Machine Learning Foundations
Week 3

Shubharup G.
BSDS, IIT Madras

October 2023

Contents

1 Visualizing linear subspaces

A linear subspace can be **visualized** as a **straight or flat** geometrical object (in general *non-curved*) that is necessarily **(1) infinite and (2) passing through the origin**.

Some examples:

1. $dim(S) = 0 \implies S$ is a point
2. $dim(S) = 1 \implies S$ is a line
3. $dim(S) = 2 \implies S$ is a plane

The dimension of a subspace is not to be confused with the dimension of the co-ordinate system in which vectors are represented. For example, taking a vertex on the ceiling of your room as the origin in $\mathbb{R}^3$, *all* points in your room have a co-ordinate representation with three dimensions i.e. three co-ordinates (x, y, z); but all the points on *the ceiling* have the form $(x, y, 0)$ i.e. with only 2 independent parameters ('free variables'), and so live in a 2-dimensional subspace aka. a plane.

A note to the reader

The algebraic and geometric perspectives are two different perceptual modes when studying Linear Algebra. Although you'll probably have to switch between them as necessary, try to connect and develop them both, in both directions e.g. algebra to be interpreted geometrically, and geometric intuition to inform the results to be shown with the formal rigour of algebra.

2 Rank and Nullity

2.1 What is rank of a matrix?

1. The **row rank** of a matrix $A \in \mathbb{M}_{m\times n}$ is the **number of linearly independent rows**. These linearly independent rows taken together is the basis of the row space of A i.e. of $\mathbb{R}(A)$. Equivalently, the **row rank is dimension of the row space** of the matrix A. Note that $rowrank \leq m$ where m is the number of rows.

$$\boxed{rowrank = dim(\mathbb{R}(A))}$$

2. The **column rank** of a matrix $A \in \mathbb{M}_{m\times n}$ is the **number of linearly independent columns**. These linearly independent columns taken together is the basis of the column space of A i.e. of $\mathbb{C}(A)$. Equivalently, the **column rank is dimension of the column space** of the matrix A. Note that $colrank \leq n$ where n is the number of columns.

$$\boxed{colrank = dim(\mathbb{C}(A))}$$

3. It is a fundamental theorem in Linear Algebra that:

$$\boxed{\text{row rank} = \text{column rank}}$$

This is called the **rank of the matrix** $= r$

2.2 What is nullity of a matrix?

1. The **null space** $\mathbb{N}(A)$ is the subspace containing all vectors: v such that $Av = \vec{0}$.

2. The **nullity** is the dimension of the null space viz.

$$\boxed{nullity = dim(\mathbb{N}(A)}$$

2.3 Rank-Nullity Theorem:

$$\boxed{\textbf{Rank(A) + Nullity(A) = n}}$$

Recall that a matrix $A \in \mathbb{M}_{m\times n}$ can be interpreted as a 'vector-to-vector' function $A : \mathbb{R}^n \to \mathbb{R}^m$ i.e. as sending input vectors in $\mathbb{R}^n$ to output vectors in $\mathbb{R}^m$.

Actually, since $Ax = v$ (say) is a linear combination of the columns of A, you can see that $v \in \mathbb{C}(A)$. So **the matrix A sends input vectors to the column space of A** i.e. the range or *image* of A: $Im(A) = \mathbb{C}(A)$, $\mathbb{C}(A) \subseteq \mathbb{R}^m$.

With respect to A, the domain of input vectors aka input space, viz. $\mathbb{R}^n$ is "'made up"[1] of two parts – two orthogonal subspaces *whose dimensions add to n* [2], and so aptly called ***orthogonal complements***:

1. **Null space** $\mathbb{N}(A)$**:**

   All vectors in this subspace get mapped to $\vec{0}$ in $\mathbb{R}^m$. Note that $\vec{0}$ belongs to every subspace, in particular to $\mathbb{C}(A)$.

2. **Row space** $\mathbb{R}(A)$**:**

   All vectors in this subspace get mapped to some non-zero vector in $\mathbb{C}(A) \subseteq \mathbb{R}^m$.

[1] Meaning that any vector $v \in V$ can always be written as the sum of two components, viz. its projections on (any choice of) a pair of orthogonal complements of V. See Section 4.3.1.

[2] In equivalent words, their Direct Sum is $\mathbb{R}^n$. See Direct Sum (Wikipedia).

2.4 How to calculate Rank

1. Use **Gaussian elimination** and find the **row-echelon form (REF)** of the matrix.

Rank = No. of pivot columns in REF = r
Nullity = No. of free variables in REF = n - r

2.5 Upper bound on Rank

$$\boxed{rank(A) \leq min(m, n)}$$

The rank can never be larger than the number of rows,
and it can never be larger than the number of columns either.

