MIT 9.520/6.860: Statistical Learning Theory **Fall 2016**

Appendix1 - Basic Math

Lorenzo Rosasco

These notes present a brief summary of some of the basic definitions from calculus that we will need in this class. Throughout these notes, we assume that we are working with the base field $\mathbb{R}$.

1.1 Structures on Vector Spaces

A **vector space** V is a set with a linear structure. This means we can add elements of the vector space or multiply elements by scalars (real numbers) to obtain another element. A familiar example of a vector space is $\mathbb{R}^n$. Given $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$, we can form a new vector $x + y = (x_1 + y_1, \ldots, x_n + y_n) \in \mathbb{R}^n$. Similarly, given $r \in \mathbb{R}$, we can form $rx = (rx_1, \ldots, rx_n) \in \mathbb{R}^n$.

Every vector space has a basis. A subset $B = \{v_1, \ldots, v_n\}$ of V is called a **basis** if every vector $v \in V$ can be expressed uniquely as a linear combination $v = c_1 v_1 + \cdots + c_m v_m$ for some constants $c_1, \ldots, c_m \in \mathbb{R}$. The cardinality (number of elements) of V is called the **dimension** of V. This notion of dimension is well defined because while there is no canonical way to choose a basis, all bases of V have the same cardinality. For example, the standard basis on $\mathbb{R}^n$ is $e_1 = (1, 0, \ldots, 0), e_2 = (0, 1, 0, \ldots, 0), \ldots, e_n = (0, \ldots, 0, 1)$. This shows that $\mathbb{R}^n$ is an n-dimensional vector space, in accordance with the notation. In this section we will be working with finite dimensional vector spaces only.

We note that any two finite dimensional vector spaces over $\mathbb{R}$ are isomorphic, since a bijection between the bases can be extended linearly to be an isomorphism between the two vector spaces. Hence, up to isomorphism, for every $n \in \mathbb{N}$ there is only one n-dimensional vector space, which is $\mathbb{R}^n$. However, vector spaces can also have extra structures that distinguish them from each other, as we shall explore now.

A **distance (metric)** on V is a function $d \colon V \times V \to \mathbb{R}$ satisfying:

- (positivity) $d(v, w) \geq 0$ for all $v, w \in V$, and $d(v, w) = 0$ if and only if $v = w$.
- (symmetry) $d(v, w) = d(w, v)$ for all $v, w \in V$.
- (triangle inequality) $d(v, w) \leq d(v, x) + d(x, w)$ for all $v, w, x \in V$.

The standard distance function on $\mathbb{R}^n$ is given by $d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$. Note that the notion of metric does not require a linear structure, or any other structure, on V; a metric can be defined on any set.

A similar concept that requires a linear structure on V is **norm**, which measures the "length" of vectors in V. Formally, a norm is a function $\|\cdot\| \colon V \to \mathbb{R}$ that satisfies the following three properties:

- (positivity) $\|v\| \geq 0$ for all $v \in V$, and $\|v\| = 0$ if and only if $v = 0$.
- (homogeneity) $\|rv\| = |r|\|v\|$ for all $r \in \mathbb{R}$ and $v \in V$.
- (subadditivity) $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$.

For example, the standard norm on $\mathbb{R}^n$ is $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$, which is also called the ℓ_2-norm. Also of interest is the ℓ_1-norm $\|x\|_1 = |x_1| + \cdots + |x_n|$, which we will study later in this class in relation to sparsity-based algorithms. We can also generalize these examples to any $p \geq 1$ to obtain the ℓ_p-norm, but we will not do that here.

Given a normed vector space $(V, \|\cdot\|)$, we can define the **distance (metric) function** on V to be $d(v, w) = \|v - w\|$. For example, the ℓ_2-norm on $\mathbb{R}^n$ gives the standard distance function

$$d(x, y) = \|x - y\|_2 = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2},$$

while the ℓ_1-norm on $\mathbb{R}^n$ gives the Manhattan/taxicab distance,

$$d(x, y) = \|x - y\|_1 = |x_1 - y_1| + \cdots + |x_n - y_n|.$$

As a side remark, we note that all norms on a finite dimensional vector space V are **equivalent**. This means that for any two norms μ and ν on V, there exist positive constants C_1 and C_2 such that for all $v \in V$, $C_1\mu(v) \leq \nu(v) \leq C_2\mu(v)$. In particular, continuity or convergence with respect to one norm implies continuity or convergence with respect to any other norms in a finite dimensional vector space. For example, on $\mathbb{R}^n$ we have the inequality $\|x\|_1/\sqrt{n} \leq \|x\|_2 \leq \|x\|_1$.

