Introduction to Machine Learning (67577)

Recitation 01
Linear Algebra

Second Semester, 2021

Contents

1 Linear Algebra

1.1 Linear Transformations

Definition 1.1 — Linear Transformation. Let $V \in \mathbb{R}^d$ and $W \in \mathbb{R}^m$ be two vectors spaces. A function $T : V \to W$ is called a linear transformation of V into W, if $\forall u, v \in V$ and $c \in \mathbb{R}$.

- Additivity: $T(u+v) = T(u) + T(v)$
- Scalar multiplication: $T(cu) = cT(u)$

For V and W of a finite dimension, any linear transformation can be represented by a matrix A. Therefore, from now and on we will focus only on finite-dimensional spaces, and implicitly refer to the matrix representing the linear transformation.

Definition 1.2 — Affine Transformation. An *affine transformation* is a transformation of the form $T(u) = Au + w$, where $u \in V, w \in W$.

Notice, that by definition an affine transformation is not a linear transformation. Notice that for a linear transformation A it holds that $A \cdot 0_V = 0_W$, but in the case of an affine transformation where $0 \neq w \in W$ then $T(0_V) = A \cdot 0_V + w \neq 0_w$.

Let us define some vector spaces associated with each linear transformation

Definition 1.3 Let A be the matrix corresponding the linear transformation $T : V \to W$. We define the:

- Kernel- (or null-) space of A as $Ker(A) := \{x \in V | Ax = 0\}$. Also denotes as $N(A)$.
- Image- (or column-) space of A as $Im(A) := \{w \in W | w = Ax, x \in V\}$. Also denotes as $Col(A)$.
- Row space of A as $Im\left(A^\top\right) := \left\{x \in V | x = A^\top w, w \in W\right\}$. Equivelantly it can be defined as the column space of $A^\top$ and therefore denoted as $Col\left(A^\top\right)$.
- Null space of $A^\top$ as $Ker\left(A^\top\right) := \left\{x \in W | A^\top x = 0\right\}$. This space is also referred to as the left null space of A.

Note that by definition, $Ker(A), Row(A) \subseteq V$ and $Im(A) \subseteq W$. Using the above definitions let us gain some insights into what these vector spaces provide us with.

Definition 1.4 Let $A \in \mathbb{R}^{m \times d}$. The rank of A is the maximum number of linearly independent rows of A and denoted by $rank(A)$.

It holds that the rank of A equals both the dimension of the columns space and of the row space of A. As such, we refer to A being of *full rank* if and only if $rank(A) = min(m, d)$. Otherwise we say that A is rank deficient.

Definition 1.5 Let $A \in \mathbb{R}^{d \times d}$ be a square matrix. A is called invertible (or non-singular) if there exists a matrix $B \in \mathbb{R}^{d \times d}$ such that $AB = I_d = BA$. We denote the inverse by A^{-1}.

Claim 1.1 Let A be a square matrix. The following are equivalent (TFAE):
- A is invertible (non-singular)
- A is full-rank
- $Det(A) \neq 0$
- $Im(A) = \mathbb{R}^m$ (i.e., the image is the whole space)
- $ker(A) = \vec{0}$

■ **Example 1.1** Consider the following scenario: Suppose we are given a set of d linearly independent linear equations, each of the form $y_i = \sum_{j=1}^{d} \mathbf{w}_j \cdot x_{ij}$, where the $x_{i,j}$'s and y_i are given while $\mathbf{w}_j$'s are unknown. We would like to find a solution for this system of equations. That is, a coefficients vector $\mathbf{w} \in \mathbb{R}^d$ that satisfies:

$$\forall i \in [d] \ y_i = \sum_{j=1}^{d} \mathbf{w}_j \cdot x_{ij} = \mathbf{w}^\top x_i$$

