STA4724: Big Data Analytics Methods

Daniel Palma

February 18, 2025

Course Update on February 6th, 2025

Dr. Tomova spoke with Dr. Ding, and had a discussion about the course. there needed to be a change but they all agreed to NOT change the course in the middle of the semester.

They agreed to figure out how to balance the pre-requisites, based on the name the course should include a lot of concepts of linear algebra, however they dont want it to become a barrier for students currently taking the course (they want students to keep attending) so they will be currently making it similar to previous semesters, consistent with the syllabus and include more practical applications, and less conceptual theory moving forward.

The last three topics, 10-12 (Projection and Isometry, Variance-Covariance Matrix, and Multivariate Normal Distribution) are not going to be covered and will be optional. today we're finishing up 8-9 (spectral decomposition, Properties and Derivations of Matrix Traces). The following homeworks and exams will not have any proof questions

Contents

1 Matrix Algebra

1.1 Definitions of Matrices and Vectors

Matrix

- a matrix is an arrangement of numbers in rectangular form
- a $J \times K$ matrix has J rows and k columns
- a Square matrix is of order $(2, 2)$ as a special case
- Vectors are subcategories of matrices that have either one row or one column

  $(1, k)$ is one rowm, and multiple columns, e.g. $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$

  $(k, 1)$ is one column, and multiple rows, e.g. $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$
- a matrix with one row and one column is the same as a scalar. $a = 5 \Leftrightarrow a = \begin{bmatrix} 5 \end{bmatrix}$

1.2 Addition, Subtraction, Multiplication

- $A + B = C$
- $A + B \Leftrightarrow B + A$
- $(A + B) + C \Leftrightarrow A + (B + C)$

Transposition An order (j, j) matrix is said to be symmetric if $A = A^T$

- $(A^T)^T \Leftrightarrow A$
- $(kA)^T \Leftrightarrow kA^T$ where k is a scalar
- $(A + B)^T \Leftrightarrow A^T + B^T$
- $kA \Rightarrow k \cdot \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \Rightarrow \begin{bmatrix} k \cdot a_1 & k \cdot a_2 & k \cdot a_3 \end{bmatrix}$
- Given matrix A of order (m, n) and matrix B of order (n, r)

  $C = A \cdot B$ is of order $(m, r) = \begin{bmatrix} C_{mr} \end{bmatrix}$ where $C_{mr} = \sum_{i=1}^{n} A_{mi} \cdot B_{ir}$

Example 1. Given the matrices $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ and $B = \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix}$, find $C = A \cdot B$

$$
\begin{aligned}
C_{11} &= 1 \cdot 7 + 2 \cdot 9 + 3 \cdot 11 = 58 \\
C_{12} &= 1 \cdot 8 + 2 \cdot 10 + 3 \cdot 12 = 64 \\
C_{21} &= 4 \cdot 7 + 5 \cdot 9 + 6 \cdot 11 = 139 \\
C_{22} &= 4 \cdot 8 + 5 \cdot 10 + 6 \cdot 12 = 154
\end{aligned}
$$

Therefore, $C = \begin{bmatrix} 58 & 64 \\ 139 & 154 \end{bmatrix}$

Properties

- $AB \neq BA$
- $A(BC) \Leftrightarrow (AB)C$
- $A(B + C) \Leftrightarrow AB + AC$
- $(AB)^T \Leftrightarrow B^T A^T$
- $A^n \Leftrightarrow A_0 \cdot A_1 \cdot ... \cdot A_{n-1}$

1.3 Diagonal and Identity Matrices

Diagonal matrix A diagonal matrix is a square matrix with zero entries except possible on the main diagonal

Example 2. $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ is a diagonal matrix. note that they dont need to be 1s, they can be any number, including zero.

In general, a diagonal matrix is given by $D_{mn} = \begin{bmatrix} d_1 & 0 & 0 & ... & 0 \\ 0 & d_2 & 0 & ... & 0 \\ 0 & 0 & d_3 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & d_n \end{bmatrix}$

Echelon Form

1. row echeleon form (ref)

   The first non-zero element in each row is called the leading entry, is always 1

   Each leading entry is in a column to the right of the leading entry in the previous row (if any)

   Rows with all zero elements are below rows with non-zero elements (if any)

2. reduced row echelon form (rref)

   any ref with the leading entry in each row is the only non-zero entry in its column.