3 Four Fundamental Subspaces

3.1 Row Space and Null Space

$$\boxed{R(A) \perp N(A)}$$

The row space is perpendicular to the null space.

3.2 Column Space and Left Null Space

$$\boxed{C(A) \perp N(A^T)}$$

The column space is perpendicular to the left null space.

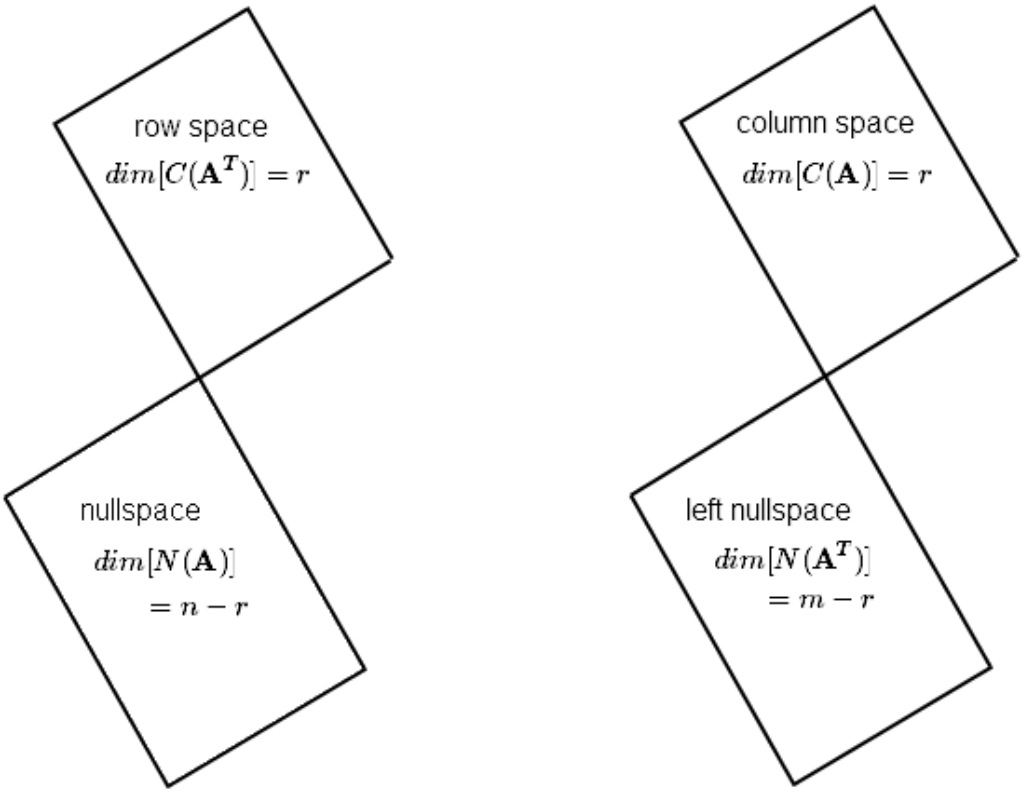

Figure : Four Fundamental Subspaces

4 Projection

Introduction

Motivation

It can be shown that finding the best fit line predictions obtained in Linear Regression by minimizing sum of squared error loss, is equivalent to projection of the label vector y onto the subspace spanned by the feature vectors.

4.1 What is a projection?

Asking for the projection of a vector $\vec{b}$ onto a vector $\vec{a}$ is asking:

What is a vector along $\vec{a}$ that is closest to $\vec{b}$?

Similarly, asking for the projection of a vector $\vec{b}$ onto a subspace U is asking:

What is a vector in the subspace U that is closest to $\vec{b}$?