Let us rearrange the equations in matrix form. Given a linear equation we will denote all it's x's by the vector $x_i \in \mathbb{R}^d$ where i denotes the numbering of the current equation. Similarly we will arrange all the y's in a vector $y \in \mathbb{R}^d$. Thus, we can represent the problem written above as follows:

$$\text{Find } \mathbf{w} \in \mathbb{R}^d \text{ such that } y = X\mathbf{w}$$

As we assumed that all linear equations are independent, the rows of X are linearly independent. Therefore, it is of full rank and there exists an invertible matrix X^{-1} such that $XX^{-1} = I$. Equipped with this observation finding $\mathbf{w}$ is simply:

$$y = X\mathbf{w} \ \Rightarrow \ X^{-1}y = X^{-1}X\mathbf{w} \ \Rightarrow , \mathbf{w} = X^{-1}y$$

■

R Let us think of each vector $x_i \in \mathbb{R}^d$ as some independent observation (or sample) we have of some phenomena. Each coordinate of x_i corresponds some measurement we have of this observation. Together with this sample we are given some response value $y_i \in \mathbb{R}$. By solving for $\mathbf{w}$ we learn the relation between the x's and y's. Now suppose we are given a new sample $x \in \mathbb{R}^d$. As we already know the relation between the xs and the ys, we can predict what is the appropriate y value it achieves.
The general problem of finding such vectors is called **Regression**. In the case where the relationship is linear it is called **Linear Regression**. We will discuss linear regression in **??**.

1.2 Norms, Inner Products and Projections

More many applications in machine learning we are interested in measuring distances between vectors or sizes of vectors, and "using" a vector (or set of vectors) on another vector. For such, let us formulate these notions.

Definition 1.6 — Metric. A function on a set $X \subseteq \mathbb{F}^k$ $d : X \times X \to \mathbb{R}_+$ is called a metric function (or distance function) iff for any $v, u, w \in X$ it holds that:

- $d(v,u) = 0 \iff v = u$
- Symmetry: $d(v,u) = d(u,v)$
- Triangle inequality $d(v,u) \leq d(v,w) + d(w,u)$.

These conditions also imply that a metric is non-negative. As such, we also call a metric function a positive-definite function. Some common metric functions are the absolute distance or the Euclidean distance.

Exercise 1.1 Let $v, u \in \mathbb{R}^k$. Show that the absolute distance, defined as the sum of absolute element-wise subtraction betwen the vectors $d(v,u) := \sum |v_i - u_i|$, is a metric function. ■

Proof. Firstly, notice that for some scalars $a, b \in \mathbb{R}$ it holds that $|a-b| = 0$ iff $a = b$. Therefore d, being a sum of non-negative elements equals zero iff all elements are zero. This takes place iff $v = u$. Next, symmetry of d is achieved through symmetry of the absolute value function. Lastly, let $v, u, w \in \mathbb{R}^k$ then

$$d(v,u) = \sum |v_i - u_i| = \sum |v_i - w_i + w_i - u_i| \leq \sum |v_i - w_i| + \sum |w_i - u_i| = d(v,w) + d(w,u)$$

■

Next, let us define the notion of a *size* of a vector.

Definition 1.7 — Norm. A norm is a function $||\cdot|| : \mathbb{R}^d \to \mathbb{R}_+$ that satisfies the following three conditions for all $a \in \mathbb{R}$ and all $u, v \in \mathbb{R}^d$:

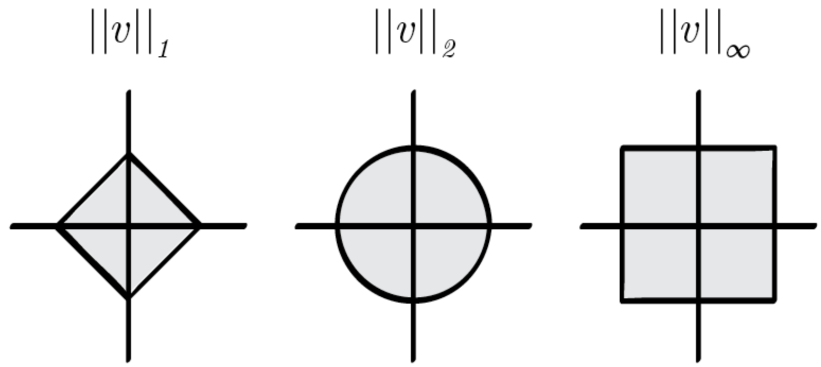
- Positive definite: $||v|| \geq 0$ and $||v|| = 0$ iff v is the zero vector.
- Positive homogeneity: $||av|| = |a| \cdot ||v||$.
- Triangle inequality: $||v+u|| \leq ||v|| + ||u||$.