Properties of Diagonal Matrices

- A diagonal matrix D is invertible if and only if all the diagonal elements are non zero.

Example 3. given $D^{-1} = \begin{bmatrix} 1/d_1 & 0 & ... & 0 \\ 0 & 1/d_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1/d_n \end{bmatrix}$

so $DD^{-1} = \begin{bmatrix} d_1 \cdot 1/d_1 & 0 & ... & 0 \\ 0 & d_2 \cdot 1/d_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & d_n \cdot 1/d_n \end{bmatrix} \to I$ which is the identity matrix

Identity Matrix The identity matrix is a square matrix, consisting of ones along the diagonal and zeros elsewhere. Typically, I is used to denote the identity matrix.

Example 4.

$$I_{nn} = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{bmatrix}$$

Properties of Identity Matrices

- $AI = IA = A$

Zero Matrix a zero matrix consists of all zero elements.

1.4 Determinant and Eigenstructure

Determinant Determinants are defined only for square matrices and scalars.

Example 5. let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $\det(A) = ad - cb$

the determinant of a matrix is denoted by $|A|$ or $\det(A)$ and is a number which encodes a lot of information about the matrix.

In general, we need to first define the cofactor $Q_{r,s}$ of each element of A, $a_{r,s}$. The cofactor of $a_{r,s}$ is $Q_{r,s} = -1^{r+s}|A_{r,s}|$ (where $|A_{r,s}|$ is the determinant of the matrix obtained by deleting the r-th row and s-th column of A).

The last step is to define the determinant of the matrix A as

$$|A| = \sum_{j=1}^{n} a_{ij} Q_{ij}$$

or

$$|A| = \sum_{i=1}^{n} a_{ij} Q_{ij}$$

Properties of Determinants

- $|I| = 1$
- if exchanging two rows of a matrix, we only need to reverse the sign of its determinant
- If we multiply one row of a matrix by a scalar k, the determinant is also multiplied by k.
- The determinant behaves like a linear function on the rows of the matrix

Lemma 6. *The geometric multiplicity of an eigenvalue is at most its algebraic multiplicity.*

Characteristic equation $det(A - \lambda I) = 0$

The Geometric Multiplicity of Eigenvalues

- It is the dimension of the linear space of its associated eigenvectors.

Let A be a $k \times k$ matrix, λ_k be one of the eigenvalues of A and denote its associated eigenspace by E_k. Then the dimension of E_k is called the geometric multiplicity of this eigenvalue λ_k

Proof. Suppose that the geometric multiplicity of λ_k is equal to g, so that there are g linearly independent eigenvectors. $x_1, ...x_g$ associated to λ_k. Randomly choose $k - g$ factors $x_{g+1}...x_k$, all having dimension $k \times l$ and such that the k column vectors $x_1, ..., x_k$ are linearly independent.

Define the $k \times k$ matrix

$$x = [x_1, ..., x_k]$$

for any g, denoted by b_g the vector that solves $xb_g = Ax_g = \lambda x_g$

Define the $k \times (k - g)$ matrix

$$B = [b_g + 1, ..., b_k]$$

and denote by C its upper $g \times (k-g)$ block, and denote by D its lower $(k-g) \times (k-g)$ block

$$B = \begin{bmatrix} C \\ D \end{bmatrix}$$

Denote by I the $k \times k$ identity matrix. for any scalar λ, we have that $(A - \lambda I)X$ $(= 0$ to find x for $\lambda_k \times)$

□

I dont understand anything she wrote after this so sorry that there isnt anything here

1.5 Inverses and Singularity

Suppose A is a square matrix. we look for an Inverse Matrix, A^{-1} of the same size, such that $AA^{-1} \Rightarrow 1$ (does nothing to a vector)

Thus, $AA^{-1}x = x$

Multiplying $Ax = b$ by A^{-1} gives $A^{-1}Ax = A^{-1}b \Rightarrow x = A^{-1}b$

If the determinant of A is non-zero, then A^{-1} exists, thus it is invertible.

