CS 4980/6980: Introduction to Data Science © Spring 2018

Lecture 2: Linear Algebra Review

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN Scribed by: Anh Nguyen and Kira Jordan

This is preliminary work and has not been reviewed by instructor. If you have comments about typos, errors, notation inconsistencies, etc., please email Anh Nguyen and Kira Jordan at `anguyen139@student.gsu.edu`, `kjordan32@student.gsu.edu`.

2.1 Lecture One: Intro to Data Science Recap

Data Science combines the disciplines of mathematics and computer science to gain meaningful insight from data sets.

We often use tables to organize the data we are studying. Each row represents a data feature (ranging from 1 to D) and each column represents a vector (ranging from 1 to N).

	Sample1	Sample 2	Sample 3	$\cdots$	Sample N
Feature 1					
Feature 2					
Feature 3					
$\vdots$					
Feature D					

Table 2.1: Data Table Layout

We normally use $\mathbf{X}$ to denote a $\mathrm{D} \times \mathrm{N}$ data matrix, and use $\mathbf{x}_\mathrm{i}$ to denote the i^{th} [column] vector in $\mathbf{X}$. We use $\mathbf{x}_\mathrm{i} \in \mathbb{R}^\mathrm{D}$ to indicate that $\mathbf{x}_\mathrm{i}$ is a vector with D real-valued entries.

2.2 Data Models

To compose a complete data model, we must select the most appropriate model (line, bar graph, etc) and model learning process. Determining the best model and learning process is dependent on the data that we are analyzing.There are two types of model learning processes- supervised learning and unsupervised learning.

(a) Supervised Learning

  (i) Uses labels (categorical data)

  (ii) Example: Multiple photographs are labeled by the subject's name. Our goal is to be able to identify the proper label for any new photograph.

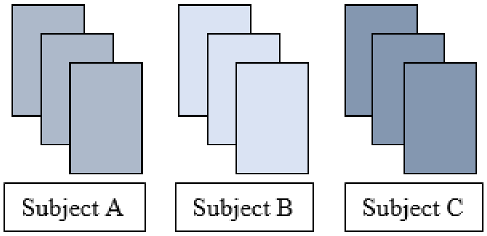

Figure 2.1: The photos are our data, and the subject names are the labels

(iii) Other applications include- identifying dog breed based on images, diagnosing diseases based on MRI scans, etc.

(iv) Algorithms for Supervised Learning:

- Logistic, Linear, and Polynomial Regression
- Support Vector Machine (SVM)
- K Nearest Neighbors
- Random Forest
- Neural Networks
- Decision Trees
- Naive Bayes

(b) Unsupervised Learning

(i) Does not use labels (data is uncategorized)

(ii) Example: We are given numerous shuffled images and we do not know who any of the subjects are. Our goal is to be able to identify that the 1st picture and the 7th picture are the same subject.

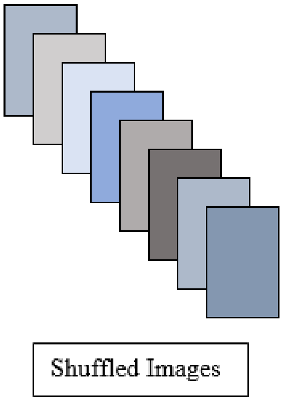

Figure 2.2: We can see that the far left image is the same as the image second from the right.

(iii) Another application for unsupervised learning is video surveillance.

(iv) Algorithms for Unsupervised Learning:

- K Means Clustering
- Hierarchical
- Expectation Management (EM)
- Principal Component Analysis (PCA)
- Correlation Analysis

2.3 Linear Algebra: Fundamentals

(a) A vector is a single column which contains D number of features. A vector is denoted with a bold, lowercase letter (i.e. $\mathbf{x}$). A non-bolded lowercase letter denotes a scaler value. Below is the standard notation for a vector of the i^{th} index with five dimensions (D values).

$$\mathbf{x_i} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

(b) A matrix is a collection of N number of vectors. A matrix is denoted with a bold, uppercase letter (i.e. $\mathbf{X}$). An individual element in a matrix's jth row and ith column is denoted (j,i). Below is the standard notation for a 4 by 5 matrix (four representing the number of features and 5 representing the number of vectors).

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} \\ x_{4,1} & x_{4,2} & x_{4,3} & x_{4,4} \end{bmatrix}$$

2.4 Linear Algebra: Operations

(a) To transpose a vector or a matrix means to switch its rows and columns. We use $\mathbf{x}^T$ or $\mathbf{X}^T$ to denote the transpose operation.

