Mathematical Foundations for Machine Learning | MathFML

Summary

CONTENTS

1. LINEAR ALGEBRA

Term 1 : **Linear combination**

$$\sum_{i=1}^{k} \lambda_i \vec{v}_i$$

1.1. STANDARD BASIS

Basis vectors $\{\vec{e}_1, \vec{e}_2, ..., \vec{e}_n\}$ in $\mathbb{R}^n$ where $\vec{e}_i = \left(0, 0, ..., \underbrace{1}_{i\text{th position}}, ..., 0\right)^T$

Example 1

$$\left[\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right] = \mathbb{R}^3$$

It can be shown that any basis of $\mathbb{R}^n$ consists of exactly n linear independent vectors and that conversely any set of n linear independent vectors is a basis of $\mathbb{R}^n$.

1.2. VECTORS

TODO:
group and clean this up

$$\lambda\vec{0} = \vec{0}$$
$$\vec{v} + \vec{0} = \vec{v}$$
$$-\vec{v} = -1 \cdot \vec{v}$$
$$-\vec{v} + \vec{v} = \vec{0}$$
$$(\lambda\mu)\vec{v} = \lambda\left(\mu\vec{v}\right) = \lambda\mu\vec{v}$$
$$\lambda\left(\vec{v} + \vec{w}\right) = \lambda\vec{v} + \lambda\vec{w}$$
$$\vec{v} + \left(\vec{u} + \vec{w}\right) = \left(\vec{v} + \vec{u}\right) + \vec{w} = \vec{v} + \vec{u} + \vec{w}$$

1.3. MATRICES

Term	Rules
Associativity	$(A + B) + C = A + (B + C) = A + B + C$ $(A \cdot B) \cdot C = A \cdot (B \cdot C) = A \cdot B \cdot C$
Distributivity	$C \cdot (A + B) = C \cdot A + C \cdot B$ $(A + B) \cdot C = A \cdot C + B \cdot C$
Commutativity	$A + B = B + A$ $A \cdot B \neq B \cdot A$

Term	Rules
Transposing	$(A^T)^T = A$ $(A+B)^T = A^T + B^T$ $(\lambda A)^T = \lambda A^T$ $(A \cdot B)^T = B^T \cdot A^T$ $A_{ij} = A^T_{ji}$
Identity matrix	$\mathbb{1} \cdot A = A \cdot \mathbb{1} = A$ for $A \in \mathbb{R}^{n \times n}$
Inverting	$A \cdot A^{-1} = A^{-1} \cdot A = \mathbb{1}$
Determinate	$\det(\lambda M) = \lambda^n \det(M), M \in \mathbb{R}^{n \times n}$ $\det\begin{pmatrix} A & * \\ 0 & B \end{pmatrix} = \det(A) \cdot \det(B)$ $\det(A \cdot B) = \det(A) \cdot \det(B)$ $\det(A^{-1}) = \frac{1}{\det(A)}$ $\det(A^T) = \det(A)$
Scalars	$(\lambda + \mu)A = \lambda A + \mu A$ $(\lambda\mu)A = \lambda(\mu A)$ $\lambda(B + C) = \lambda B + \lambda C$ $\lambda(BC) = (\lambda B)C = B(\lambda)C$

1.3.1. Unit matrices

Let M be the set of all 2×2 matrices. How could the basis of M look? $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$

Unit matrices:

$$E_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, E_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, E_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, E_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

$$M = \{E_{11}, E_{12}, E_{21}, E_{22}\}$$

$$(E_{11})_{22} = 0, (E_{11})_{11} = 1$$

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^3, (\vec{e}_1)_1 = 1$$

1.3.2. Kronecker Delta

$$\delta_{ij} = \begin{cases} 1, i = j \\ 0, \text{otherwise} \end{cases}$$

Example 2

$$\delta_{23} = 0, \delta_{11} = 1, (\vec{e}_1)_j = \delta_{1j} = \begin{cases} 1, 1 = j \\ 0, \text{otherwise} \end{cases}$$

$$\delta_{m1}\delta_{m2} = 0$$
$$\delta_{kl} - \delta_{lk} = 0$$
$$(e_k)_i = \delta_{ki} = \delta_{ik}$$
$$(E_{kl})_{ij} = \delta_{ki}\delta_{lj}$$
$$(E_{rst})_{ijk} = \delta_{ri}\delta_{sj}\delta_{tk}$$