I also dont know what she wrote for this sorry guys

1.6 Systems of Equations

Systems of linear equations. A system of K linear equations in L unknown variables is a set of equations of the form:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1L}x_L &= b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2L}x_L &= b_2 \\ &... \\ a_{K1}x_1 + a_{K2}x_2 + ... + a_{KL}x_L &= b_K \end{aligned}$$

and can be represented by the matrix equation $Ax = b$ where A is a $K \times L$ matrix, x is a $L \times 1$ vector, and b is a $K \times 1$ vector.

Definition 7. An Augmented Matrix is a matrix obtained by appending the columns of two matrices that have the same number of rows.

Let A be a $K \times L$ matrix, B is a $K \times M$ matrix. The augmented matrix of A and B is denoted by $[A|B]$ and is a $K \times (L + M)$ matrix. and is obtained by appending the columns of B to the right of A.

Homogeneous System The vector of constants on the right-hand side of the equation is zero. $Ax = 0$

Non-Homogeneous System The vector of constants on the right-hand side of the equation is not zero. $Ax = b$

By elementary row operations, a non-Homogeneous system can be transformed into an $Rx = b_R$ where the coefficient matrix R is in row echelon form. If R has a zero row R_i with $b_{Ri} \neq 0$, then the system has no solution.

Partitioned System Suppose that we have a $K \times L$ row ecehlon form matrix R with r basic columns and the last $L - r$ columns are non-basic. Partition the matrix into two blocks $R = [B \quad N]$ where B is a $K \times r$ matrix and N is a $K \times (L - r)$ matrix, Similarly partition the vector of variables into two blocks $x = [x_B \quad x_N]$ where x_B is a $r \times 1$ vector and x_N is a $(L - r) \times 1$ vector.

1.7 Singular Value Decomposition (SVD)

Calculation The columns of u are the eigenvectors corresponding to the nonzero values of MM^T

I did not get the rest of these notes

Proof. Since $MM^T = (u\Sigma v^t) \cdot (v\Sigma u^T)$ □

1.8 Spectral Decomposition

Calculation Let A be a square $n \times n$ matrix (λ_i, v_i).

$$A = \begin{bmatrix} v_1 & v_2 & ... & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ ... \\ v_n \end{bmatrix} = \sum_{i=1}^{n} \lambda_i v_i v_i^T$$

Cholesky Decomposition

$$A = VA^{\frac{1}{2}}(VA^{\frac{1}{2}})^T, A^{\frac{1}{2}} = diag(\lambda_1^{\frac{1}{2}}, ..., \lambda_n^{\frac{1}{2}}) = LL^*$$

$$A_{tv} = \sum_{u=1}^{k} L_{tu}(L^*)_{uv} = \sum_{u=1}^{k} L_{tu}L_{vu}$$

i didnt get the rest

1.9 Properties and Derivations of Matrix Traces

Vector Norms For any vector $x \in \mathbb{R}^n$, the l_2 norm of x is given by

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Inner Product for any $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, the angle θ can be computed by

$$\cos(\theta) = \frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1^2 x_2^2}\sqrt{y_1^2 y_2^2}}$$

let $\langle x, y \rangle$ be the inner product of x and y, then

$$\langle x, y \rangle = x^T y$$

where

$$\langle x, y \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} y_{ij}$$

Matrix Trace For any square matrix $x \in \mathbb{R}^{n \times n}$, the matrix trace $tr()$ is the sum of the diagonal entries of

$$tr(x) = \sum_{i=1}^{n} x_{ii}$$

where $x_{ii}, \forall i \in [n]$ is the (i,i)-th entry of x. Thus, $tr(x) = tr(x^T)$

Properties of Matrix Trace

- $tr(x + y) = tr(x) + tr(y)$
- $tr(\alpha x + \beta y) = \alpha tr(x) + \beta tr(y)$
- $tr(xy) = tr(yx)$
- $tr(x^T x) = ||x||_F^2$

1.10 Projection and Isometry

OPTIONAL - SKIPPED

1.11 Variance-Covariance Matrix

OPTIONAL - SKIPPED

1.12 Multivariate Normal Distribution

OPTIONAL - SKIPPED

2 Machine Learning

Artificial Intelligence In the broadest sense, it is intelligence exhibited by machines & computer systems.