We can think of this size in the sense of vector's *distance* from the origin, under some distance function defined by the norm. A few commonly used norms are:

- Absolute norm (ℓ_1): $||v||_1 := \sum |v_i|$.
- Euclidean norm (ℓ_2): $||v||_2 := \sqrt{\sum x_i^2}$.
- Infinity norm: $||x||_\infty := max_i |v_i|$.

(R) The absolute and Euclidean norms are part of a wider family of norms called the L_p norms defined as $||v||_p := \left(\sum |v_i^p|\right)^{1/p}, \quad p \in \mathbb{N}$.

Definition 1.8 Let V be a vector space and $||\cdot||$ be a norm over this space. The unit ball of $||\cdot||$ is defined as the set of vectors such that: $B_{||\cdot||} = \{v \in V : ||v|| \leq 1\}$.



Now that we have defined the notions of distances and sizes of vectors, we want to define what it means to “apply“ some vector on another.

Definition 1.9 — Inner Product. An inner product space is a vector space V over $\mathbb{R}$ together with a map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}_+$ satisfying that $\forall v, u, w \in V, \alpha \in \mathbb{R}$:

- Symmetry: $\langle v, u \rangle = \langle u, v \rangle$
- Linearity: $\langle \alpha v + w, u \rangle = \alpha \langle v, u \rangle + \langle w, u \rangle$
- Non-negativity : $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0 \iff v = 0$

Notice the similarity between the definition of a norm and of an inner product. In fact, given an inner-product space, we are also given a norm on this space.

Claim 1.2 — Induced Norm. Let H be an inner product space. Then the function $||\cdot|| : H \to \mathbb{R}_+$ is defined $\forall v \in H$ by $||v|| = \langle v, v \rangle^{\frac{1}{2}}$ is a norm on H.

Exercise 1.2 Let $v, u \in V$. Show that $\langle v, u \rangle = ||v|| \, ||u|| \cos\theta$, where θ is the angle between v, u. ■

Proof. Recall the Law of Cosines: in a triangle with lengths a, b, c, then

$$c^2 = a^2 + b^2 - 2ab\cos\theta$$

By applying the cosine law to the triangle defined by v and u and $v - u$ we see that:

$$||v-u||^2 = ||v||^2 + ||u||^2 - 2\,||v|| \cdot ||u|| \cdot \cos\theta$$

On the other hand we also know that:

$$||v-u||^2 = \langle v-u, v-u \rangle = \langle v, v \rangle - 2\langle v, u \rangle + \langle u, u \rangle = ||v||^2 + ||u||^2 - 2\langle v, u \rangle$$

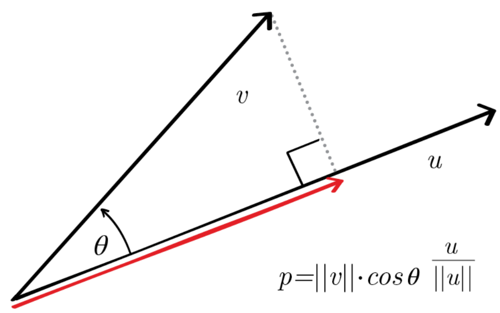
Hence, we conclude that:

$$||v|| \cdot ||u|| \cdot \cos\theta = \langle v, u \rangle$$

■

From the above, we have an expression for the angle between two vectors, using the inner-product. We can therefore define what it means to project one vector onto the other. Using the identity of $\cos\theta$:

$$p = ||v||\cos\theta \cdot \frac{u}{||u||} = ||v|| \frac{\langle v,u\rangle}{||v||\cdot||u||} \cdot \frac{u}{||u||} = \frac{\langle v,u\rangle}{||u||^2} \cdot u$$



Definition 1.10 — Vector Projection. A projection of a vector v onto a vector u, is a vector p of length $||v||\cos\theta$ in the direction of u.

