Expectation, Variance, Covariance

If P is a probability distribution (rather than a general measure), then two more special cases of interest are obtained for particular choices of functions f in (B.15). If f is the identity on $\mathbb{R}^N$, we get the *expectation* $\mathbf{E}[x]$. If $f(x) = (x - \mathbf{E}[x])^2$ (on $\mathbb{R}$), we obtain the *variance* of x, denoted by $\mathrm{var}(x)$. In the N-dimensional case, the functions $f_{ij}(x) = (x_i - \mathbf{E}[x_i])(x_j - \mathbf{E}[x_j])$ lead to the *covariance* $\mathrm{cov}(x_i, x_j)$. For a data set $\{x_1, \ldots, x_m\}$, the matrix $(\mathrm{cov}(x_i, x_j))_{ij}$ is called the *covariance matrix*.

B.1.4 Stochastic Processes

A *stochastic process* y on a set $\mathcal{X}$ is a random quantity indexed by $x \in \mathcal{X}$. This means that for every x, we get a random quantity $y(x)$ taking values in $\mathbb{R}$, or more generally, in a *set* $\mathcal{R}$. A stochastic process is characterized by the joint probability distributions of y on arbitrary finite subsets of $\mathcal{X}$; in other words, of $(y(x_1), \ldots, y(x_m))$.[11]

A *Gaussian process* is a stochastic process with the property that for any $\{x_1, \ldots, x_m\} \subset \mathcal{X}$, the random quantities $(y(x_1), \ldots, y(x_m))$ have a joint Gaussian distribution with mean μ and covariance matrix K. The matrix elements K_{ij} are given by a covariance kernel $k(x_i, x_j)$.

When a Gaussian process is used for learning, the *covariance function* $k(x_i, x_j) := \mathrm{cov}(y(x_i), y(x_j))$ essentially plays the same role as the kernel in a SVM. See Section 16.3 and [587, 596] for further information.

B.2 Linear Algebra

B.2.1 Vector Spaces

We move on to basic concepts of linear algebra, which is to say the study of vector spaces. Additional detail can be found in any textbook on linear algebra (e.g., [170]). The feature spaces studied in this book have a rich mathematical structure, which arises from the fact that they allow a number of useful operations to be carried out on their elements: addition, multiplication with scalars, and the product between the elements themselves, called the dot product.

What's so special about these operations? Let us, for a moment, go back to our earlier example (Chapter 1), where we classify sheep. Surely, nobody would come up with the idea of trying to add two sheep, let alone compute their dot product. The set of sheep does not form a vector space; mathematically speaking, it could be argued that it does not have a very rich structure. However, as discussed in Chapter 1 (cf. also Chapter 2), it is possible to *embed* the set of all sheep into a dot product space such that we can think of the dot product as a measure of

11. Note that knowledge of the finite-dimensional distributions (fdds) does not yield complete information on the properties of the sample paths of the stochastic process; two different processes which have the same fdds are known as *versions* of one another.

the similarity of two sheep. In this space, we can perform the addition of two sheep, multiply sheep with numbers, compute hyperplanes spanned by sheep, and achieve many other things that mathematicians like.

Vector Space

Definition B.4 (Real Vector Space) *A set $\mathcal{H}$ is called a* vector space *(or* linear space*) over $\mathbb{R}$ if addition and scalar multiplication are defined, and satisfy (for all $\mathbf{x}, \mathbf{x}', \mathbf{x}'' \in \mathcal{H}$, and $\lambda, \lambda' \in \mathbb{R}$)*

$$\mathbf{x} + (\mathbf{x}' + \mathbf{x}'') = (\mathbf{x} + \mathbf{x}') + \mathbf{x}'', \tag{B.19}$$

$$\mathbf{x} + \mathbf{x}' = \mathbf{x}' + \mathbf{x} \in \mathcal{H}, \tag{B.20}$$

$$0 \in \mathcal{H},\ \mathbf{x} + 0 = \mathbf{x}, \tag{B.21}$$

$$-\mathbf{x} \in \mathcal{H},\ -\mathbf{x} + \mathbf{x} = 0, \tag{B.22}$$

$$\lambda\mathbf{x} \in \mathcal{H}, \tag{B.23}$$

$$1\mathbf{x} = \mathbf{x}, \tag{B.24}$$

$$\lambda(\lambda'\mathbf{x}) = (\lambda\lambda')\mathbf{x}, \tag{B.25}$$

$$\lambda(\mathbf{x} + \mathbf{x}') = \lambda\mathbf{x} + \lambda\mathbf{x}', \tag{B.26}$$

$$(\lambda + \lambda')\mathbf{x} = \lambda\mathbf{x} + \lambda'\mathbf{x}. \tag{B.27}$$

The first four conditions amount to saying that $(\mathcal{H}, +)$ is a commutative group.[12]

We have restricted ourselves to vector spaces over $\mathbb{R}$. The definition in the complex case is analogous, both here and in most of what follows. Any non-empty subset of $\mathcal{H}$ that is itself a vector space is called a *subspace* of $\mathcal{H}$.

