1 Mathematical basics

1.1 Sets, vectors, matrices

A set of k items can be represented as $\{i_1, i_2, \ldots, i_k\}$. We can also represent this set by $\{i_l\}_{l=1}^k$. In our course, i_l, for $l = 1, 2, \ldots, k$ is often a real number or an integer. The Cartesian product of two sets S_1, S_2 is written as $S_1 \times S_2$ and is defined as $\{s = (s_1, s_2) \,|\, s_1 \in S_1, s_2 \in S_2\}$. This definition can be readily generalized to the Cartesian product of K sets. If we take the Cartesian product of a set with itself k times, we can use the notation S^k.

The set of real numbers is $\mathbb{R}$. The set of n dimensional real numbers is $\mathbb{R}^n$. An open interval is represented as $(a, b) \subset \mathbb{R}$, with $a, b \in \mathbb{R}$, $a < b$. A vector is an ordered finite list of numbers. If there are n elements in the list, then the vector is in $\mathbb{R}^n$ and it is written as a vertical array, surrounded by square or curved brackets, e.g. $a \in \mathbb{R}^n$ is written as

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \quad \text{or} \quad \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

In the above, a_i, $i = 1, \ldots, n$ are referred to as elements, entries or components of the vector a. The vector $a \in \mathbb{R}^n$ can also be written as the elements separated by commas and surrounded by parentheses, $a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$. Observe that this notation for vectors can be confusing, as (a_1, a_2) can also refer to an open interval. However, the former is $(a_1, a_2) \in \mathbb{R}^2$, whereas the latter is $(a_1, a_2) \subset \mathbb{R}$. Usually, it is clear from the context which of the two is meant, but it is a good practice to always write the domain of objects to avoid such confusion.

A matrix $A \in \mathbb{R}^{m \times n}$ is a rectangular array of numbers with m rows and n columns. It is written between rectangular brackets or parentheses:

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ & \vdots & & \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \quad \text{or} \quad \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ & \vdots & & \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

The element a_{ij} refers to the entry of the matrix at the i-th row and j-th column. A^T denotes the transpose of the matrix A. Observe that a vector in $\mathbb{R}^n$ can be interpreted as an n by 1 matrix. A matrix with only one row, that is, size 1 by n, is called a row vector. The inner product between vectors $a, b \in \mathbb{R}^n$ is written as $a^T b$.

You are expected to be familiar with the basics of matrix and vector operations. For a review of basic linear algebra concepts, see Chapters 1 and 6 here. Furthermore, you are required to be familiar with eigenvalues, determinant and inverse of a matrix.

Exercise. For $a = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $M = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, compute the following:

- $a^T b$, $b^T a$, $a^T M$, Ma, $\sqrt{a^T a}$, $\sqrt{b^T b}$.
- the determinant of M.
- the eigenvalues of M.
- the inverse of M if it exists.

Solution. The results of the products are

$$a^T b = \begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = 0*1+1*1=1,$$
$$b^T a = \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} = 1*0+1*1=1,$$
$$a^T M = \begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\1 & 1\end{bmatrix} = \begin{bmatrix}(0*1+1*1) & (0*0+1*1)\end{bmatrix} = \begin{bmatrix}1 & 1\end{bmatrix},$$
$$Ma = \begin{bmatrix}1 & 0\\1 & 1\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}(1*0+0*1)\\(1*0+1*1)\end{bmatrix} = \begin{bmatrix}0\\1\end{bmatrix},$$
$$\sqrt{a^T a} = \sqrt{\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix}} = \sqrt{0*0+1*1} = 1,$$
$$\sqrt{b^T b} = \sqrt{\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}} = \sqrt{1*1+1*1} = \sqrt{2}.$$

The determinant of a 2×2 matrix $\begin{bmatrix}a & b\\c & d\end{bmatrix}$ is $ad-bc$. Thus, the determinant of M is

$$\det(M) = 1*1-0*1=1, \tag{1}$$

The eigenvalues of M are the values λ, such that the determinant of the matrix $M-\lambda I = \begin{bmatrix}1-\lambda & 0\\1 & 1-\lambda\end{bmatrix}$ is equal to zero:

$$\det(M-\lambda I) = (1-\lambda)*(1-\lambda)-0*1 = (1-\lambda)^2 = 0. \tag{2}$$

The only eigenvalue is $\lambda = 1$ and it has multiplicity 2.
The inverse of a 2×2 matrix $A = \begin{bmatrix}a & b\\c & d\end{bmatrix}$ exists if $\det(A)\neq 0$ and is equal to

$$A^{-1} = \frac{1}{\det(A)}\begin{bmatrix}d & -b\\-c & a\end{bmatrix}.$$

The determinant of M is different from zero according to (1), and its inverse is equal to

$$M^{-1} = \frac{1}{\det(M)}\begin{bmatrix}1 & 0\\-1 & 1\end{bmatrix} = \begin{bmatrix}1 & 0\\-1 & 1\end{bmatrix}.$$