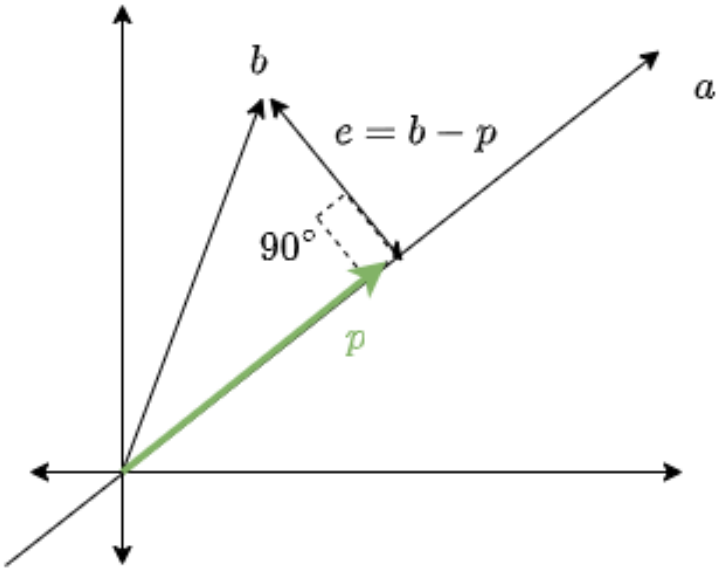

Figure : Projection onto a vector

4.2 Projection onto a vector

$$\boxed{\boldsymbol{p = proj_a(b) = (\frac{a^T b}{a^T a})a}}$$

The above is of the form
(scaling co-efficient $\times$ the vector to be projected on).

Alternatively, you can write it as:

$$\boxed{\boldsymbol{p = proj_a(b) = (\frac{aa^T}{a^T a})b}}$$

The above is in the **projection matrix** form

where the projection matrix

$$\boxed{\mathbb{P}_a = (\frac{aa^T}{a^T a})}$$

is of general relevance,
and will be encountered again in the subsequent section.

Observations:

- For a projection matrix $\mathbb{P}_a$ for a non-zero vector $\vec{a}$, $rank(\mathbb{P}_a) = 1$. [3]
  - The i^{th} row of aa^T is a^T scaled by the the i^{th} entry of $\vec{a}$.

[3]The **dyadic product** of two non-zero vectors $u, v \neq \vec{0}$ is a 2-nary operation defined as $u \otimes v = uv^T$. The product uv^T is a matrix, and $rank(uv^T) = 1$. Such a matrix is called a *dyadic matrix*, or simply a 'dyad'. See chapter on Week 6 for additional relevance: viz. the $k-$rank approximation of a matrix using SVD is a truncated k-sum of dyads.

– To see,

if $\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$, then $aa^T = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \begin{bmatrix} — & a^T & — \end{bmatrix} = \begin{bmatrix} — & (a_1)a^T & — \\ — & (a_2)a^T & — \\ & \vdots & \\ — & (a_n)a^T & — \end{bmatrix}$

– So the rows in aa^T are all just scalar multiples of the same vector viz. a^T, so (if $\vec{a} \neq \vec{0}$) the rank of aa^T is 1.

– $a^T a$ is a scalar (non-zero if a is non-zero), and dividing any matrix, in particular aa^T by a (non-zero) scalar does not change its rank.

4.3 Projection onto a subspace

4.3.1 Subspace with known orthogonal basis

Suppose $U = span(u_1, u_2...u_k) = span(B)$ (say) and $u_i \perp u_j$ i.e. ${u_i}^T u_j = 0 \; \forall i \neq j$ i.e. B is a linearly independent spanning set a.k.a a basis.

Then:

$$\boxed{\boldsymbol{p = proj_U(b) = \sum_{u_i \in B} (\frac{u_i^T b}{u_i^T u_i}) u_i}}$$

Projection onto a subspace with known orthogonal basis
is the sum of the projections on the basis elements.

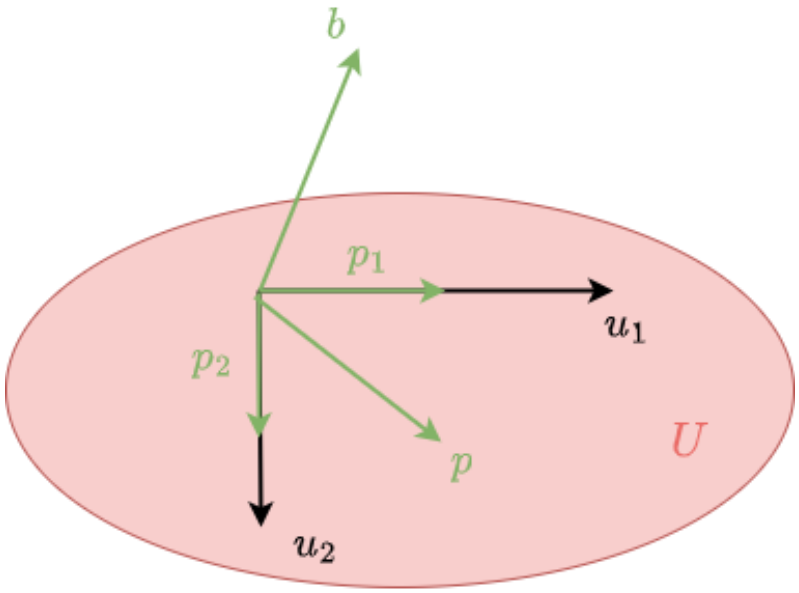

Figure : Projection onto a subspace with known basis

Observations:

- U is in a co-ordinate system of some dimension, say an n-dimensional co-ordinate system: i.e. $U \subseteq \mathbb{R}^n$. If in fact $span(u_1, u_2...u_k)$ is the whole $\mathbb{R}^n$ i.e. $U = \mathbb{R}^n$, then the 'projection' is just a change of basis.
  - Projection onto orthogonal complements of $\mathbb{R}^n$ is a special case of this.

