Lecture 4: Vector Spaces, Spans and Bases *

10-606 Mathematical Foundations for Machine Learning

1 The vector space $\mathbb{R}^n$

The best-known example of a vector space is $\mathbb{R}^n$: the space of n-element vectors of real numbers. The first important properties of $\mathbb{R}^n$ are:

- It is *closed* under addition and scalar multiplication of vectors: for any pair of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and scalar $a \in \mathbb{R}$, $a\mathbf{x} + \mathbf{y} \in \mathbb{R}^n$.
- The following axioms are satisfied $\forall$ vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and scalars $a, b \in \mathbb{R}$:
  1. Associativity of addition: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$
  2. Associativity of scalar multiplication: $(ab)\mathbf{x} = a(b\mathbf{x})$
  3. Commutativity of addition: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
  4. Existence of an additive identity: $\mathbf{x} + \mathbf{0} = \mathbf{x}$ where $\mathbf{0}$ is the vector of n zeros.
  5. Existence of additive inverses: $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ where $\mathbf{0}$ is the additive identity.
  6. Existence of scalar multiplication identity: $(1)\mathbf{x} = \mathbf{x}$
  7. Distributivity of scalar multiplication over vector addition: $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$
  8. Distributivity of scalar multiplication over scalar addition: $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$

The properties above make $\mathbb{R}^n$ a *vector space* (technically a vector space over the *field* $\mathbb{R}$ but that last bit is usually omitted).

In addition, $\mathbb{R}^n$ also has the following properties:

- It has an inner product, written $\mathbf{x} \cdot \mathbf{y}$ or $\langle \mathbf{x}, \mathbf{y} \rangle$, with $\mathbf{x} \cdot \mathbf{y} \in \mathbb{R}$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- This inner product satisfies the following properties $\forall$ vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and scalars $a, b \in \mathbb{R}$:
  1. Symmetry: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$
  2. Bilinearity (linear in each argument with the other fixed): $\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle$
  3. Positive-definiteness: $\mathbf{x} \cdot \mathbf{x} \geq 0$ and $\mathbf{x} \cdot \mathbf{x} = 0 \Leftrightarrow \mathbf{x} = 0$.

Much like the first set of properties define a vector space, this second set of properties make $\mathbb{R}^n$ an *inner product* space.

*Compiled on Tuesday 6th June, 2023 at 14:00

We can go even further and note that $\mathbb{R}^n$ has a final set of desirable properties:

- We can use the inner product to define a norm $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$. This norm lets us define other useful concepts like the convergence of a sequence of vectors to a limit point.
- Under the above norm, every Cauchy sequence converges. That is, for a sequence $\mathbf{x}_t$ with $t = 1, 2, \ldots$, if $\|\mathbf{x}_t - \mathbf{x}_{t+1}\| \to 0$ as $t \to \infty$, then $\mathbf{x} = \lim_{t\to\infty} \mathbf{x}_t$ exists and $\mathbf{x} \in \mathbb{R}^n$.

These last two properties make $\mathbb{R}^n$ a *complete* inner product space. Both of these are stronger than just being a vector space: they are extensions of a vector space.

1.1 Orthogonal and normal vectors

In an inner product space such as $\mathbb{R}^n$, two vectors $\mathbf{x}$ and $\mathbf{y}$ are called *orthogonal* if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Geometrically, orthogonal vectors form a right angle with one another.

A vector $\mathbf{x}$ is called *normal* if it has unit length: $\langle \mathbf{x}, \mathbf{x} \rangle = 1$. Two vectors are *orthonormal* if they are orthogonal and both are normal.