Linear Combination

Among the things we can do in a vector space are *linear combinations*,

$$\sum_{i=1}^{m} \lambda_i \mathbf{x}_i, \text{ where } \lambda_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{H}, \tag{B.28}$$

Convex Combination

and *convex combinations*,

$$\sum_{i=1}^{m} \lambda_i \mathbf{x}_i, \text{ where } \lambda_i \geq 0, \sum_i \lambda_i = 1, \mathbf{x}_i \in \mathcal{H}. \tag{B.29}$$

Span

The set $\{\sum_{i=1}^{m} \lambda_i \mathbf{x}_i | \lambda_i \in \mathbb{R}\}$ is referred to as the *span* of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

A set of vectors $\mathbf{x}_i$, chosen such that none of the $\mathbf{x}_i$ can be written as a linear combination of the others, is called *linearly independent*. A set of vectors $\mathbf{x}_i$ that allows us to uniquely write each element of $\mathcal{H}$ as a linear combination is called a

Basis

basis of $\mathcal{H}$. For the uniqueness to hold, the vectors have to be linearly independent. All bases of a vector space $\mathcal{H}$ have the same number of elements, called the

Dimension

dimension of $\mathcal{H}$.

$\mathbb{R}^N$

The standard example of a finite-dimensional vector space is $\mathbb{R}^N$, the space of column vectors $([\mathbf{x}]_1, \ldots, [\mathbf{x}]_N)^\top$, where the $^\top$ denotes the transpose. In $\mathbb{R}^N$,

12. Note that (B.21) and (B.22) should be read as existence statements. For instance, (B.21) states that there exists an element, denoted by 0, with the required property.

addition and scalar multiplication are defined element-wise. The canonical basis of $\mathbb{R}^N$ is $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$, where for $j = 1, \dots, N$, $[\mathbf{e}_j]_i = \delta_{ij}$. Here δ_{ij} is the Kronecker symbol;

Kronecker δ_{ij}

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{B.30}$$

A somewhat more abstract example of a vector space is the space of all real-valued functions on a domain $\mathcal{X}$, denoted by $\mathbb{R}^{\mathcal{X}}$. Here, addition and scalar multiplication are defined by

$$(f + g)(x) := f(x) + g(x), \tag{B.31}$$

$$(\lambda f)(x) := \lambda f(x). \tag{B.32}$$

We shall return to this example below.

Linear Map

Linear algebra is the study of vector spaces and *linear maps* (sometimes called *operators*) between vector spaces. Given two real vector spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, the latter are maps

$$L : \mathcal{H}_1 \to \mathcal{H}_2 \tag{B.33}$$

that for all $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$, $\lambda, \lambda' \in \mathbb{R}$ satisfy

$$L(\lambda \mathbf{x} + \lambda' \mathbf{x}') = \lambda L(\mathbf{x}) + \lambda' L(\mathbf{x}'). \tag{B.34}$$

It is customary to omit the parentheses for linear maps; thus we normally write $L\mathbf{x}$ rather than $L(\mathbf{x})$.

Let us go into more detail, using (for simplicity) the case where $\mathcal{H}_1$ and $\mathcal{H}_2$ are identical, have dimension N, and are written $\mathcal{H}$. Due to (B.34), a linear map L is completely determined by the values it takes on a basis of $\mathcal{H}$. This can be seen by writing an arbitrary input as a linear combination in terms of the basis vectors $\mathbf{e}_j$, and then applying L;

$$L \sum_{j=1}^{N} \lambda_j \mathbf{e}_j = \sum_{j=1}^{N} \lambda_j L \mathbf{e}_j. \tag{B.35}$$

The image of each basis vector, $L\mathbf{e}_j$, is in turn completely determined by its expansion coefficients A_{ij}, $i = 1, \dots, N$;

$$L\mathbf{e}_j = \sum_{i=1}^{N} A_{ij} \mathbf{e}_i. \tag{B.36}$$

Matrix

The coefficients (A_{ij}) form the *matrix* A of L with respect to the basis $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$. We often think of linear maps as matrices in the first place, and use the same symbol to denote them. The *unit* (or *identity*) matrix is denoted by $\mathbf{1}$. Occasionally, we also use the symbol $\mathbf{1}$ as the identity map on arbitrary sets (rather than vector spaces).