High Profile appliations of AI include

- Advanced Web Search Engines
- Recommendation Systems
- Virtual Assistants
- Autonomous Vehicles
- Generative AI (ChatGPT)

Machine Learning is a subset of Artificial Intelligence (AI) that focuses on algorithms that allows computer to learn from data and improve their performance over time without being explicitly programmed, It leverages the processing power to analyze vast amounts of information that traditional methods wouldnt handle efficiently.

Deep Learning is a subset of Machine Learning (ML) that specifically utilizes artificial neural networks to process complex data sets.

Machine Learning with Big Data The Observed Input-Output mapping

2.1 Two-Layer Neural Networks

The Basic problem

$$\{(y_i, x_i)\}_{i \leq i \leq n}, iid$$

joint distribution $\mathbb{P}$

Without a loss of generality, we can describe the relationship between y and x as

$$y_i = f(x_i) + z_i$$

Example 8. Fitting a polynomial of degree k.

$$\hat{f}(x, \theta) = \Sigma_{\alpha:|\alpha| \leq k} \quad \theta_{\alpha_1}, ..., \alpha_d x_1^{\alpha_1}, ..., x_d^{\alpha_d}, |\alpha| = \Sigma_{i \leq d} \alpha_i$$

$$R(\theta) = E\{[f(x) - \hat{f}(x, \theta)]^2\} + \sigma^2 = E\{[y - \hat{f}(x, \theta)]^2\}$$

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{f}(x_i, \theta)]^2$$

Simple neural network model

$$\hat{f}(x,\theta) = \frac{1}{n}\sum_{i=1}^{n} \sigma_*(x,\theta_i)$$

where

$$\begin{cases} N: \text{ Number of hidden units (neurons)} \\ \sigma_* : \mathbb{R}\times\mathbb{R}^{\mathbb{D}} \to \mathbb{R} \text{ is an activation function.} \\ \theta_i \in \mathbb{R}^{\mathbb{D}} \text{ are the parameters of the model. } \theta = (\theta_1, ..., \theta_n) \end{cases}$$

Example 9. Sigmoid function $\sigma(x) = \frac{1}{1+e^{-2x}}$
Rectified Linear Unit $\sigma(x) = max(x, 0)$

Theorem 10. *Assume* $E(f(x)^2) < \infty$*. and assume* $\sigma : \mathbb{R} \to \mathbb{R}$ *is continuous with* $\lim_{x\to\infty} \sigma(x) \to 1$*.* $\lim_{x\to-\infty} \sigma(x) \to 0$*. Then, for any* ϵ*, there exists* $N = N(\epsilon)$

$$inf_{(a_i,b_i,w_i)} E\{[f(x) - \frac{1}{n}\sum_{i=1}^{n} \sigma(\langle w_i, x\rangle + b_i)]\} \leq \epsilon$$

Proof. SKIPPED □

Theorem 11. *Let*

$$f(x) = \int e^{i\langle w,z\rangle} F(w)\mathrm{d}w$$

$$\sigma : \mathbb{R} \to \mathbb{R} \textit{ with} \lim_{t\to\infty} \sigma(t) = 1, \lim_{t\to-\infty} \sigma(t) = 0$$

There exists a network of the form

$$\mathcal{L} = \int \frac{1}{n}\Sigma\alpha_i\sigma(\langle w, x\rangle + b_i) : n \in \mathbb{N}, \alpha_i, b_i \in \mathbb{R}, w_i \in \mathbb{R}^D$$

with

$$N(\epsilon) = \frac{1}{\epsilon}(2r\int ||w||_2 |F(w)| dw)^2$$

achieving error

$$E\{[f(x) - \mathcal{L}(x)]^2\} \leq \epsilon$$

Stochastic Gradient Descent Goal: minimize $R(\theta)$

Algorithm: $\theta^{k+1} = \theta^k + s_k v_k$ where S_k is the step size and v_k is the direction of the descent, $n_k = -\nabla R(\theta^k), v_k = -\nabla R(\theta^k) + Z^k$