4.3.2 Projection onto column space of a matrix

$$\boxed{\boldsymbol{p = proj_{C(A)}(b) = [A(A^TA)^{-1}A^T]b}}$$

where the projection matrix

$$\boxed{\boldsymbol{\mathbb{P}_{C(A)} = [A(A^TA)^{-1}A^T]}}$$

Observations:

- When A is just a column vector, the projection matrix can be seen to be identical to the projection matrix in Section 2.1, as projecting onto the column space is just projecting onto the one single column vector.
  - This is valid in general when $dim(C(A)) = 1$ irrespective of the number of columns, in which case a is any basis vector for the 1-dimensional i.e. linear column-space of A.
- For a projection matrix $\mathbb{P}_{C(A)}$ for the column space $C(A)$ of a matrix A: $rank(\mathbb{P}_{C(A)}) = dim(C(A)) = rank(A)$.

4.4 Properties of the projection matrix $\mathbb{P}$

1. **Symmetric**: $\mathbb{P} = \mathbb{P}^T$
2. **Idempotence**: $\mathbb{P}^2 = \mathbb{P}$: *re-projecting the projection leaves it unchanged.*

5 Linear Regression as Projection

Let X be the data matrix, with data points stacked as columns. If $X \in \mathbb{M}_{m\times n}$ then this is to be interpreted as saying there are n datapoints (no. of columns) each with m features (no. of rows).

We have a label vector $y \in \mathbb{R}^n$ which is the set of actual i.e. empirically observed outputs for each of the n datapoints.

$w \in \mathbb{R}^m$ is a vector of weights for each of the m features. For each datapoint, the predicted output is a weighted sum of its feature values.

$\hat{y_j} = \sum_{i=1}^{m} w_i y_{ji} = {y_j}^T w$

The vector of all the predicted values $\hat{y}$ can be written as:

$$\boxed{\hat{y} = X^T w}$$

The squared error loss is given by:

- In simple sum form:

$$\boxed{\mathcal{L}(w) = \sum_{j=1}^{n} (\hat{y_j} - y_j)^2 = \sum_{j=1}^{n} ({y_j}^T w - y_j)^2}$$

- In vector form:

$$\boxed{\mathcal{L}(w) = ||\hat{y} - y||^2 = ||X^T w - y||^2}$$

Then it can be shown *directly* with the gradient-based method of optimization i.e. by setting $\nabla_w \mathcal{L}(w) = 0$ that:

1. The optimal w^* which minimizes $\mathcal{L}(w)$ is $w^* = (XX^T)Xy$; and so . . .
2. The optimal predictions $\hat{y}$ are:

$$\boxed{\hat{y} = (X^T w^*) = [X^T (XX^T)^{-1} X]y}$$

By directly, it is meant that 1. and 2. above are derived without any reference to the idea of projections discussed in the previous sections.

But observe the form of the R.H.S of 2. Think of the projection matrix in Section 2.2.2, with $A = X^T$ and $b = y$.

That is to say, the vector of (optimal / error-minimizing) predicted outputs is a projection of the vector of actual (label) outputs onto the column space of X^T i.e. onto the subspace spanned by the feature vectors of the data matrix.

It is deserving of a moment for appreciation, that this is a rather non-trivial finding[4] : this establishes a **geometric interpretation of linear regression**, viz. as a projection.

[4]As elementary as it may seem, that is only after the fact has been laid as a credible certainty, eliminating that strength of persuasion which was historically necessary to prove it for the first time in knowledge, as a fresh addition to the extent of human knowledge. Furthermore, if additionally the steps have been laid out in clear view. Epistemic advancement, with mathematical discovery as a special case, may be better appreciated, in both its process and outcomes, upon learning to draw the distinction between prospective and retrospective predictability (latter terms as used by N.N. Taleb).

References

Wikipedia

Wikipedia contributors. *Direct Sum (Wikipedia)*. URL: https://en.wikipedia.org/w/index.php?title=Direct_sum.