Another structure that we can introduce to a vector space is the inner product. An **inner product** on V is a function $\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{R}$ that satisfies the following properties:

- (symmetry) $\langle v, w \rangle = \langle w, v \rangle$ for all $v, w \in V$.
- (linearity) $\langle r_1 v_1 + r_2 v_2, w \rangle = r_1 \langle v_1, w \rangle + r_2 \langle v_2, w \rangle$ for all $r_1, r_2 \in \mathbb{R}$ and $v_1, v_2, w \in V$.
- (positive-definiteness) $\langle v, v \rangle \geq 0$ for all $v \in V$, and $\langle v, v \rangle = 0$ if and only if $v = 0$.

For example, the standard inner product on $\mathbb{R}^n$ is $\langle x, y \rangle = x_1 y_1 + \cdots + x_n y_n$, which is also known as the *dot product*, written $x \cdot y$.

Given an inner product space $(V, \langle \cdot, \cdot \rangle)$, we can define the norm of $v \in V$ to be $\|v\| = \sqrt{\langle v, v \rangle}$. It is easy to check that this definition satisfies the axioms for a norm listed above. On the other hand, not every norm arises from an inner product. The necessary and sufficient condition that has to be satisfied for a norm to be induced by an inner product is the **parallelogram law**:

$$\|v + w\|^2 + \|v - w\|^2 = 2\|v\|^2 + 2\|w\|^2.$$

If the parallelogram law is satisfied, then the inner product can be defined by **polarization identity**:

$$\langle v, w \rangle = \frac{1}{4}\Big(\|v + w\|^2 - \|v - w\|^2\Big).$$

For example, you can check that the ℓ_2-norm on $\mathbb{R}^n$ is induced by the standard inner product, while the ℓ_1-norm is not induced by an inner product since it does not satisfy the parallelogram law.

A very important result involving inner product is the following **Cauchy-Schwarz inequality**:

$$\langle v, w \rangle \leq \|v\|\|w\| \text{ for all } v, w \in V.$$

Inner product also allows us to talk about orthogonality. Two vectors v and w in V are said to be **orthogonal** if $\langle v, w \rangle = 0$. In particular, an **orthonormal basis** is a basis $v_1, \ldots, v_n$ that

is orthogonal ($\langle v_i, v_j \rangle = 0$ for $i \neq j$) and normalized ($\langle v_i, v_i \rangle = 1$). Given an orthonormal basis $v_1, \ldots, v_n$, the decomposition of $v \in V$ in terms of this basis has the special form

$$v = \sum_{i=1}^{n} \langle v, v_i \rangle v_i.$$

For example, the standard basis vectors $e_1, \ldots, e_n$ form an orthonormal basis of $\mathbb{R}^n$. In general, a basis $v_1, \ldots, v_n$ can be orthonormalized using the Gram-Schmidt process.

Given a subspace W of an inner product space V, we can define the **orthogonal complement** of W to be the set of all vectors in V that are orthogonal to W,

$$W^{\perp} = \{v \in V \mid \langle v, w \rangle = 0 \text{ for all } w \in W\}.$$

If V is finite dimensional, then we have the **orthogonal decomposition** $V = W \oplus W^{\perp}$. This means every vector $v \in V$ can be decomposed uniquely into $v = w + w'$, where $w \in W$ and $w' \in W^{\perp}$. The vector w is called the **projection** of v on W, and represents the unique vector in W that is closest to v.

1.2 Matrices

In addition to talking about vector spaces, we can also talk about operators on those spaces. A **linear operator** is a function $L \colon V \to W$ between two vector spaces that preserves the linear structure. In finite dimension, every linear operator can be represented by a matrix by choosing a basis in both the domain and the range, i.e. by working in coordinates. For this reason we focus the first part of our discussion on matrices.