1.2 Functions and linearity

A function from a *domain* D to a *co-domain* C is written as $f: D\to C$. This function takes an element $d\in D$ and returns a unique element $f(d)\in C$. The co-domain is sometimes referred to as the range of the function. However, in some texts, the range refers specifically to the set $\{c\in C \,|\, c = f(d) \text{ for some } d\in D\}$. For example, the *softmax function*, often used in machine learning, and denoted by $\sigma:\mathbb{R}^K\to\mathbb{R}^K$ is given as

$$\sigma(z) = \left(\frac{e^{z_1}}{\sum_{i=1}^K e^{z_i}}, \frac{e^{z_2}}{\sum_{i=1}^K e^{z_i}}, \ldots, \frac{e^{z_K}}{\sum_{i=1}^K e^{z_i}}\right).$$

Its range is $(0,1)^K \subset \mathbb{R}^K$. Recall: $(0,1)^K$ is the Cartesian product of $(0,1)$ with itself K times.

A *linear function* $f : \mathbb{R}^n \to \mathbb{R}^m$ is a function such that for any $a, b \in \mathbb{R}^n$, $\alpha, \beta \in \mathbb{R}$, $f(\alpha a + \beta b) = \alpha f(a) + \beta f(b)$. An affine function $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear function with an offset. That is, f is affine if there exists $c \in \mathbb{R}^m$ such that $f - c$ is linear.

Exercise. (a) Verify that for any $w \in \mathbb{R}^{d+1}$, the function $f : \mathbb{R}^d \to \mathbb{R}$, defined as $f(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$ is affine; (b) verify that the function $f : \mathbb{R}^d \to \mathbb{R}^m$ defined as $f(x) = Ax$ is linear. And hence, the function $f(x) = Ax + b$, with $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ is affine.

Solution. (a) To prove that the function f is affine, we need to prove that there exists $c \in \mathbb{R}$ for which $f - c$ is linear. In this case, the offset is w_0, thus we can simply show that $f(x) - w_0$ is linear:

$$\begin{aligned} f(\alpha a + \beta b) - w_0 =& w_1(\alpha a_1 + \beta b_1) + \cdots + w_d(\alpha a_d + \beta b_d) \\ =& \alpha(w_1 a_1 + \cdots + w_d a_d) + \beta(w_1 b_1 + \cdots + w_d b_d) \\ =& \alpha\left(f(a) - w_0\right) + \beta\left(f(b) - w_0\right). \end{aligned}$$

(b) We need to prove that $f(\alpha a + \beta b) = \alpha f(a) + \beta f(b)$ for any $a, b \in \mathbb{R}^d$, $\alpha, \beta \in \mathbb{R}$:

$$\begin{aligned} f(\alpha a + \beta b) =& A(\alpha a + \beta b) \\ =& \alpha A a + \beta A b \\ =& \alpha f(a) + \beta f(b). \end{aligned}$$

We showed that $f(x) = Ax$ is linear. To show that $f(x) = Ax + b$ is affine, we need to show that there exists $c \in \mathbb{R}^m$ such that $f(x) - c$ is linear. By selecting $c = b$ we obtain $f(x) - c = f(x) - b = Ax + b - b = Ax$, which we proved to be a linear function. Thus $f(x) = Ax + b$ is affine.

Advanced. Observe that the above exercises shows that every matrix gives rise to a linear function. For this reason, matrices are also referred to as linear maps. The converse of the above statement can also be shown. Namely, every linear function mapping finite dimensional domain to a finite dimensional co-domain can be represented by a matrix.

Linear functions can be defined on more general spaces than the Euclidean spaces (example, space of continuously differentiable functions, which is infinite dimensional). Linear functions between infinite dimensional spaces that you have seen before include differentiation and integration. This is useful for studying linear dynamical systems or optimization problems. However, to verify this fact and appreciate its usefulness, you would first need to learn about infinite dimensional vector spaces, which is beyond this course.

1.3 Supervised learning basics

Consider a dataset $\{(x^i, y^i)\}_{i=1}^N$ with N data points.

- The *independent variables* are denoted by $x^1, x^2, \ldots, x^N$, with $x^i = (x_1^i, x_2^i, \ldots, x_d^i)$ being a vector in $\mathbb{R}^d$. Here, x_j^i for $j = 1, \ldots, d$ is referred to as a component, field or a feature of the vector. The vectors $x^i \in \mathbb{R}^d$ are also sometimes referred to as predictors, covariates, explanatory variables or feature vectors.