Notice, that for the special case where $\theta = 90^\circ$ we get $\langle v,u\rangle = 0$. In this case we say that the vectors v,u are "orthogonal", and use the notation: $v \perp u$. If v,u are also unit vectors we say that the vectors v,u are "orthonormal" to each other.

Definition 1.11 An orthogonal matrix is a square matrix whose columns are unit vectors orthogonal to one another (i.e. they are orthonormal vectors) and whose rows are unit vectors orthogonal to one another.

Lemma 1.3 Let $A \in \mathbb{R}^{d\times d}$ orthogonal matrix, then

$$AA^\top = I = A^\top A$$

Putting together the definitions of a vector projection and orthogonal matrices we can define the notion of orthogonal projecting a vector onto some linear subspace.

Definition 1.12 Let V be a k-dimensional subspace of $\mathbb{R}^d$, and let $v_1,\ldots,v_k$ be an orthonormal basis of V. Define $P = \sum_{i=1}^k v_i v_i^\top$. The matrix P is an ***orthogonal projection matrix*** onto the subspace V.

The following lemma summarizes some useful properties of orthogonal projection matrices.

Lemma 1.4 Let $v_1,\ldots,v_k$ be a set of orthonormal vectors, and let $P = \sum_{i=1}^k v_i \otimes v_i^\top = \sum_{i=1}^k v_i v_i^\top$. P has the following properties:

- P is symmetric
- $P^2 = P$

- The eigenvalues of P are either 0 or 1. $v_1, \ldots, v_k$ are the eigenvectors of P which correspond to the eigenvalue 1.
- $(I-P)P = 0$
- $\forall x \in \mathbb{R}^d$ and $\forall u \in V$, $||x-u|| \geq ||x-Px||$
- $x \in V \Rightarrow Px = x$

R Notice that the definition of the projection matrix includes a sum of outer products

1.3 Matrix Decompositions

Matrix factorizations/decompositions are a strong tool with many theoretical as well as practical usages. It often appears in many different machine learning approaches, some of which we will encounter.

Definition 1.13 Let A be a square matrix. A is diagonalizable if there exists an invertible matrix P such that $P^{-1}AP$ is diagonal.

Next, we would like to see if we could represent A as the multiplication of orthogonal matrices, and a diagonal one.

Definition 1.14 — Eigenvector and Eigenvalue. Let A a square matrix. We say that a vector $0 \neq v \in V$ is an eigenvector of A corresponding to an eigenvalue $\lambda \in \mathbb{R}$ if $Av = \lambda v$.

Claim 1.5 Let A be a square symmetric matrix. Then there exists an orthonormal basis $u_1, \ldots, u_n \in \mathbb{R}^d$ of eigenvectors of A.

Theorem 1.6 — EVD. Let $A \in \mathbb{R}^{d\times d}$ be a real symmetric matrix. Then there exist an orthonormal matrix $U \in \mathbb{R}^{d\times d}$ and a diagonal matrix D such that, $D_{i,i}$, $i = 1..n$ are the eigenvalues of A and $A = UDU^\top$.

This decomposition of A is called Eigenvalues Decomposition (EVD). It is widely used and has some strong properties. For example, notice that it is very easy to compute high powers of A: $A^k = UDU^\top \cdot UDU^\top \cdot UDU^\top = UD^kU^\top$. It is also very easy to compute the inverse of A, if it exists: $A^{-1} = UD^{-1}U^\top$.

A drawback of the EVD is the restriction to square symmetric matrices. Though this is a rich family of matrices we would like to derive some useful decomposition for non-symmetric and even non-square matrices.

Definition 1.15 Let $(V, ||\cdot||)$ be a normed space. We say that $v \in V$ is a unit vector iff $||v|| = 1$.