If V is n-dimensional and W is m-dimensional, then a linear map $L \colon V \to W$ is represented by an $m \times n$ matrix A whose columns are the values of L applied to the basis of V. The **rank** of A is the dimension of the image of A, and the **nullity** of A is the dimension of the kernel of A. The **rank-nullity theorem** states that $\text{rank}(A) + \text{nullity}(A) = m$, the dimension of the domain of A. Also note that the transpose of A is an $n \times m$ matrix $A^{\top}$ satisfying

$$\langle Av, w \rangle_{\mathbb{R}^m} = (Av)^{\top} w = v^{\top} A^{\top} w = \langle v, A^{\top} w \rangle_{\mathbb{R}^n}$$

for all $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$.

Let A be an $n \times n$ matrix with real entries. Recall that an **eigenvalue** $\lambda \in \mathbb{R}$ of A is a solution to the equation $Av = \lambda v$ for some nonzero vector $v \in \mathbb{R}^n$, and v is the **eigenvector** of A corresponding to λ. If A is symmetric, i.e. $A^{\top} = A$, then the eigenvalues of A are real. Moreover, in this case the **spectral theorem** tells us that there is an orthonormal basis of $\mathbb{R}^n$ consisting of the eigenvectors of A. Let $v_1, \ldots, v_n$ be this orthonormal basis of eigenvectors, and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues. Then we can write

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^{\top},$$

which is called the **eigendecomposition** of A. We can also write this as

$$A = V \Lambda V^{\top},$$

where V is the $n \times n$ matrix with columns v_i, and Λ is the $n \times n$ diagonal matrix with entries λ_i. The orthonormality of $v_1, \ldots, v_n$ makes V an orthogonal matrix, i.e. $V^{-1} = V^{\top}$.

A symmetric $n \times n$ matrix A is **positive definite** if $v^\top A v > 0$ for all nonzero vectors $v \in \mathbb{R}^n$. A is **positive semidefinite** if the inequality is not strict (i.e. ≥ 0). A positive definite (resp. positive semidefinite) matrix A has positive (resp. nonnegative) eigenvalues.

Another method for decomposing a matrix is the **singular value decomposition** (SVD). Given an $m \times n$ real matrix A, the SVD of A is the factorization

$$A = U \Sigma V^\top,$$

where U is an $m \times m$ orthogonal matrix ($U^\top U = I$), Σ is an $m \times n$ diagonal matrix, and V is an $n \times n$ orthogonal matrix ($V^\top V = I$). The columns $u_1, \ldots, u_m$ of U form an orthonormal basis of $\mathbb{R}^m$, and the columns $v_1, \ldots, v_n$ of V form an orthonormal basis of $\mathbb{R}^n$. The diagonal elements $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$ in Σ are nonnegative and called the **singular values** of A. This factorization corresponds to the decomposition

$$A = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^\top.$$

This decomposition shows the relations between σ_i, u_i, and v_i more clearly: for $1 \leq i \leq \min\{m, n\}$,

$$\begin{aligned} A v_i &= \sigma_i u_i & \qquad A A^\top u_i &= \sigma_i^2 u_i \\ A^\top u_i &= \sigma_i v_i & \qquad A^\top A v_i &= \sigma_i^2 v_i \end{aligned}$$

This means the u_i's are eigenvectors of $AA^\top$ with corresponding eigenvalues σ_i^2, and the v_i's are eigenvectors of $A^\top A$, also with corresponding eigenvalues σ_i^2.

Given an $m \times n$ matrix A, we can define the **spectral norm** of A to be largest singular value of A,

$$\|A\|_{\text{spec}} = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(AA^\top)} = \sqrt{\lambda_{\max}(A^\top A)}.$$

Another common norm on A is the **Frobenius norm**,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} = \sqrt{\text{trace}(AA^\top)} = \sqrt{\text{trace}(A^\top A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}.$$

However, since the space of all matrices can be identified with $\mathbb{R}^{m \times n}$, the discussion in Section 1.1 still holds and all norms on A are equivalent.