Knowledge check

1. **Select one**: Which of the following describes the relationship between $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.
   a) $\mathbf{x}$ and $\mathbf{y}$ are not orthogonal.
   b) $\mathbf{x}$ and $\mathbf{y}$ are orthogonal but not orthonormal.
   c) $\mathbf{x}$ and $\mathbf{y}$ are orthonormal.
   - **Answer**: b, $\langle \mathbf{x}, \mathbf{y} \rangle = (1)(1) + (1)(-1) = 0$ but $\langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle = 2$.
2. **Select one**: Which of the following describes the relationship between $\mathbf{x} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.
   a) $\mathbf{x}$ and $\mathbf{y}$ are not orthogonal.
   b) $\mathbf{x}$ and $\mathbf{y}$ are orthogonal but not orthonormal.
   c) $\mathbf{x}$ and $\mathbf{y}$ are orthonormal.
   - **Answer**: cm $\langle \mathbf{x}, \mathbf{y} \rangle = (0)(1) + (-1)(0) = 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0^2 + (-1)^2 = 1 = (1)^2 + 0^2 = \langle \mathbf{y}, \mathbf{y} \rangle$.
3. **Select one**: Which of the following describes the relationship between $\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$.
   a) $\mathbf{x}$ and $\mathbf{y}$ are not orthogonal.
   b) $\mathbf{x}$ and $\mathbf{y}$ are orthogonal but not orthonormal.
   c) $\mathbf{x}$ and $\mathbf{y}$ are orthonormal.
   - **Answer**: a, $\langle \mathbf{x}, \mathbf{y} \rangle = 1 \neq 0$ and if two vectors are not orthogonal, then they cannot be orthonormal.

1.2 Other vector spaces

All of these properties make $\mathbb{R}^n$ a nice formal system to work with. But sometimes we need to work with objects that are not in $\mathbb{R}^n$, e.g., matrices or functions. We can still use some of the same tricks when working with these objects, by abstracting out the important properties that we like from $\mathbb{R}^n$.

This sort of abstraction is very important in machine learning: it lets us take algorithms that are designed to pick out some "optimal" or best element of $\mathbb{R}^n$ (often called a *parameter vector*) and generalize them to work on other classes of objects such as matrices or functions. For some classes of objects, this is the main way that we know how to design effective learning algorithms.

We define a vector space to be a set V (whose elements are called *vectors*) together with operations of addition and scalar multiplication that behave in the usual way (i.e., copying the important properties from $\mathbb{R}^n$), e.g., $a\mathbf{x} + \mathbf{y} \in V$ for $a \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in V$.

Note that the scalar a is still a real number, even though we've changed the vectors that it is multiplying; the set of scalars doesn't change when we go from $\mathbb{R}^n$ to a general vector space V. More precisely, we're going to be talking about a "vector space over the reals", which means that our scalars are elements of $\mathbb{R}$. [1]

Some examples of other vector spaces are:

- The set of matrices $\mathbb{R}^{m \times n}$ is a vector space, if we interpret addition and scalar multiplication elementwise.
- We can make the set of real functions of one argument $\mathbb{R} \to \mathbb{R}$ into a vector space, if we define addition and scalar multiplication to operate separately on each possible argument to our functions. For example, we would define $(f+3g)(x) = f(x)+3g(x)$; this is the usual interpretation of addition and scalar multiplication of functions.
- The real numbers themselves are a vector space, though kind of a trivial one as the vectors and scalars come from the same set.

In addition, one vector space can be contained inside another one: $U \subseteq V$; in this case U is called a *subspace* of V. It is typically easy to verify whether or not some subset of V is a subspace (and hence a vector space itself) because many of the properties of V are directly inherited by all of its subsets. Formally, a subset U of a vector space V is a subspace if and only if:

- It contains the additive identity, $\mathbf{0}$.
- U itself is closed under addition and scalar multiplication i.e., for any pair of vectors $\mathbf{x}, \mathbf{y} \in U$ and scalar $a \in \mathbb{R}$, $a\mathbf{x} + \mathbf{y} \in U$.

Almost all of the vector spaces that we deal with in machine learning will be complete inner product spaces, so we don't need to worry too much about checking all of the required properties (see above). But these properties are worth keeping in the back of our minds, since they can cause algorithms to fail in inconvenient ways in the rare cases they are not satisfied.

[1] You can define vector spaces that use different kinds of scalars, e.g., complex numbers $\mathbb{C}$ or integers modulo a prime but those are uncommon in machine learning.

