Vector Spaces (Lab 1)

BST 235: Advanced Regression and Statistical Learning

Alex Levis, Fall 2019

1 Some motivation

In the first lecture, we laid out the basic regression framework that we will study in this course. At hand we have a collection of inputs $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathcal{X} \subseteq \mathbb{R}^d$, a corresponding collection of outputs $Y_1, \ldots, Y_n \in \mathcal{Y}$ (for now we consider $\mathcal{Y} \subseteq \mathbb{R}$), and we are tasked vaguely with using this data to construct a function $g : \mathcal{X} \to \mathcal{Y}$ that can predict new outputs given new inputs, in some 'good' way. We introduced two possible probabilistic frameworks for describing how the data arise,

- Random design: $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n) \overset{\text{iid}}{\sim} P$
- Fixed design: $(\mathbf{X}_1, \ldots, \mathbf{X}_n) \sim P^*$, and given $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, the outputs $Y_1, \ldots, Y_n$ are independently drawn from $P_{Y|\mathbf{X}}$, induced by the joint distribution P. In particular, we assume that for $i \in \{1, \ldots, n\}$,

$$Y_i \perp\!\!\!\perp ((Y_1, \mathbf{X}_1), \ldots, (Y_{i-1}, \mathbf{X}_{i-1}), (Y_{i+1}, \mathbf{X}_{i+1}), \ldots, (Y_n, \mathbf{X}_n)) \,\big|\, \mathbf{X}_i.$$

Using the notions of *loss* and *risk*, we decided that in a certain sense, the conditional mean $g_P(\mathbf{X}) = \mathbb{E}_P[Y \mid \mathbf{X}]$, called the *regression function*, was the optimal choice for g (how precisely was this decided?). Since we cannot possibly know this function exactly, a good idea is to use the data at hand to construct a *regression estimator* $\widehat{g}_n$ to approximate g_P, and subsequently *empirical risk minimization* (ERM) was proposed as a reasonable way to choose $\widehat{g}_n$. Unfortunately, even in the simple setting where $\mathbf{X}$ is a one-dimensional continuous variable, this approach fails unless we make further assumptions on the form of $\mathbb{E}_P[Y \mid \mathbf{X}]$.

Indeed, we will spend much of this course assuming a *linear model* for $\mathbb{E}_P[Y \mid \mathbf{X}]$. Specifically, we will postulate the existence of $\boldsymbol{\beta}(P) \in \mathbb{R}^d$ such that

$$\mathbb{E}_P[Y \mid \mathbf{X}] = \mathbf{X}^T \boldsymbol{\beta}(P) = \sum_{j=1}^{d} \beta_j(P) X_j.$$

Strictly speaking, with this modeling restriction and the mechanism of ERM, we could go through the brute force calculus to derive the least squares estimator $\widehat{\boldsymbol{\beta}}$ and end the story with the regression estimator $\widehat{g}_n(\mathbf{X}) = \mathbf{X}^T \widehat{\boldsymbol{\beta}}$. However, doing only this would be foolish, as it would ignore the deep and beautiful structure of both the statistical model, and the optimal estimators that fall out. The mathematical language best suited for understanding linear models is, quite naturally, linear algebra, and this is what we will explore in the coming weeks.

2 Fundamental vector space concepts

Our study of linear algebra in BST 235 will cover the following broad topics:

- Vector spaces, subspaces, basis and dimension
- Inner products and norms, orthogonal projection
- Linear maps and matrices, rank, inverses
- Spectral / singular value decomposition of matrices, generalized inverses

Today, we will take as our goal to understand the ideas in the first bullet point.

Definition 1. Suppose we have a non-empty set V, a field $\mathbb{F}$ and operations $\oplus : V \times V \to V$, $\odot : \mathbb{F} \times V \to V$. The triple $(V, \oplus, \odot)$, or just V if clear from context, is called a *vector space* if the following axioms are satisfied.

(1) **Vector addition**:

  (a) associativity: $(v_1 \oplus v_2) \oplus v_3 = v_1 \oplus (v_2 \oplus v_3)$, for all $v_1, v_2, v_3 \in V$,

  (b) identity element: $\exists\, 0_V \in V$ such that $v \oplus 0_V = v$, for all $v \in V$,

  (c) commutativity: $v_1 \oplus v_2 = v_2 \oplus v_1$, for all $v_1, v_2 \in V$,

  (d) inverse element: $\forall\, v \in V, \exists -v \in V$ such that $-v \oplus v = 0_V$.

(2) **Scalar multiplication**:

  (a) associativity: $(a_1 \cdot a_2) \odot v = a_1 \odot (a_2 \odot v)$, for all $a_1, a_2 \in \mathbb{F}$, $v \in V$,

  (b) identity element: $1_{\mathbb{F}} \odot v = v$, for all $v \in V$,

  (c) distributivity wrt vector addition: $a \odot (v_1 \oplus v_2) = (a \odot v_1) \oplus (a \odot v_2)$, $\forall a \in \mathbb{F}$, $v_1, v_2 \in V$.

  (d) distributivity wrt to field addition: $(a_1 + a_2) \odot v = (a_1 \odot v) \oplus (a_2 \odot v)$, $\forall a_1, a_2 \in \mathbb{F}$, $v \in V$.

