Note 2: Review of Basic Linear Algebra

Math 198: Math for Machine Learning

1 Topic: Vector Spaces, Linear Maps, and Matrices

Linear algebra touches nearly every facet of machine learning. Broadly, linear algebra is the study of *vector spaces* and the maps between them, *linear transformations*.

1.1 Vector Spaces and Subspaces

A (real) *vector space* is a set V that is closed under finite vector addition and scalar multiplication and that satisfies the following axioms:

(a) Associativity of addition: $(\mathbf{x}+\mathbf{y})+\mathbf{z}=\mathbf{x}+(\mathbf{y}+\mathbf{z})$;

(b) Additive identity: There exists an identity element $\mathbf{0}\in V$ such that $\mathbf{x}+\mathbf{0}=\mathbf{x}$ for all $\mathbf{x}\in V$;

(c) Additive inverses: For every $\mathbf{x}\in V$ there exists an element $-\mathbf{x}$ such that $\mathbf{x}+(-\mathbf{x})=\mathbf{0}$;

(d) Commutativity of addition[1]: $\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{x}$;

(e) Associativity of scalar multiplication: $a(b\mathbf{x})=(ab)\mathbf{x}$;

(f) Distributivity: $a(\mathbf{x}+\mathbf{y})=a\mathbf{x}+a\mathbf{y}$ and $(a+b)\mathbf{x}=a\mathbf{x}+b\mathbf{x}$;

(g) Multiplicative identity: $1\mathbf{x}=\mathbf{x}$, where $1\in\mathbb{R}$.

By convention, we refer to the vector space as V and to an element of V as a *vector*. Some vector spaces we'll be working with are

- $\mathbb{R}$ and $\mathbb{R}^d$, the spaces of one- or d-dimensional vectors over the real numbers
- $\mathbb{R}^{m\times n}$, the space of $m\times n$ matrices with real entries
- $\mathbb{P}_n$, the space of n^{th}-degree polynomials on $\mathbb{R}$ with real coefficients.

A *subspace* of a vector space V is a subset $U\subseteq V$ such that U is a vector space under the same addition and scalar multiplication operations. Subspaces are easy to characterize: A nonempty subset $U\subseteq V$ is a subspace iff U contains $\mathbf{0}$ and is closed under addition and scalar multiplication. No need to check the other axioms – since they are met in V, they are met in U.

1.2 Basis and Dimension

We'll quickly run through some key definitions. Let V be a vector space. Given $\mathbf{x}_1,...,\mathbf{x}_k\in V$, a *linear combination* of $\mathbf{x}_1,...,\mathbf{x}_k$ is any vector of the form $a_1\mathbf{x}_1+...+a_k\mathbf{x}_k$, where $a_i\in\mathbb{R}$. Note that saying V is closed under finite addition and scalar multiplication is equivalent to saying that V is closed under taking linear combinations. Given $A\subseteq V$, define the *span* of A, denoted $\text{span}(A)$, to be the set of linear combinations of vectors in A. A nonzero set of vectors $\{\mathbf{x}_1,...,\mathbf{x}_k\}\subseteq V$ is said to be *linearly independent* if there do not exist scalars $a_1,...,a_k$, all nonzero, such that $a_1\mathbf{x}_1+...+a_k\mathbf{x}_k=0$, i.e. you can't write any of the vectors as a nontrivial linear combination of the others. The definition implies that any set of vectors containing $\mathbf{0}$ is not linearly independent.

[1] Note that axioms (a) through (d) say that $(V,+)$ is an abelian group.