Knowledge check

1. **True or False**: Let $U = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq 1\}$. U is a subspace of $\mathbb{R}^n$.
   - **Hint**: First, try to describe the set U in words.
   - **Answer**: False. In words, U is the set of all vectors with norm less than or equal to 1 or geometrically, all vectors in the unit hypersphere centered around the origin. While this subset does contain the additive identity $\mathbf{0}$, it is not closed under addition nor scalar multiplication. As a simple counterexample consider the vector of a 1 followed by $n-1$ 0's and the vector of $n-1$ 0's followed by a 1: both of these have unit norm and are in U but their sum has norm $\sqrt{2}$ and is thus, not in U.
2. **True or False**: Let $U = \{\mathbf{0}\}$. U is a subspace of $\mathbb{R}^n$.
   - **Answer**: True. Trivially, U contains the additive identity. Furthermore, $a\mathbf{0} + \mathbf{0} = \mathbf{0} + \mathbf{0} = \mathbf{0} \in U \; \forall \; a \in \mathbb{R}$.

2 Span

Given a set of vectors B, their *span* is the set of all their linear combinations:

$$\text{span}(B) = \{w_1\mathbf{b}_1 + w_2\mathbf{b}_2 + \ldots + w_n\mathbf{b}_n \mid w_i \in \mathbb{R}, \mathbf{b}_i \in B\}$$

Note that the set-builder notation discards duplicates: if $\sum_i v_i\mathbf{a}_i = \sum_i w_i\mathbf{b}_i$ for two different lists of weights v_i and w_i and vectors $\mathbf{a}_i$ and $\mathbf{b}_i$, we only include the linear combination vector once.

The span of any set of vectors B is a vector space: we can add any two vectors in the span by adding their linear-combination weights, and scalar-multiply a vector in the span by scalar-multiplying its weights. Furthermore, if the elements of B come from some vector space V, then their span is a subspace of V: $\text{span}(B) \subseteq V$. For example, if $|B| = 2$ and the elements of B are nonzero 3-element vectors, then the span of B is a line or a plane in $\mathbb{R}^3$.

Note that the above definition makes sense even if B has infinitely many elements: we include all linear combinations of *finitely many* elements of B. This is sometimes called the *finite span* to avoid ambiguity.

3 Basis

A *basis* of a vector space V is a set of vectors $B = \{\mathbf{b}_1, \mathbf{b}_2, \ldots\}$ with two properties:

- It spans our vector space, $V = \text{span}(B)$. That is, any vector in V can be represented as a linear combination of basis vectors from B: $\forall\, \mathbf{v} \in V$, $\exists$ scalars $x_1, x_2, \ldots$ such that $\mathbf{v} = x_1\mathbf{b}_1 + x_2\mathbf{b}_2 + \ldots$.
- It's *minimal* or as small as possible: removing any element from B destroys the previous property.

The elements of B are called *basis vectors*.

Given a basis, we can use the scalars $x_1, x_2, \ldots$ as a concrete representation of the abstract vector $\mathbf{v}$:

$$\mathbf{x} \in V \quad \leftrightarrow \quad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

This is called a *coordinate representation*.

For example, in a 2d map, we could pick as our basis one vector pointing up in the image and one vector pointing right, each with a length that corresponds to one meter in the physical world. Or we could pick one vector pointing north in the image and one pointing east, with a length that corresponds to 1km in the real world. We could even pick one vector pointing southeast and one north-by-northwest, although this would violate the usual convention of preserving lengths and angles from the real world in our map coordinates (see the definition of orthonormal bases below).

It's a fundamental theorem that any vector space has a basis, and that every basis has the same number of basis vectors (though this number might be infinite). The number of basis vectors is called the *dimension* of the corresponding vector space. This is the same idea of dimension that we're used to: a line is a one-dimensional vector space, a plane is a two-dimensional vector space, and so forth.

