverse: How is it computed? Where did we use the pseudo-inverse?

   SOLUTION: Given a system of equations, with unknown $\mathbf{x}$, we can find the solution (or least squares solution if A is not invertible) by using the pseudo-inverse of A. That is, given the system of equations and the *reduced* SVD (where the rank is determined using some level of variance or equivalently, some level of energy),
   $$A\mathbf{x} = \mathbf{b} \quad \text{with} \quad A = U\Sigma V^T$$
   then $A^\dagger = V\Sigma^{-1}U^T$, where Σ is a square diagonal matrix with non-zero values along the diagonal (so that its inverse is found by taking the reciprocal of each diagonal element).

   Then,
   $$A^\dagger A\mathbf{x} = A^\dagger\mathbf{b} \quad \Rightarrow \quad V\Sigma^{-1}U^T U\Sigma V^T\mathbf{x} = A^\dagger\mathbf{b} \quad \Rightarrow \quad VV^T\mathbf{x} = A^\dagger\mathbf{b}$$
   This shows that the (least squares) solution is the projection of $\mathbf{x}$ into the column space of V (which is the row space of A).

4. Lines of best fit: How did we get the error measures?

   Draw a picture: The first line came from measuring error vertically between the actual y value and the $y-$value coming from the line. The second line came from measuring error by orthogonally projecting each point to the line. The third line (median-median) did not have an error function.

   Give a short derivation of two of the algorithms (error in $y-$coordinate and the median - median line). (See the notes)

   Be able to give the derivation of the second error function (orthogonally projecting the data to the line). (See your homework)

5. Finding the best linear function:

   (a) 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 $AX + b = Y$).

   (b) Hebb's Rule (the biological version) See the notes

   (c) Hebb's Rule (the version with no feedback on p. 6) See the notes.

   (d) Will the rule on p. 6 converge? (Exercises on p. 6) No. See the exercises right after the rule.

   (e) The failure leads to Widrow-Hoff (p. 7)

6. Derivatives (Appendix A)

   Be sure you can linearize different kinds of functions (like the examples, p. 5-6)

   Be able to write a quadratic as $\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + \mathbf{c}$, and take the first and second "derivatives" (e.g., find the gradient and Hessian)- Like exercises 7-8 of the homework.

   Be able to explain the method of gradient descent, and explain by approximately how much we drop (in terms of the function) at each step. You might show it in one dimension.

7. Back to linear learning algorithms: How is it that Widrow-Hoff is (approximately) gradient descent? (Be explicit, starting with the error function).

   Full gradient descent would require computation of the error function, which in this case is the Mean Square Error (MSE). With p data points $\mathbf{x}^{(i)}$, targets $\mathbf{t}^{(i)}$, and network output $\mathbf{y}^{(i)} = W\mathbf{x}^{(i)} + \mathbf{b}$,

   $$E_{mse} = \frac{1}{p}\sum_{j=1}^{p} \|\mathbf{t}^{(j)} - \mathbf{y}^{(j)}\|^2$$

Computing this quantity for gradient descent would require all p data points. In online training, we do not have all data. Therefore, the error is estimated by using the only the current points $\mathbf{t}^{(j)}$ and $\mathbf{y}^{(j)}$.

Using the error based only on this point as an estimate of E_{mse}, we look at Widrow-Hoff (or modified Hebb) as performing a kind of gradient descent.

8. What is translationally invariant data?

   SOLUTION: Not covered yet. You may skip this section for the exam.

9. Best subspaces as feature extraction: If we have p data "points" (really vectors) in $\mathbb{R}^n$, then looking for a small set of "template vectors" (or feature vectors) so that each point is a linear combination of the features is the same as finding the best basis for the data set.

Review Questions

1. We had three lines of best fit- Two of them were designed to minimize error functions- What were the error functions (also show them graphically)?