A *basis* for V is a set $B = \{\mathbf{x}_1, ..., \mathbf{x}_d\} \subseteq V$ such that (i) B is linearly independent and (ii) $\text{span}(B) = V$. Intuitively, (i) ensures that B doesn't have too many vectors, and (ii) ensures that B has enough vectors to write every $\mathbf{x} \in V$ as a linear combination of vectors in B. Some facts about bases:

- Does every vector space have a basis? Yes, if we assume Zorn's lemma[2] holds.
- Bases are not unique: You can check that $\{(1,0),(0,1)\}$ and $\{(1,1),(1,-1)\}$ both form bases for $\mathbb{R}^2$.
- Given a subset $S \subset V$, S could fail to be a basis because it has too many vectors (i.e. it's not linearly independent), it doesn't have enough vectors (i.e. it doesn't span V), or a combination of the two (too few and linearly dependent). But these problems are easy to fix: we can always create a basis from S by adding vectors until S spans V and/or removing vectors until S is linearly independent.
- Most importantly: Every basis of V has the same number of vectors; this number is known as the *dimension* of V, denoted $\dim V$. Dimension is unique.

We will work almost exclusively with finite-dimensional vector spaces[3]. The standard basis for the d-dimensional vector space $\mathbb{R}^d$ is $\{e_1, ..., e_d\}$, where $e_1 = (1, 0, ..., 0)$, $e_2 = (0, 1, 0, ..., 0)$, etc. From now on, assume that every vector space is finite-dimensional unless stated otherwise.

1.3 Inner Products, Orthogonality, and Norms

For a real vector space V, an *inner product* is a map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying

(a) Linearity in the first coordinate: $\langle a\mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$

(b) Symmetry: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$

(c) Positive semi-definite: $\forall v \in V, \langle v, v \rangle \geq 0$; $\langle v, v \rangle = 0 \iff v = \mathbf{0}$

By symmetry, the inner product is linear in both coordinates. A vector space equipped with an inner product is called an *inner product space*. Note that an inner product induces a *norm* (size) on V given by $||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$, which in turn induces a *metric* (distance) on V given by $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$. For our purposes, we will make use of the standard inner product on $\mathbb{R}^d$, the dot product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{d} x_i y_i.$$

Inner products allow us to define the notion of orthogonality. Two vectors $\mathbf{x}, \mathbf{y}$ are *orthogonal* if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. This will be important when we cover linear approximation.

Norms allow us to assign a "size" to each vector. In $\mathbb{R}^d$, there is an important family of norms called the ℓ^p-norms (a.k.a. p-norms). For $p \in \mathbb{Z}, p \geq 1$, define

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{d} (x_i)^p \right)^{\frac{1}{p}}.$$

Note that $||\mathbf{x}||_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$, the norm induced by the dot product.

[2] Zorn's lemma $\iff$ Axiom of Choice $\iff$ every vector space has a basis. When the existence of something is shown with Zorn's lemma, it is often difficult to construct an example of it. Can you exhibit a basis for $\mathbb{R}$ as a vector space over $\mathbb{Q}$?

[3] Infinite dimensional vector spaces are usually spaces of functions, e.g. $C(\mathbb{R}, \mathbb{R})$, the space of continuous functions from $\mathbb{R}$ to $\mathbb{R}$ with pointwise addition and s.m.. Their study is known as *functional analysis*.

1.4 Linear Maps and Isomorphism

Let V, W be vector spaces. A function $T : V \to W$ is said to be a *linear map* if

$$T(a\mathbf{x} + \mathbf{y}) = aT(\mathbf{x}) + T(\mathbf{y})$$

That is, a linear map preserves vector addition and scalar multiplication.

Associated with T are two important subspaces, the range and the kernel. The range (a.k.a. the image) of T, denoted $\text{ran}(T)$ or $\text{Im}(T)$, is given by $\text{ran}(T) = \{\mathbf{y} \in W : \mathbf{y} = T(\mathbf{x}) \text{ for some } \mathbf{x} \in V\}$. The kernel of T, denoted $\ker(T)$, is given by $\ker(T) = \{\mathbf{x} \in V : T(\mathbf{x}) = \mathbf{0}_W\}$. The image of T is a subspace of W, and the kernel of T is a subspace of V. An important result, the Rank-Nullity theorem, states that $\dim(\text{ran}(T)) + \dim(\ker(T)) = \dim(V)$.