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \qquad \mathbf{x^T} = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \qquad \mathbf{X^T} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

(b) To trace a matrix means to take the sum of its diagonal elements (beginning at the top left element and going down to the bottom right). This operation is only applicable to square matrices.

The notation of the trace of a matrix is represented as $tr(\mathbf{X}) = \sum_{i=1}^{n} x_{ii} = x_{11} + x_{22} + ... + x_{nn}$

Example:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix} \qquad tr(\mathbf{X}) = \sum_{i=1}^{n} x_{ii} = 1 + 6 + 11 + 16 = 34$$

(c) A square matrix ($\mathbf{X}$) is said to be invertible if: $\mathbf{X} \cdot \mathbf{X^{-1}} = \mathbf{I}$

To invert a (2 x 2) matrix $\mathbf{X}$, swap the positions of $x_{1,1}$ and $x_{2,2}$, make $x_{1,2}$ and $x_{2,1}$ negative, and divide by the determinant value- which is computed as: $((x_{11} \cdot x_{2,2})) - (x_{1,2} \cdot x_{2,1})$.

$$\mathbf{X^{-1}} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \frac{1}{|\mathbf{X}|} \begin{bmatrix} x_{22} & -x_{12} \\ -x_{21} & x_{11} \end{bmatrix}$$

(d) We are able to multiply vectors and matrices by themselves or by scaler values.

(i) Vectors:
We multiply two vectors of the same size together to get the inner product (or dot product). Given two vectors, $\mathbf{x}$ and $\mathbf{y}$, we multiply $\mathbf{y}$ by $\mathbf{x^T}$ to get the value of $\mathbf{x} \cdot \mathbf{y}$.

$$\mathbf{x} = \begin{bmatrix} 4 \\ 7 \\ 2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}$$

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x^T} \cdot \mathbf{y}$$

$$\mathbf{x^T} = \begin{bmatrix} 4 & 7 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 4 & 7 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} = ((4 \cdot 3) + (7 \cdot 0) + (2 \cdot 1)) = 14$$

In scaler multiplication, we can multiply each element in a vector by the given scaler value to result in a new vector.

$$4 \cdot \begin{bmatrix} 0 \\ -6 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \cdot 0 \\ 4 \cdot -6 \\ 4 \cdot 2 \\ 4 \cdot 4 \end{bmatrix} = \begin{bmatrix} 0 \\ -24 \\ 8 \\ 16 \end{bmatrix}$$

(ii) Matrices:
It is only possible to multiply two matrices if the number of columns in the first matrix is equal to the number of rows in the second matrix.

$$\text{Not Applicable for Multiplication: } \begin{bmatrix} 4 & 6 & 14 & 8 \\ 15 & 0 & -3 & 62 \\ -8 & 1 & 0 & 7 \end{bmatrix} \cdot \begin{bmatrix} 5 & 0 & 11 & 79 \\ 6 & 22 & 3 & 6 \\ 34 & 45 & 0 & 58 \end{bmatrix}$$

$$\text{Applicable for Multiplication: } \begin{bmatrix} 17 & 28 \\ 3 & 65 \\ 0 & 21 \end{bmatrix} \cdot \begin{bmatrix} -3 & 0 & 61 \\ 49 & 29 & 3 \end{bmatrix}$$

Example:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} (1\cdot 7 + 2\cdot 10) & (1\cdot 8 + 2\cdot 11) & (1\cdot 9 + 2\cdot 12) \\ (3\cdot 7 + 4\cdot 10) & (3\cdot 8 + 4\cdot 11) & (3\cdot 9 + 4\cdot 12) \\ (5\cdot 7 + 6\cdot 10) & (5\cdot 8 + 6\cdot 11) & (5\cdot 9 + 6\cdot 12) \end{bmatrix} = \begin{bmatrix} 27 & 30 & 33 \\ 61 & 68 & 75 \\ 95 & 106 & 117 \end{bmatrix}$$

Scaler multiplication on matrices works the same as it does on vectors. Simply multiply each element of the matrix by the scaler value to result in a matrix.
Example:

$$2 \cdot \begin{bmatrix} -3 & 0 & 6 \\ 4 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 2\cdot -3 & 2\cdot 0 & 2\cdot 6 & \\ 2\cdot 4 & 2\cdot 2 & 2\cdot 3 & \end{bmatrix} = \begin{bmatrix} -6 & 0 & 12 \\ 8 & 4 & 6 \end{bmatrix}$$

2.5 Linear Algebra: Norms

A vector's norm is used as a means to quantify how big or small a vector is. There are different methods of finding a norm value.