Definition 1.16 Let $A \in \mathbb{R}^{m\times d}$ and let $v \in \mathbb{R}^d$, $u \in \mathbb{R}^m$ be unit vectors. We say that v, u are right- and left singular vectors of A, respectively, corresponding to a singular value $\sigma \in \mathbb{R}_+$ if $Av = \sigma u$.

Theorem 1.7 — Singular Value Decoposition (SVD). Let $A \in \mathbb{R}^{m\times d}$ be a real matrix. A can be written as a singular value decomposition of the form $A = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{d\times d}$ are orthonormal matrices, and $\Sigma \in \mathbb{R}^{m\times d}$ is a diagonal matrix **with non-negative values. These are called the singular values of A.**

Claim 1.8 Let $A = U\Sigma V^\top$ be an SVD of a matrix A. It holds that the columns of U and the rows of $V^\top$ are the left- and right singular vectors of A, corresponding to the singular values present on the diagonal of Σ.

Suppose that $rank(A) = r$. This means that the number of non-zero singular values is r, and notice that $r \le \min\{d, m\}$. When $m \le d$ then A and Σ are both wide matrices (they have more columns than rows):

$$A = U\Sigma V^\top = \begin{bmatrix} | & & | & & | \\ u_1 & \cdots & u_r & \cdots & u_m \\ | & & | & & | \end{bmatrix} \left[\begin{array}{ccc|ccc} \sigma_1 & \cdots & 0 & & & \\ \vdots & \ddots & \vdots & & 0 & \\ 0 & \cdots & \sigma_r & & & \\ \hline & & & 0 & \cdots & 0 \\ & 0 & & \vdots & \ddots & \vdots \\ & & & 0 & \cdots & 0 \end{array}\right] \begin{bmatrix} - & v_1^\top & - \\ & \vdots & \\ - & v_r^\top & - \\ & \vdots & \\ - & v_d^\top & - \end{bmatrix}$$

Since $\sigma_{r+1}, \ldots \sigma_m$ are all zero, and any off diagonal element of Σ is zero, the left- and right singular values with indices greater than r are multiplied by zeros and do not take part in the final construction of the matrix A. Their purpose is in expanding the set of left- and right singular vectors to form a basis for $\mathbb{R}^m$ and $\mathbb{R}^d$ respectively. This means that the important information carried by the SVD about the matrix A is actually contained in a smaller $r \times r$ matrix, sometimes called the **compact SVD of** A, which we can write as:

$$A = \tilde{U}\tilde{\Sigma}\tilde{V}^\top = \overbrace{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix}}^{m\times r} \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \overbrace{\begin{bmatrix} - & v_1^\top & - \\ & \vdots & \\ - & v_r^\top & - \end{bmatrix}}^{d\times r}$$

To avoid cluttered notations we will drop the $\tilde{\cdot}$ notation and refer to U, Σ, V in the compact form.

The two decompositions seen above and connected to one another. The following lemma the SVD of A to the EVD of $AA^\top$ and $A^\top A$. In particular, it shows that the SVD of A can be calculated in polynomial time in m and d.

Lemma 1.9 Let $A = U\Sigma V^\top$ be an SVD of $A \in \mathbb{R}^{m\times d}$. Then $AA^\top = U\Sigma\Sigma^\top U^\top$ is an EVD of $AA^\top$, and $A^\top A = V\Sigma^\top\Sigma V^\top$ is an EVD of $A^\top A$.

This means that the eigenvalues of $AA^\top$ and $A^\top A$ equal to the square of the singular values of A. In addition, as the orthogonal matrices of the EVD contain the eigenvectors of the matrix, the eigenvectors of $AA^\top$ are the left singular values of A while the eigenvectors of $A^\top A$ are the right singular values of A.

R Note however, that the inverse claim is not correct. Take, for example, $A = U_1 \Sigma V^\top$ with $U_1 \equiv -U$. Both relations, $AA^\top = U\Sigma\Sigma^\top U^\top$ and $A^\top A = V\Sigma^\top \Sigma V^\top$ are still EVD's but $A \neq U\Sigma V^\top$.