The linearity of linear maps makes them interact nicely with the structure of the vector spaces involved. An important property of linear maps is that their behavior is determined completely be their action on a basis for the domain. Let $T : V \to W$ be a linear map and $B = \{\mathbf{x}_1, ..., \mathbf{x}_d\}$ a basis for V. Suppose that we know $T(\mathbf{x}_i)$ for all $i = 1, ..., d$. Choose some arbitrary vector $\mathbf{v} \in V$. Since B is a basis, we can write $\mathbf{v}$ as a linear combination $\mathbf{v} = a_1\mathbf{x}_1 + ... + a_d\mathbf{x}_d$. Then, by linearity, $T(\mathbf{v}) = a_1T(\mathbf{x}_1) + ... + a_dT(\mathbf{x}_d)$.

How can we say that two arbitrary vector spaces are "the same"? We use the notion of an *isomorphism*. The following are equivalent statements about a linear map $T : V \to W$:

(i) T is one-to-one and onto (in other words, T is a bijection)

(ii) T is an isomorphism (a bijective linear map)

(iii) T has an inverse, T^{-1}

(iv) V, W have the same dimension and $\ker(T) = \mathbf{0}_V$ (i.e. T is one-to-one)

(v) V, W have the same dimension and $\text{ran}(T) = W$ (i.e. T is onto)

(vi) Applying T to each element of a basis for V results in a basis for W

The vector spaces V and W are *isomorphic* if there exists an isomorphism between them, in which case we write $V \cong W$. The above equivalences imply that two vector spaces are isomorphic if and only if they share the same dimension. Thus, every d-dimensional vector space is isomorphic to $\mathbb{R}^d$. Given a d-dimensional vector space V, how do we exhibit such an isomorphism to $\mathbb{R}^d$? By choosing a basis $B = \{\mathbf{x}_1, ..., \mathbf{x}_d\}$ and letting $T(\mathbf{x}_i) = e_i$. If $\dim V = n$ and $\dim W = m$, we can always identify V with $\mathbb{R}^n$ and W with $\mathbb{R}^m$ if needed.

1.5 Matrices

The key idea of this section is that we can concretely represent linear maps between finite-dimensional vector spaces as matrices. Given a map $T : \mathbb{R}^n \to \mathbb{R}^m$, we can form a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ such that $\mathbf{Ax} = T(\mathbf{x})$ by setting the i^{th} column of A to be the column vector $T(\mathbf{e}_i) \in \mathbb{R}^m$. In other words,

$$\mathbf{A} = \begin{bmatrix} T(\mathbf{e}_1) & \dots & T(\mathbf{e}_n) \end{bmatrix}.$$

To see that this construction works, recall that the action of T is

$$\mathbf{x} = x_1\mathbf{e}_1 + ... + x_n\mathbf{e}_n \mapsto T(\mathbf{x}) = x_1T(\mathbf{e}_1) + ... + x_nT(\mathbf{e}_n) = \mathbf{Ax}.$$

Note that we necessarily define A with respect to an ordered basis for $\mathbb{R}^n$ and $\mathbb{R}^m$. In fact, any time we write out the elements of a vector or matrix, we do so with respect to some ordered basis. For example, the i-th column of a matrix A represents the action of that matrix on the i-th basis vector. To limit confusion, you can always assume that a matrix or vector is being written with respect to the standard bases unless otherwise noted.

2 Applications: Projections

2.1 Motivation

Recall from note 1 that the goal of Ordinary Least Squares is to determine a weight vector $\mathbf{w}$ such that $\mathbf{Xw} \approx \mathbf{y}$ for our data matrix $\mathbf{X}$ and observations y. Suppose that $y \in \text{range}(\mathbf{X})$ and that $\mathbf{X}$ is invertible.[4] Then we could solve directly: $\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$. Of course, this scenario is rarely, if ever, seen in practice. In general, we will not be able to come up with an exact solution for the equation $\mathbf{Xw} = \mathbf{y}$; instead, we seek a weight vector $\mathbf{w}$ such that $\mathbf{Xw}$ is the best approximation to $\mathbf{y}$ in the range of $\mathbf{X}$.[5] To do so, we will first introduce the notion of an *orthogonal projection*.