General form of norm for vector $X \in R^D$ is $\|X\|_p = (\sum_{j=1}^{D} X_j^p)^{\frac{1}{p}}$

For case $p = 2$, it becomes Euclidean distance or the length the vector X in R^D $\|X\|_2 =\sqrt{\sum_{j=1}^{D} X_j^p}$

For case $p = 1$, it is the sum of length of each dimension. Hence, it is also called Manhattan distance. $\|X\|_2 = \sum_{j=1}^{D} |X_j|$

Matrix norm

L2 norm $\|X\|_2 :=$ largest eigenvalue of X Frobenius Norm $\|X\|_F :=\sqrt{\sum_{i=1} N \sum_{j=1}^{D} X_{ij}^2}$

In data science, the norm value plays a critical role in determining how different or how similar our data is.

2.6 Optimization Review

Basic terms:

$x^* = \underset{x}{\text{argmax}} f(x)$ means for all x in domain of function f, find value x^* such that $f(x^*)$ is the maximum.

$x^* = \underset{x}{\text{argmin}} f(x)$ means for all x in domain of function f, find value x^* such that $f(x^*)$ is the minimum.

Case 1: Simple convex/concave function with one optimal solution

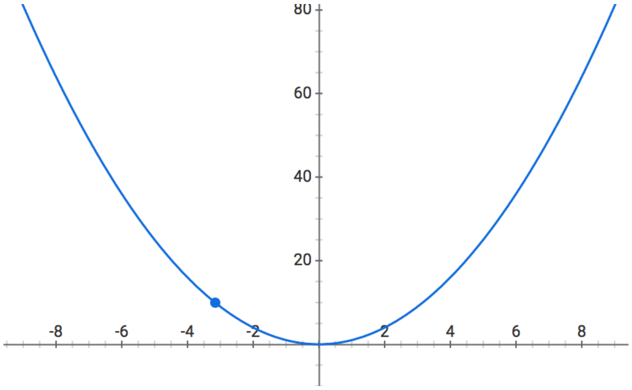

Figure 2.3: function $y = x^2$ has one minimum at $x = 0$

Steps to find optimal point for convex/concave function:

Step 1: Find $\frac{df}{dx}$

Step 2: Solve $\frac{df}{dx} = 0$

Case 2: Multivariable convex function

In this case, find minimum/maximum of f(X) where **X** is a matrix. $X* := \underset{X \in R^{DxN}}{\text{argmax}} f(X)$

Step 1: Calculate gradient

$$\nabla(f(X)) = \frac{\partial f}{\partial X} =< .., \frac{\partial f}{\partial X_i,}, .. > \tag{2.1}$$

Step 2: Set gradient to zero and find solution for the equation

$$\nabla(f(X)) = 0 \tag{2.2}$$

Gradient for multivariable function is complicated, there is some special cases such as:

$$\nabla(tr(AX)) = A \tag{2.3}$$

$$\nabla(tr(X^T AX)) = 2X^T A \tag{2.4}$$

Case 3: A function which has multiple critical points

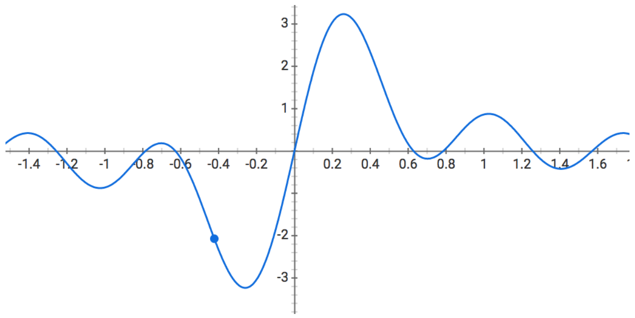

Figure 2.4: The function above has multiple solutions when its derivative is equal to zero

When the equation $\frac{df}{dx}(X) = 0$ has multiple solutions, finding the global minimum or maximum is an open problem with no exact solution. For one variable functions, the approximated function could be reconstructed by many sampling many points.

However, for multivariable functions with high dimensions, sampling is expensive. Instead, the gradient descent/ascent approach is used to find the optimal point in the case of convex/concave functions or the local optimal point in case of functions with multiple critical points. The idea is the gradient vector points to the steepest direction, then by going to the inverse direction of gradient vector, we ascend toward the local minimum.

$$X_{i+1} = X_i - \alpha * \nabla(f(X_i)) \tag{2.5}$$