- The *labels* or *dependent variables* are denoted by $y^1, y^2, \ldots, y^N$. They are also referred to as targets, outcomes or response variables. In a classification problem $y^i \in \{1, 2, \ldots, K\}$ where K denotes the number of classes, also referred to as categories. In a regression problem, $y^i \in \mathbb{R}^m$. However, often we consider the case of $m = 1$, because we learn a model for each component of vector y separately.

- We use the notation $\Phi : \mathbb{R}^d \to \mathbb{R}^p$ to denote a so-called *feature mapping*. This is simply a transformation of the data. For example, for $x \in \mathbb{R}^2$, we can define $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$ as follows. We can consider $\Phi_1(x) = 1$ as the constant feature, $\Phi_2(x) = x_1^2$ as a quadratic mapping for the first coordinate of x and $\Phi_3(x) = \sin(x_2)$ as a sinusoidal mapping of the second coordinate of x. We can represent this feature mapping by $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$, with $\Phi = (\Phi_1, \Phi_2, \Phi_3)$ and $\Phi : x \mapsto (1, x_1^2, \sin(x_2))$.
- We often use the constant feature to augment our independent variables from $x \in \mathbb{R}^d$ to $\bar{x} = (1, x_1, x_2, \ldots, x_d) \in \mathbb{R}^{d+1}$. Thus, an affine function $f(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$ can be written as a linear function with domain $\mathbb{R}^{d+1}$ by defining $w = (w_0, w_1, w_2, \ldots, w_d) \in \mathbb{R}^{d+1}$ and writing $f(x) = w^T \bar{x}$. As almost always we consider the extension of the feature vector by constant feature, with an abuse of notation, we often write $f(x) = w^T x$ instead of $w^T \bar{x}$.

Exercise. Define $X \in \mathbb{R}^{N \times (d+1)}$ as a matrix whose rows are the data vectors: $\bar{x}^i = (1, x_1^i, x_1^i, \ldots, x_d^i)$ for $i = 1, 2, \ldots, N$. With this notation, an affine function $f : \mathbb{R}^{d+1} \to \mathbb{R}$ acting on each data point gives rise to $\hat{y}^i = w^T \bar{x}^i$, $i = 1, \ldots, N$.

- Let $\hat{y}$ denote$(\hat{y}^1, \hat{y}^2, \ldots, \hat{y}^N) \in \mathbb{R}^N$. Verify that $\hat{y} = Xw$.

  Solution: Since each component of $\hat{y}$ is computed as $\hat{y}^i = w^T \bar{x}^i = (\bar{x}^i)^T w$, we can express $\hat{y}$ as:

$$\hat{y} = \begin{bmatrix} \hat{y}^1 \\ \hat{y}^2 \\ \vdots \\ \hat{y}^N \end{bmatrix} = \begin{bmatrix} (\bar{x}^1)^T w \\ (\bar{x}^2)^T w \\ \vdots \\ (\bar{x}^N)^T w \end{bmatrix} = \begin{bmatrix} (\bar{x}^1)^T \\ (\bar{x}^2)^T \\ \vdots \\ (\bar{x}^N)^T \end{bmatrix} w = Xw.$$

- Let y^i denote the true label of the data $x^i \in \mathbb{R}^d$. Verify that we can then write the error in predicting y^i using a linear model as $e^i = \hat{y}^i - y^i$, and thus, the error vector, namely, the vector of errors for each data point as $e = Xw - y$.

  Solution: The error for predicting y^i using the linear model is: $e^i = \hat{y}^i - y^i$. Therefore, the error vector $e \in \mathbb{R}^N$ is the difference between the vector of predicted values $\hat{y}$ and the true labels $y = (y^1, y^2, \ldots, y^N) \in \mathbb{R}^N$:

$$e = \hat{y} - y = Xw - y$$

  Thus, the error vector $e = Xw - y$ represents the difference between the predicted values and the actual labels for each data point.

Exercise. If each label y^i is a vector in $\mathbb{R}^m$, we would learn an affine function for each element of y^i. Thus, we would have m weight vectors for our linear predictor. We use the notation $w_{l,j}$ to refer to the j-th entry of the l-th weight vector $w_l \in \mathbb{R}^{d+1}$. Thus, we have $j = 0, 1, \ldots, d$ and $l = 1, 2, \ldots, m$. Verify that with this notation $y_l^i = w_l^T \bar{x}^i$, for $l \in \{1, \ldots, m\}$ and $i \in \{1, \ldots, N\}$.

Solution: In this case, we are predicting multiple outputs for each data point, which means that each label $y^i = (y_1^i, y_2^i, \ldots, y_m^i) \in \mathbb{R}^m$ is a vector with m components. The prediction for the l-th component, $\hat{y}_l^i$, is given by:

$$\hat{y}_l^i = w_{l,0} + w_{l,1} x_1^i + w_{l,2} x_2^i + \cdots + w_{l,d} x_d^i = w_l^T \bar{x}^i.$$