1.3.3. Matrix product

$$A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times r}, A \cdot B \in \mathbb{R}^{n\times r}$$

$$A \cdot B = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{im} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ b_{i1} & b_{i2} & \dots & b_{ir} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mr} \end{pmatrix}$$

$$(A \cdot B)_{11} = a_{11}b_{11} + a_{12}b_{12} + \dots + a_{1m}b_{m1} = \sum_{k=1}^{m} a_{1k}b_{k1}$$

$$(A \cdot B)_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$$

Shorthand:

$$(A \cdot B)_{ij} = \sum_{k} a_{ik}b_{kj}$$

Example 3 : **Let X be a matrix for which X^{-1} exists. Prove that the inverse of X^T exists and is $(X^{-1})^T$**

$$\mathbb{1}^T = \mathbb{1}$$
$$\Rightarrow (X^{-1}X)^T = \mathbb{1}$$
$$\Leftrightarrow X^T(X^{-1})^T = \mathbb{1}$$

1.3.4. Tensors

Vectors = tensors of rank 1, matrices = tensors of rank 2

The standart basis for $\mathbb{R}^{n\times m\times l}$ consists of $n \cdot m \cdot l$ basis vectors E_{ijk} where $1 \le i \le n, 1 \le j \le m, 1 \le k \le l$, such that all components of E_{ijk} are zero except at position i, j, k where the value of the component is 1.

2. FUNCTIONS OF SEVERAL VARIABLES

Let D and R be two sets. A mapping $f : D \to R$ that associates to each element $x \in D$ exactly one element of $f(x) \in R$ is called function.

Term	*Definition*
Domain of f	D

Term	*Definition*
Range of f	R
Real valued function	$R \subset \mathbb{R}$
Vector valued function	$R \subset \mathbb{R}^m$
A function of n variables	$D \subset \mathbb{R}^n$

2.1. IMAGE CLASSIFICATION

A **128 × 256** pixel image I with **3** color channels (RGB) can be represented as a **128 × 256 × 3** matrix (tensor of rank 3). The red value of a pixel at row **100**, column **12** can be indexed with $I_{100,12,1}$. A function to determine whether the image was taken at day (**0**) or night (**1**) would have the following signature:

$$f : \mathbb{R}^{128\times256\times3} \to \{0, 1\}$$

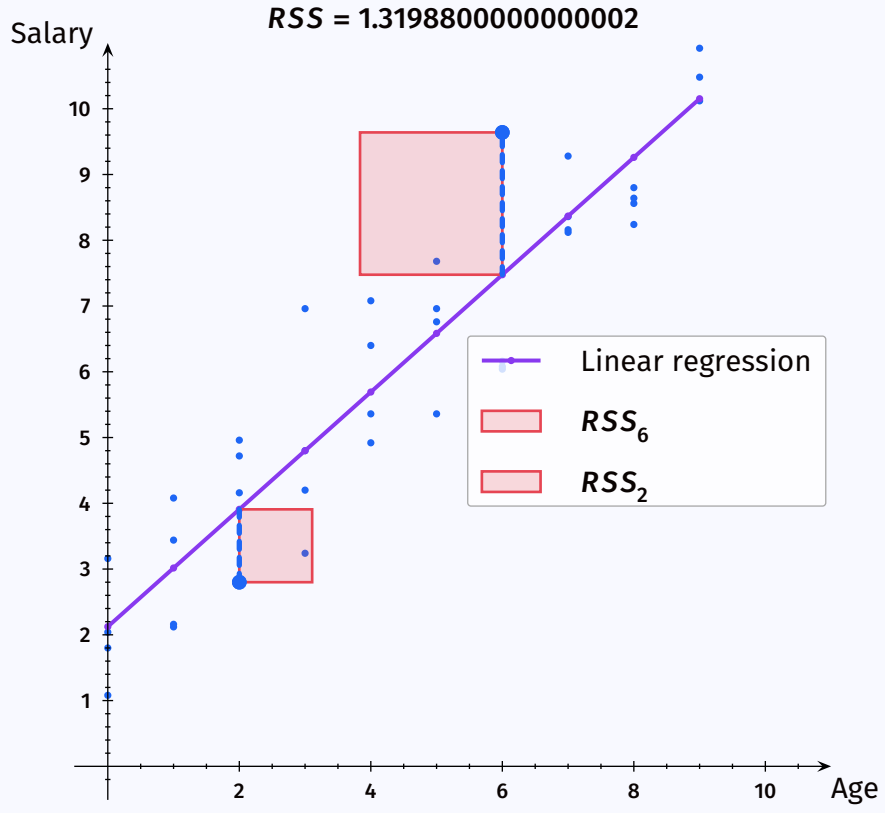
2.2. RESIDUAL SUM OF SQUARE

Residual sum of square (RSS) is a statistical method that helps identify the level of discrepancy in a dataset not predicted by a regression model. Thus, it measures the variance in the value of the observed data when compared to its predicted value as per the regression model.