Coordinate representations in a given basis are *unique*: given a vector $\mathbf{x} \in V$, there is only one linear combination of basis vectors that yields $\mathbf{x}$. The coordinate representation depends on which basis we pick: if we choose a different basis B', the representation of $\mathbf{x}$ can look totally different. However, it's important to remember that the underlying vector $\mathbf{x}$ hasn't changed, only our representation of it.

It is slightly tricky to define basis and dimension in infinite-dimensional vector spaces. The difficulty comes in whether to allow convergent infinite sums of basis vectors (our above definition does not). It only makes sense to allow infinite sums if we have an appropriate notion of convergence; we don't always have such a notion, but we do in many important cases, including all complete inner product spaces. Depending on whether we do allow infinite sums, the size of a basis might be different.

3.1 Linear independence

We said above that a basis has to be minimal: if any vector $\mathbf{b}_i$ can be removed from B while keeping $\mathrm{span}(B) = V$, then B is not a basis. More generally, for any list of vectors X, if we can remove some vector $\mathbf{x}_i \in X$ without changing $\mathrm{span}(X)$, the vectors in X are called *linearly dependent*. This means that the $\mathbf{x}_i$ can itself be expressed as a linear combination of the other vectors $\mathbf{x}_j \in X$ for $j \neq i$. On the other hand, if no vector can be removed from X without changing $\mathrm{span}(X)$, then the vectors are called *linearly independent*.

For example, the vectors

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

are linearly dependent. One way to show this is that we can represent the last vector as the sum of the first two. On the other hand, any two of these vectors are linearly independent.

Geometrically, linear independence is a generalization of being collinear or coplanar: that is, we are asking whether any vector lies in the line, plane, or other subspace defined by the other vectors.

3.2 Orthonormal bases

If the elements of a basis B are orthonormal (that is, if $\langle \mathbf{b}_i, \mathbf{b}_j = 0$ when $i \neq j$ and $\langle \mathbf{b}_i, \mathbf{b}_i \rangle = 1$ for all i) then we say that B is an orthonormal basis. Orthonormal bases are important since they preserve lengths and angles. That is, if $\mathbf{x}, \mathbf{y} \in V$ are abstract vectors and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are their concrete representations under an orthonormal basis, then $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{u} \cdot \mathbf{v}$.

This is a crucial property if we're working in an inner product space. For an arbitrary vector space, lengths and angles might not even be defined, so there's no point in trying to preserve them. But for an inner product space, lengths and angles are an intrinsic property of the space. If we use a non-orthonormal basis, the standard inner product in $\mathbb{R}^n$ will give us wrong answers for the lengths and angles between vectors. Instead, if we want to reason about lengths and angles, either we need to switch to an orthonormal basis or we need to define a non-standard inner product for $\mathbb{R}^n$.

So, if we have a non-orthonormal basis for an inner product space, and if we want to use it to build a concrete representation, we need to be careful. That being said, given any basis for an inner product space, we can construct an orthonormal one. One method for doing so is Gram-Schmidt orthonormalization.

Knowledge check For the following questions, let $B = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$.

1. **True or False**: $\mathrm{span}(B) = \mathbb{R}^2$

   item **Answer**: True. Any vector in $\mathbb{R}^2$ can be expressed as some linear combination of the vectors in B.
2. **True or False**: B is a basis for $\mathbb{R}^2$
   - **Answer**: False. B is not minimal; one easy way to see this is by observing that $\mathbb{R}^2$ is a 2-dimensional space so any set of 3 vectors cannot be a basis. Alternatively, you could note that the vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is a linear combination of the other two so the vectors are not linearly independent.
3. **True or False**: Any two vectors from B form a basis for $\mathbb{R}^2$
   - **Answer**: True. You can verify that any pair of vectors from B are linearly independent and because $\mathbb{R}^2$ is a 2-dimensional space, any two linearly independent vectors from $\mathbb{R}^2$ can be a basis.
4. **True or False**: Any two vectors from B form an orthonormal basis for $\mathbb{R}^2$
   - **Answer**: False. $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ are orthonormal but the norm of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is $\sqrt{2}$ so any pair of vectors containing $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ cannot be an orthonormal basis.