2.2 Projectors

Suppose we have some vector space V, some subspace $W \subset V$, and some element $\mathbf{v} \in V$ such that $\mathbf{v} \notin W$. Define the orthogonal projection of $\mathbf{v}$ in W, $\mathbf{v}_w$, to be the vector in W which is closest to $\mathbf{v}$:

$$\mathbf{v}_w = \arg\min_{\mathbf{w} \in W} ||\mathbf{v} - \mathbf{w}||$$

How can we go about finding such a vector? The first step is to note that $\mathbf{v}_w$ is the closest vector in W to $\mathbf{v}$ if and only if $\mathbf{v} - \mathbf{v}_w$ is orthogonal to every $\mathbf{w} \in W$.

Proof. Fix some arbitrary $\mathbf{w} \in W$, and define the function[6] $f_w(t) = ||\mathbf{v} - (\mathbf{v}_w + t\mathbf{w})||^2$. Then f is the square of the distance between $\mathbf{v}_w + t\mathbf{w}$, a vector in W, and v. It should be clear that f is minimized when $t = 0$. So, the derivative of f_w at $t = 0$ is 0. To determine the derivative of f_w, we first expand it by rewriting it as an inner product:

$$\begin{aligned} f_w(t) &= \langle (\mathbf{v} - \mathbf{v}_w) - t\mathbf{w}, (\mathbf{v} - \mathbf{v}_w) - t\mathbf{w} \rangle \\ &= \langle \mathbf{v} - \mathbf{v}_w, \mathbf{v} - \mathbf{v}_w \rangle - 2\langle \mathbf{v} - \mathbf{v}_w, t\mathbf{w} \rangle + \langle t\mathbf{w}, t\mathbf{w} \rangle \\ &= ||\mathbf{v} - \mathbf{v}_w||^2 - 2t\langle \mathbf{v} - \mathbf{v}_w, \mathbf{w} \rangle + t^2||\mathbf{w}||^2 \end{aligned}$$

We then take the derivative with respect to t:

$$f'_w(t) = -2\langle \mathbf{v} - \mathbf{v}_w, \mathbf{w} \rangle + 2t||\mathbf{w}||^2$$

and so

$$0 = f'_w(0) = -2\langle \mathbf{v} - \mathbf{v}_w, \mathbf{w} \rangle$$

and so $\mathbf{v} - \mathbf{v}_w$ is orthogonal to $\mathbf{w}$. Since our choice of $\mathbf{w}$ was arbitrary, we conclude that $\mathbf{v} - \mathbf{v}_w$ is orthogonal to every vector in W. To prove the converse, note that f_w is quadratic in its input and non-negative; so if $\mathbf{v} - \mathbf{v}_w$ is orthogonal to every vector in W, then $f'_w(0) = 0$, and so $t = 0$ must be the global minimum of f_w for all $\mathbf{w}$; so, $||\mathbf{v} - (\mathbf{v}_w + t\mathbf{w})||^2$ is minimized for $t = 0$, and thus $\mathbf{v}_w$ is the closest vector in W to $\mathbf{v}$. □

This proof has an important corollary. We have proven that $\mathbf{v} - \mathbf{v}_w$ is orthogonal to the subspace W. Let $W^\top$ denote the set of all vectors in V which are orthogonal to W; so, $\mathbf{v} - \mathbf{v}_w \in W^\top$. It turns out that $W^\top$ is itself a subspace such that $W \oplus W^\top = V$.