In this book, we assume elementary knowledge of matrix algebra, including the

Matrix Product

matrix product, corresponding to the composition of two linear maps,

$$(AB)_{ij} = \sum_{n=1}^{N} A_{in} B_{nj}, \tag{B.37}$$

Transpose

and the *transpose* $(A^\top)_{ij} := A_{ji}$.

Inverse and Pseudo-Inverse

The *inverse* of a matrix A is written A^{-1} and satisfies $AA^{-1} = A^{-1}A = \mathbf{1}$. The *pseudo-inverse* $A^\dagger$ satisfies $AA^\dagger A = A$. While every matrix has a pseudo-inverse, not all have an inverse. Those which do are called *invertible* or *nonsingular*, and their inverse coincides with the pseudo-inverse. Sometimes, we simply use the notation A^{-1}, and it is understood that we mean the pseudo-inverse whenever A is not invertible.

B.2.2 Norms and Dot Products

Thus far, we have explained the linear structure of spaces such as the feature space induced by a kernel. We now move on to the *metric* structure. To this end, we introduce concepts of length and angles.

Definition B.5 (Norm) *A function* $\|\cdot\| : \mathcal{H} \to \mathbb{R}_0^+$ *that for all* $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$ *and* $\lambda \in \mathbb{R}$ *satisfies*

$$\|\mathbf{x} + \mathbf{x}'\| \leq \|\mathbf{x}\| + \|\mathbf{x}'\|, \tag{B.38}$$

$$\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|, \tag{B.39}$$

$$\|\mathbf{x}\| > 0 \text{ if } \mathbf{x} \neq 0, \tag{B.40}$$

Norm

is called a norm on $\mathcal{H}$. *If we replace the "$>$" in (B.40) by "$\geq$," we are left with what is called a* semi-norm.

Metric

Any norm defines a *metric* d via

$$d(\mathbf{x}, \mathbf{x}') := \|\mathbf{x} - \mathbf{x}'\|; \tag{B.41}$$

likewise, any semi-norm defines a *semi-metric*. The (semi-)metric inherits certain properties from the (semi-)norm, in particular the triangle inequality (B.39) and positivity (B.40).

While every norm gives rise to a metric, the converse is not the case. In this sense, the concept of the norm is stronger. Similarly, every *dot product* (to be introduced next) gives rise to a norm, but not vice versa.

Before describing the dot product, we start with a more general concept.

Definition B.6 (Bilinear Form) *A* bilinear form *on a vector space* $\mathcal{H}$ *is a function*

$$Q : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$$
$$(\mathbf{x}, \mathbf{x}') \to Q(\mathbf{x}, \mathbf{x}') \tag{B.42}$$

with the property that for all $\mathbf{x}, \mathbf{x}', \mathbf{x}'' \in \mathcal{H}$ *and all* $\lambda, \lambda' \in \mathbb{R}$, *we have*

$$Q((\lambda \mathbf{x} + \lambda' \mathbf{x}'), \mathbf{x}'') = \lambda Q(\mathbf{x}, \mathbf{x}'') + \lambda' Q(\mathbf{x}', \mathbf{x}''), \tag{B.43}$$

$$Q(\mathbf{x}'', (\lambda \mathbf{x} + \lambda' \mathbf{x}')) = \lambda Q(\mathbf{x}'', \mathbf{x}) + \lambda' Q(\mathbf{x}'', \mathbf{x}'). \tag{B.44}$$

If the bilinear form also satisfies

$$Q(\mathbf{x}, \mathbf{x}') = Q(\mathbf{x}', \mathbf{x}) \tag{B.45}$$

for all $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$*, it is called* symmetric.

