Week 2 - Matrix Review, SVD and PCA

Matrix Review

1. Basic concepts and notation

Consider the following system of equations:

$4x_1 - 5x_2 = -13$

$-2x_1 + 3x_2 = 9$

In matrix notation, we can write the system more compactly as:

$Ax = b$

with

$A = \begin{bmatrix} 4 & 5 \\ -2 & 3 \end{bmatrix}, b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$

- $A \in \mathbb{R}^{m \times n}$: a matrix with m rows and n columns, where the entries of A are real numbers.
- $x \in \mathbb{R}^n$: a vector with n entries. By convention, a n-dimensional vector is ofter thought of as a matrix with n rows and 1 column (column vector). If we want to explicitly represent a row vector, we write x^T.

2. Matrix multiplication

$A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$

$C = AB \in \mathbb{R}^{m \times p}$

where

$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$

Matrix multiplication is:

- associative: $(AB)C = A(BC)$
- distributive: $A(B + C) = AB + AC$
- not commutative: $AB \neq BA$

3. Operations and Properties

3.1 Identity matrix

$I \in \mathbb{R}^{n\times n}$, a square matrix with ones on the diagonal and zeros everywhere else.

For all $A \in \mathbb{R}^{m\times n}$,

$AI = A = IA$

3.2 Transpose

Given a matrix $A \in \mathbb{R}^{m\times n}$, its transpose, written $A^T \in \mathbb{R}^{n\times m}$.

Properties:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A+B)^T = A^T + B^T$

3.3 Trace

The trace of a square matrix $A \in \mathbb{R}^{n\times n}$ is denoted as $tr(A)$, which is the sum of diagonal elements in the matrix:

$trA = \sum_{i=1}^{n} A_{ii}$

Properties:

- $trA = trA^T$, for $A \in \mathbb{R}^{n\times n}$
- $tr(A+B) = trA + trB$, for $A, B \in \mathbb{R}^{n\times n}$
- $tr(tA) = t\, trA$, for $A \in \mathbb{R}^{n\times n}, t \in \mathbb{R}$
- $trAB = trBA$ for A, B such that AB is square.
- $trABC = trBCA = trCAB$ for A, B, C such that ABC is square.

3.4 Norms

A norm of a vector is informally a measure of the "length" of the vector.

ℓ_p norms:

$||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$

ℓ_2 norm (Euclidean):

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

Norms can also be defined for matrices, such as the Frobenius norm,

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} = \sqrt{tr(A^T A)}$$

3.5 Rank

A set of vectors $\{x_1, x_2, \ldots x_n\} \subset \mathbb{R}^m$ is said to be (linearly) independent if no vector can be represented as a linear combination of the remaining vectors.

For example, the vectors

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix}, x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

are linearly dependent because $x_3 = -2x_1 + x_2$

properties:

- For $A \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(A) \leq \min(m, n)$. If $\mathrm{rank}(A) = \min(m, n)$, then A is said to be full rank.
- For $A \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(A) = \mathrm{rank}(A^T)$
- For $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, $\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B))$
- For $A, B \in \mathbb{R}^{\mathrm{m\times n}}$, $\mathrm{rank}(A + B) \leq \mathrm{rank}(A) + \mathrm{rank}(B)$

3.6 Inverse

The inverse of a square matrix $A \in \mathbb{R}^{n\times n}$ is denoted as A^{-1}, and is the unique matrix such that,

$$A^{-1}A = I = AA^{-1}$$

*Note: not all matrices have inverse. e.g. non-square matrices.

We say that A is invertible or non-singular if A^{-1} exists and non-invertible or singular otherwise.

Properties:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$

3.7 Orthogonal Matrices

Two vectors $x, y \in \mathbb{R}^n$ are orthogonal if $x^T y = 0$.

A square matrix $U \in \mathbb{R}^{n\times n}$ is orthogonal if all its columns are orthogonal to each other and are normalised (the columns are then referred to as being orthnormal)

$$U^T U = I = UU^T$$

If U is not square ($U \in \mathbb{R}^{m\times n}, n < m$), but its columns are still orthonormal, then $U^T U = I$, but $UU^T \neq I$.