   See the notes above. Can you write the error functions?

$$E_1 = \frac{1}{2p}\sum_{j=1}^{p}(y_j - (mx_j + b))^2$$

$$E_2 = \frac{1}{2p}\sum_{j=1}^{p}\frac{(ax + by + c)^2}{a^2 + b^2}$$

   (You could leave the 2 out of the denominator. It is only there to cancel when you compute the derivatives, but the critical points do not change).

2. Illustrate the median-median line (you may use a calculator) given the data below:

x	-1	2	1	0	6	3	5	-2
y	5	-1	1	2	-8	-3	-7	7

   SOLUTION: Be able to do this by hand. First re-order the $x's$ (remember to also re-order the y's!). Using the first and last group of three points, find the equation of the line between $(-1, 5)$ and $(5, -7)$, which is $y = -2x + 3$. Now check the middle median: $(3/2, 0)$. In this particular case, the line goes exactly through the median, so no further shift is needed (otherwise, shift by 1/3 towards the middle point).

3. 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, $UU^T = VV^T = I$.

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

5. 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

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

6. 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.

7. Illustrate the technique of gradient descent using

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

TYPO: Please make the function $f(x,y) = x^2 + y^2 - xy + 2$.

(a) Find the minimum.

Solve for the critical point (the origin). You can show that it is a minimum by using the second derivatives test or by forming A so that $f(x,y) = \mathbf{x}^T A\mathbf{x}$ (A should be symmetric). For example,

$$f(x,y) = \mathbf{x}^T \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix} \mathbf{x} + 2$$

Then the gradient is

$$\nabla f = A\mathbf{x}$$

and the only time this is zero is at the origin. The Hessian matrix is the matrix A, whose eigenvalues are $3/2, 1/2$ (both positive).

(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).

8. 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

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

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

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

   SOLUTION:

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

10. Given just one data point:

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

   Initializing W and **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.

   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.

11. If a time series is given by:

$$x = \{1, 2, 0, 3, 4, 5, 2, 1, 0, 3, 4\}$$

   Give the result of performing lag 2:

   SOLUTION:

$$\begin{bmatrix} 1 & 2 & 0 & 3 & 4 & 5 & 2 & 1 & 0 \\ 2 & 0 & 3 & 4 & 5 & 2 & 1 & 0 & 3 \\ 0 & 3 & 4 & 5 & 2 & 1 & 0 & 3 & 4 \end{bmatrix}$$

12. If the time series is periodic with period k, what happens when we perform a lag $k-1$?

$$x = \{x_1, x_2, \ldots, x_k, x_1, x_2, \ldots, x_k, \ldots\}$$

SOLUTION: It is fixed (never changes).

13. Be sure you can provide justifications for statements 3-5, p. 96 of Chapter 6 (best basis)- You actually did this for a specific 2-dimensional case in Exam 1.

    3. Writing vector ϕ in terms of the basis given by the $\mathbf{v}$'s, and if we let the vector $\mathbf{a}$ be the vector of coordinates, we have

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

    Now, since the columns of V are orthonormal,

    $$\phi^T\phi = \mathbf{a}^T V^T V\mathbf{a} = \mathbf{a}^T I\mathbf{a} = \mathbf{a}^T\mathbf{a}$$

    (Side note: This is a statement of the Pythagorean Theorem, as we discussed in class).

    4. Recall that the vectors V are the eigenvectors of C (that is, $C = V\Lambda V^T$), and substituting, we have:

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

    Which is the expression shown in the text.

    5. The eigenvalue λ_1 is the largest eigenvalue. Therefore,

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

    Let $b_i = \frac{a_i^2}{a_1^2+a_2^2+\cdots+a_n^2}$. Then $0 \le b_i \le 1$, and we can write the expression as:

    $$\frac{\lambda_1 a_1^2 + \cdots \lambda_n a_n^2}{a_1^2 + a_2^2 + \cdots + a_n^2} = \lambda_1 b_1 + \cdots \lambda_n b_n \le \lambda_1(b_1 + b_2 + \cdots b_n) = \lambda_1$$

14. 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.