Dot Product

Definition B.7 (Dot Product) *A* dot product *on a vector space* $\mathcal{H}$ *is a symmetric bilinear form,*

$$\begin{aligned} \langle ., . \rangle : \mathcal{H} \times \mathcal{H} &\to \mathbb{R} \\ (\mathbf{x}, \mathbf{x}') &\mapsto \langle \mathbf{x}, \mathbf{x}' \rangle , \end{aligned} \tag{B.46}$$

that is strictly positive definite*; in other words, it has the property that for all* $\mathbf{x} \in \mathcal{H}$*,*

$$\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \text{ with equality only for } \mathbf{x} = 0. \tag{B.47}$$

Definition B.8 (Normed Space and Dot Product Space) *A* normed space *is a vector space endowed with a norm; a* dot product space *(sometimes called* pre-Hilbert space*) is a vector space endowed with a dot product.*

Any dot product defines a corresponding norm via

$$\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}. \tag{B.48}$$

Cauchy-Schwarz

We now describe the *Cauchy-Schwarz inequality*: For all $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$,

$$| \langle \mathbf{x}, \mathbf{x}' \rangle | \leq \|\mathbf{x}\| \|\mathbf{x}'\|, \tag{B.49}$$

with equality occurring only if $\mathbf{x}$ and $\mathbf{x}'$ are linearly dependent. In some instances, the left hand side can be much smaller than the right hand side. An extreme case

Orthogonality

is when $\mathbf{x}$ and $\mathbf{x}'$ are *orthogonal*, and $\langle \mathbf{x}, \mathbf{x}' \rangle = 0$.

Basis Expansion

One of the most useful constructions possible in dot product spaces are *orthonormal basis expansions*. Suppose $\mathbf{e}_1, \ldots, \mathbf{e}_N$, where $N \in \mathbb{N}$, form an *orthonormal set*; that is, they are mutually orthogonal and have norm 1. If they also form a basis of $\mathcal{H}$, they are called an *orthonormal basis (ONB)*. In this case, any $\mathbf{x} \in \mathcal{H}$ can be written as a linear combination,

$$\mathbf{x} = \sum_{j=1}^{N} \langle \mathbf{x}, \mathbf{e}_j \rangle \, \mathbf{e}_j. \tag{B.50}$$

The standard example of a dot product space is again $\mathbb{R}^N$. We usually employ the canonical dot product,

$$\langle \mathbf{x}, \mathbf{x}' \rangle := \sum_{i=1}^{N} [\mathbf{x}]_i [\mathbf{x}']_i = \mathbf{x}^\top \mathbf{x}', \tag{B.51}$$

and refer to $\mathbb{R}^N$ as the *Euclidean space of dimension* N. Using this dot product and the canonical basis of $\mathbb{R}^N$, each coefficient $\langle \mathbf{x}, \mathbf{e}_j \rangle$ in (B.50) just picks out one entry from the column vector $\mathbf{x}$, thus $\mathbf{x} = \sum_{j=1}^{N} [\mathbf{x}]_j \mathbf{e}_j$.

Pythagorean Theorem

A rather useful result concerning norms arising from dot products is the *Pythagorean Theorem*. In its general form, it reads as follows:

Theorem B.9 (Pythagoras) *If* $\mathbf{e}_1, \ldots, \mathbf{e}_q$ *are orthonormal (they need not form a basis), then*

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{q} \langle \mathbf{x}, \mathbf{e}_i \rangle^2 + \left\| \mathbf{x} - \sum_{i=1}^{q} \langle \mathbf{x}, \mathbf{e}_i \rangle \mathbf{e}_i \right\|^2. \tag{B.52}$$

Now that we have a dot product, we are in a position to summarize a number of useful facts about matrices.

- It can readily be verified that for the canonical dot product, we have

$$\langle \mathbf{x}, A\mathbf{x}' \rangle = \left\langle A^\top \mathbf{x}, \mathbf{x}' \right\rangle \tag{B.53}$$

for all $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$

Symmetric Matrices

- Matrices A such that $A = A^\top$ are called *symmetric*. Due to (B.53), they can be swapped between the two arguments of the canonical dot product without changing its value
- Symmetric matrices A that satisfy

$$\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0 \tag{B.54}$$

for all $\mathbf{x} \in \mathcal{H}$ are called *positive definite* (cf. Remark 2.16 for a note on this terminology)