Proof. We first show that $W^\top$ is a subspace of V. Fix $\mathbf{v_1}, \mathbf{v_2} \in W^\top$ and $a \in \mathbb{R}$. Then, for all $\mathbf{w} \in W$,

$$\langle \mathbf{v_1} + \mathbf{v_2}, \mathbf{w} \rangle = \langle \mathbf{v_1}, \mathbf{w} \rangle + \langle \mathbf{v_2}, \mathbf{w} \rangle = 0 + 0 = 0$$

$$\langle a\mathbf{v_1}, \mathbf{w} \rangle = a\langle \mathbf{v_1}, \mathbf{w} \rangle = a0 = 0$$

[4] Since a matrix $\mathbf{A}$ represents a linear map T, A is invertible if T is invertible, and the inverse of $\mathbf{A}$, $\mathbf{A^{-1}}$, represents T^{-1} with respect to the same bases as $\mathbf{A}$.

[5] We will not yet present a probabilistic motivation for our idea of "closeness"; this will be done in the probability section of the course.

[6] Do not confuse the w in $\mathbf{v}_w$ and the $\mathbf{w}$ in the $t\mathbf{w}$ term; $\mathbf{v}_w$ is the orthogonal projection of $\mathbf{v}$ in W, and $\mathbf{w}$ is some arbitrary vector in W.

so $\mathbf{v_1} + \mathbf{v_2} \in W^\top$ and $a\mathbf{v_1} \in W^\top$ and so $W^\top$ is a subspace. We now show that $W \oplus W^\top = V$. To do so, we show that any vector $\mathbf{v} \in V$ can be decomposed into the sum of two vectors, one in W, and one in $W^\top$. Of course, $\mathbf{v} = \mathbf{v}_w + (\mathbf{v} - \mathbf{v}_w)$; since $\mathbf{v}_w \in W$ and $\mathbf{v} - \mathbf{v}_w \in W^\top$, we have $V = W \oplus W^\top$. □

Suppose V has dimension n and W has dimension k. Then by the corollary, $W^\top$ has dimension $n - k$. Furthermore, suppose we have some orthogonal[7] basis (likely non-standard) for V, $\beta = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, such that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is a basis for W. Then $\{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$ is a basis for $W^\top$.[8] Consider some vector $\mathbf{v} \in V$. We can write $\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$ for appropriate coefficients α_i. Define $\mathbf{v}_w = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i$. Then

$$\mathbf{v} - \mathbf{v}_w = \sum_{i=k+1}^{n} \alpha_i \mathbf{v}_i \in W^\top$$

and so $\mathbf{v}_w = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i$ is the orthogonal projection of $\mathbf{v}$ in W. So we have reduced the problem of finding the closest approximation to $\mathbf{v}$ to the problem of finding an orthogonal basis for our subspace, W.

2.3 Conclusion

In section 2.1, we concluded that we seek a weight vector $\mathbf{w}$ such that $\mathbf{Xw}$ is the best approximation to $\mathbf{y}$ in the range of $\mathbf{X}$. Recall from note 1 that $\mathbf{X}$ is an $n \times d$ matrix, $\mathbf{y}$ is an n-dimensional vector, and $\mathbf{w}$ is a d-dimensional vector. Suppose that $\mathbf{X}$ is *full rank*, that is, $\dim(\text{range}(\mathbf{X})) = d$. In the language of section 2.2, we have that $V = \mathbb{R}^n$, $W = \text{range}(\mathbf{X}) \cong \mathbb{R}^d$, $\mathbf{v} = \mathbf{y}$, and $\mathbf{v}_w = \mathbf{Xw}$. In the coming weeks, we will complete this derivation using special classes of matrices, and then confirm that it behaves as we would expect by developing the class of *projection matrices*.

[7] Such a basis is guaranteed to exist. We can turn any basis into an orthonormal basis using Gram-Schmidt Orthonormalization, which is out of scope for this course because it is computationally horrifying. So, we can take any basis for W, orthonormalize it, and then extend it to an orthonormal basis for V to obtain the desired basis for any finite vector space.

[8] The proof for this is trivial, but the fact that our basis for V is orthogonal is essential.