Property:

Operate on a vector with an orthogonal matrix will not change its Euclidean norm, i.e.,

$$||Ux||_2 = ||x||_2$$

3.8 Determinant

The determinant of a square matrix $A \in \mathbb{R}^{n\times n}$, is a function det: $\mathbb{R}^{n\times n} \rightarrow \mathbb{R}$, and is denoted $|A|$ or det A.

$$|[a_{11}]| = a_{11}$$

$$\left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right| = a_{11}a_{22} - a_{12}a_{21}$$

$$\left|\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right| = a_{11}(a_{22}a_{33} - a_{23}a_{32}) + a_{12}(a_{23}a_{31} - a_{21}a_{33}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

3.9 Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n\times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of A and $x \in \mathbb{C}^n$ is the corresponding eigenvector if:

$$Ax = \lambda x,\ x \neq 0$$

which means, multiplying A by the vector x results in a new vector that points in the same direction as x, but scaled by a factor λ.

Rewrite the equation to state that (λ, x) is an eigenvalue-eigenvector pair of A if

$(\lambda I - A)x$, $x \neq 0$

But $(\lambda I - A)x = 0$ has a non-zero solution to x if and only if $(\lambda I - A)$ has a non-empty nullspace, which is only the case if $(\lambda I - A)$ is singular, i.e.,

$|(\lambda I - A)| = 0$

Properties:

- The trace of A is equal to the sum of its eigenvalues, $trA = \sum_{i=1}^{n} \lambda_i$
- The determinant of A is equal to the product of its eigenvalues, $|A| = \prod_{i=1}^{n} \lambda_i$
- The rank of A = the number of non-zero eigenvalues of A

http://cs229.stanford.edu/summer2019/cs229-linalg.pdf

Singular Value Decomposition (SVD)

Assume $A \in \mathbb{R}^{n \times p}$

SVD is a method of decomposing a matrix into three other matrices:

$A = USV^T$

where:

$U \in \mathbb{R}^{n \times n}$, $S \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{p \times p}$

$U^TU = I$, $V^TV = I$ (i.e. U and V are orthogonal)

Where the columns of U are the left singular vectors. S is diagonal, and V^T has rows that are the right singular vectors. The SVD represents an expansion of the original data in a coordinate system where the covariance matrix is diagonal.

Calculate SVD:

- finding eigenvalues and eigenvectors of AA^T and A^TA
- The eigenvectors of A^TA make up the columns of V.
- The eigenvectors of AA^T make up the columns of U.
- The singular values in S are square roots of eigenvalues from AA^T or A^TA. The singular values are the diagonal entries of the S and are arranged in descending order.

Principal Component Analysis (PCA)

PCA is a technique widely used for dimension reduction, data compression, feature extraction and data visualisation.

Two equivalent definitions of PCA:

- Maximum Variance Formulation:

  project the data onto a lower dimensional space such tat the variance of the projected data is maximised.

- Minimum Error Formulation:

  project the data onto a lower dimensional space such that the mean squared distance between data points and their projections (average projection cost) is minimised.

Algorithm

- Step 1: substract the mean data $\bar{x}$ from original data, i.e. $\boldsymbol{z} = \boldsymbol{x} - \bar{\boldsymbol{x}}$
- Step 2: compute the scatter matrix $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T$
- Step 3: compute the eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$ and eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ by using SVD of S. Then $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k]$ is the projection matrix
- Step 4: $y_i = \boldsymbol{U}^T \boldsymbol{x}_i$

How to determine k?

- Percentage of variance retained: $P(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq t$

*Proof in the formulation of Maximum variance formulation

Let $\boldsymbol{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{1 \leq i \leq n}$ be a set of observations.

Goal: project $\boldsymbol{X}$ onto a k dimensional subspace ($k < d$) such that the variance of the projected data is maximised.

Proof for $k = 1$:

Let $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ and $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$

Let $\boldsymbol{u}_1$ be the basis of the 1 dimensional subspace, and $\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$

$\hat{\mathrm{VAR}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_1^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}}))^2 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_1^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{u}_1 = \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1$

The problem becomes:

$\max\limits_{\boldsymbol{u}_1} \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1$

s.t. $\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$

This is equivalent to:

$\min\limits_{\boldsymbol{u}_1} \; -\boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1$

s.t. $\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$

Rewrite into Lagrangian function:

$\mathcal{L} = -\boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 + \lambda_1(\boldsymbol{u}_1^T\boldsymbol{u}_1 - 1)$

According to KKT conditions,

$\frac{\partial \mathcal{L}}{\partial \boldsymbol{u}_1} = -\boldsymbol{S}\boldsymbol{u}_1 + \lambda_1\boldsymbol{u}_1 = 0$

$\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1$

Thus, $\boldsymbol{u}_1$ is a eigenvector of $\boldsymbol{S}$ and λ_1 is its eigenvalue. Note that

$-\boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 = -\lambda_1\boldsymbol{u}_1^T\boldsymbol{u}_1 = -\lambda_1$

λ_1 is the largest eigenvalue of $\boldsymbol{S}$.