- Another interesting class of matrices are the *unitary* (or *orthogonal*) matrices. A unitary matrix U is characterized by an inverse U^{-1} that equals its transpose $U^\top$. Unitary matrices thus satisfy

$$\langle U\mathbf{x}, U\mathbf{x}' \rangle = \left\langle U^\top U\mathbf{x}, \mathbf{x}' \right\rangle = \left\langle U^{-1}U\mathbf{x}, \mathbf{x}' \right\rangle = \langle \mathbf{x}, \mathbf{x}' \rangle \tag{B.55}$$

for all $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$; in other words, they leave the canonical dot product invariant

- A final aspect of matrix theory of interest in machine learning is matrix diagonalization. Suppose A is a linear operator. If there exists a basis $\mathbf{v}_1, \ldots, \mathbf{v}_N$ of $\mathcal{H}$ such that for all $i = 1, \ldots, N$,

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i, \tag{B.56}$$

Eigenvalue and Eigenvector

with $\lambda_i \in \mathbb{R}$, then A can be *diagonalized*: written in the basis $\mathbf{v}_1, \ldots, \mathbf{v}_N$, we have $A_{ij} = 0$ for all $i \neq j$ and $A_{ii} = \lambda_i$ for all i. The coefficients λ_i are called *eigenvalues*, and the $\mathbf{v}_i$ *eigenvectors*, of A

Let us now consider the special case of symmetric matrices. These can always be diagonalized, and their eigenvectors can be chosen to form an orthonormal basis with respect to the canonical dot product. If we form a matrix V with these eigenvectors as columns, then we obtain the diagonal matrix as $VAV^\top$.

Rayleigh's principle states that the smallest eigenvalue $\lambda_{\min}$ coincides with the

minimum of

$$R(\mathbf{v}) := \frac{\langle \mathbf{v}, A\mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}. \tag{B.57}$$

The minimizer of R is an eigenvector with eigenvalue $\lambda_{\min}$. Likewise, the largest eigenvalue and its corresponding eigenvector can be found by maximizing R.
Functions $f : I \to \mathbb{R}$, where $I \subset \mathbb{R}$, can be defined on symmetric matrices A with eigenvalues in I. To this end, we diagonalize A and apply f to all diagonal elements (the eigenvalues).
Since a symmetric matrix is positive definite if and only if all its eigenvalues are nonnegative, we may choose $f(x) = \sqrt{x}$ to obtain the unique *square root* $\sqrt{A}$ of a positive definite matrix A.

Many statements about matrices generalize in some form to operators on spaces of arbitrary dimension; for instance, Mercer's theorem (Theorem 2.10) can be viewed as a generalized version of a matrix diagonalization, with eigenvectors (or eigenfunctions) ψ_j satisfying $\int_{\mathcal{X}} k(x, x')\psi_j(x')\, d\mu(x') = \lambda_j \psi_j(x)$.

B.3 Functional Analysis

Functional analysis combines concepts from linear algebra and analysis. Consequently, it is also concerned with questions of convergence and continuity. For a detailed treatment, cf. [429, 306, 112].

Cauchy Sequence

Definition B.10 (Cauchy Sequence) *A sequence* $(\mathbf{x}_i)_i := (\mathbf{x}_i)_{i \in \mathbb{N}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots)$ *in a normed space* $\mathcal{H}$ *is said to be a* Cauchy sequence *if for every* $\epsilon > 0$*, there exists an* $n \in \mathbb{N}$ *such that for all* $n', n'' > n$*, we have* $\|\mathbf{x}_{n'} - \mathbf{x}_{n''}\| < \epsilon$.

A Cauchy sequence is said to converge *to a point* $\mathbf{x} \in \mathcal{H}$ *if* $\|\mathbf{x}_n - \mathbf{x}\| \to 0$ *as* $n \to \infty$.

Banach / Hilbert Space

Definition B.11 (Completeness, Banach Space, Hilbert Space) *A space* $\mathcal{H}$ *is called* complete *if all Cauchy sequences in the space converge.*

A Banach space *is a complete normed space; a* Hilbert space *is a complete dot product space.*

The simplest example of a Hilbert space (and thus also of a Banach space) is again $\mathbb{R}^N$. More interesting Hilbert spaces, however, have *infinite* dimensionality. A number of surprising things can happen in this case. To prevent the nasty ones, we generally assume that the Hilbert spaces we deal with are *separable*,[13] which means that there exists a countable dense subset. A *dense subset* is a set S such that each element of $\mathcal{H}$ is the limit of a sequence in S. Equivalently, the completion of

13. One of the positive side effects of this is that we essentially only have to deal with one Hilbert space: all separable Hilbert spaces are equivalent, in a sense that we won't define presently.