Remark 1. In this course, we will exclusively consider $\mathbb{F} = \mathbb{R}$, and we refer to V as a *real vector space*. In this case we can unambiguously write 0 and 1 for $0_{\mathbb{F}}$ and $1_{\mathbb{F}}$, respectively. Note that these axioms were historically landed upon due to their efficiency — they are minimal, but imply all the properties we would expect. As an exercise, you might show the following consequences of the above definition:

- If $u \oplus v = u \oplus w$ then $v = w$ (cancellation).
- The zero vector 0_V is unique, as are additive inverses.
- For any $v \in V$, $0 \odot v = 0_V$.
- For any $a \in \mathbb{F}$, $a \odot 0_V = 0_V$.
- $(-1) \odot v = -v$, for all $v \in V$.

Finally, note that we will often use $+$ instead of $\oplus$, av for $a \odot v$, and infer whether we mean vector or field operations based on the context.

Definition 2. Suppose $(V, \oplus, \odot)$ is a vector space over the field $\mathbb{F}$. We say that $U \subseteq V$ is a *linear subspace*, or *subspace*, of V if $(U, \oplus, \odot)$ is a vector space over $\mathbb{F}$, where $\oplus$ and $\odot$ are restricted to U.

Lemma 1. Given vector space $(V, \oplus, \odot)$ over $\mathbb{F}$, the set $U \subseteq V$ is a subspace of V if and only if

(i) $U \neq \emptyset$ (equivalently, check $0_V \equiv 0_U \in U$),

(ii) $u, v \in U \implies u \oplus v \in U$ (closure under addition), and

(iii) $a \in \mathbb{F}, v \in U \implies av \in U$ (closure under scalar multiplication).

Now that we have defined vector spaces as sets of elements for which addition and scalar multiplication are well-behaved, we can now introduce two dual concepts that characterize collections of vectors. One concept is *span*, which describes the overall expressiveness of a set of vectors. The other concept is *linear independence*, which pertains to the non-redundancy in such a collection. Formally, these are defined as follows:

Definition 3. Suppose $(V, \oplus, \odot)$ is a vector space over $\mathbb{F}$. The *linear span* of the collection of vectors $\{v_1, \ldots, v_k\} \subseteq V$ is the set of all linear combinations of these vectors,

$$(a_1 \odot v_1) \oplus \cdots \oplus (a_k \odot v_k) =: \sum_{\ell=1}^{k} a_\ell v_\ell.$$

We write

$$\mathscr{L}(v_1, \ldots, v_k) = \left\{ \sum_{\ell=1}^{k} a_\ell v_\ell \,\middle|\, a_1, \ldots, a_k \in \mathbb{F} \right\}.$$

We say that the vectors $v_1, \ldots, v_k$ *span* a subspace $U \subseteq V$ if $\mathscr{L}(v_1, \ldots, v_k) = U$.

The vectors $v_1, \ldots, v_k$ are called *linearly independent* if

$$0_V = \sum_{\ell=1}^{k} a_\ell v_\ell \implies 0 = a_1 = \cdots = a_k,$$

i.e., there is no non-trivial way to combine the vectors to yield 0_V. If there exist, $a_1, \ldots, a_k \in \mathbb{F}$ not all zero such that $0_V = \sum_{\ell=1}^{k} a_\ell v_\ell$, then $v_1, \ldots, v_k$ are called *linearly dependent*.

Definition 4. Let $(V, \oplus, \odot)$ be a vector space over $\mathbb{F}$. A collection $\{v_1, \ldots, v_k\} \subseteq V$ is called a *basis* for V if $v_1, \ldots, v_k$ span V and are linearly independent.

Remark 2. A basis for a vector space essentially comprises a coordinate system — a minimal (cf. linearly independent) set of points that are sufficiently expressive (cf. spanning) to describe the whole space through linear combinations.

Two issues arise immediately: existence and uniqueness. With a slight expanding of the above definitions of span and linear independence to account for infinite sets, it can be shown that bases always exist. On the other hand, coordinate systems are not unique, and therefore neither are bases. Nevertheless, the number of elements in a basis is a constant for a given vector space, as we will demonstrate via a key lemma in class. This will allow us to define dimension, $\dim(V)$, as the cardinality of an arbitrary basis for V.

3 Exercises

Exercise 1. Convince yourself that the following are real vector spaces, by specifying natural choices for $\oplus$, $\odot$, 0_V, and additive inverses:

(a) $\mathbb{R}^d = \{[x_1 \cdots x_d]^T \,|\, x_1, \ldots, x_d \in \mathbb{R}\}$, for $d \in \mathbb{N}$

(b) For a probability space $(\Omega, \mathcal{A}, P)$, the space

$$L_2(P) = \{X : \Omega \to \mathbb{R} \,|\, X \text{ measurable}, \mathbb{E}_P(X^2) < \infty\}.$$

Exercise 2. Taking $V = \mathbb{R}^2$, give examples of a non-empty subset $U \subseteq V$ that is *not* a subspace, but that *does* satisfy

(a) Closure under addition, and remaining vector space properties

(b) Closure under scalar multiplication, and remaining vector space properties

Exercise 3. Let V be a vector space, and $U, W \subseteq V$ be two linear subspaces. Show that

(a) $U \cap W$ is a linear subspace of V.

(b) $U \cup W$ is a subspace iff $U \subseteq W$ or $W \subseteq U$.

Exercise 4. Let V be a vector space, and suppose $\{v_1, \ldots, v_k\}$ are linearly independent. Show that $0_V \notin \{v_1, \ldots, v_k\}$.

Exercise 5. Let V be a vector space, and suppose $\{v_1, \ldots, v_k\}$ are linearly dependent. Show that there exists $1 \leq j \leq k$, and scalars $a_1, \ldots, a_{j-1}$ such that

$$v_j = \sum_{\ell=1}^{j-1} a_\ell v_\ell,$$

where a sum from 1 to 0 is defined as 0_V.