$$RSS = \sum_{i=1}^{N} \underbrace{(y_i - f(x_i))^2}_{\text{Represented as red squares in the example}}$$

$$RSS(a, b) = \sum_{i=1}^{N} (y_i - (mx_i + b))^2 \geq 0, RSS : \mathbb{R}^2 \to \mathbb{R}$$

Example 4 : **Linear regression of salaries by age**



2.3. VISUALIZING FUNCTIONS

If $f : D \to R$ is a function from $D \subset \mathbb{R}^n$ to $R \subset \mathbb{R}^m$ its graph is defined as:

$$\text{graph}\,(f) = \left\{\left(\vec{x}, \vec{y}\right) \in D \times R \;\middle|\; x \in D \wedge \vec{y} = f\left(\vec{x}\right)\right\} \subset D \times R \subset \mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^{n+m}$$

Due to the fact that the visual imagination of humans is limited to at most 3 dimensions, the graph of a function is a useful concept for graphical illustration if and only if $n + m \leq 3$.

Example 5

The graph of the function

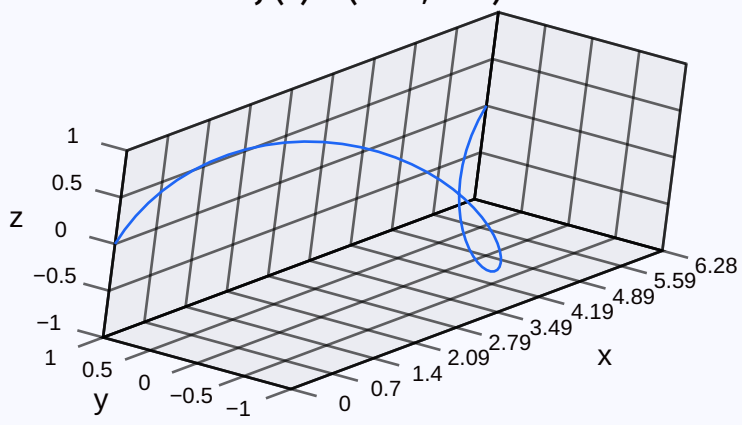
$$f : \begin{cases} \mathbb{R} \to \mathbb{R}^2 \\ t \mapsto (\cos t, \sin t) \end{cases}$$

is

$$\begin{aligned} \text{graph}\,(f) &= \left\{(t, x, y) \in \mathbb{R}^3 \;\middle|\; (x, y) = f(t) \wedge t \in \mathbb{R}\right\} \\ &= \left\{(t, \cos t, \sin t) \in \mathbb{R}^3 \;\middle|\; t \in \mathbb{R}\right\} \end{aligned}$$

It is displayed below and illustrates how f maps t to the corresponding values

$$f(t) = (\cos t, \sin t)$$



Example 6

The graph of the function

$$g : \begin{cases} \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\} \to \mathbb{R}^2 \\ (x,y) \qquad\qquad\qquad\qquad \mapsto \sqrt{1 - x^2 - y^2} \end{cases}$$

is

$$\begin{aligned} \text{graph}\,(g) &= \left\{ (x,y,z) \in \mathbb{R}^3 \;\middle|\; z = \sqrt{1 - x^2 - y^2} \wedge x^2 + y^2 \leq 1 \right\} \\ &= \left\{ \left(x, y, \sqrt{1 - x^2 - y^2}\right) \in \mathbb{R}^3 \;\middle|\; x^2 + y^2 \leq 1 \right\} \end{aligned}$$

It is displayed below and illustrates how f maps (x,y) to the corresponding values

2.3.1. Curves and surfaces

Term	*Definition*
Curves	Vector valued functions of one variable, i.e. they map one-dimensional input to multi-dimensional output
n-dimensional curve	A continuous function $f : I \to Z$ that maps an interval $I \subset \mathbb{R}$ into a subset $Z \subset \mathbb{R}^n$
Surfaces	Real valued functions of multiple variables, i.e. they map multi-dimensional input to one dimensional output
n-dimensional hyper-surface	A continuous real valued function $f : D \to Z$ that maps a sufficiently large subset $D \subset \mathbb{R}^n$ into a subset $Z \subset \mathbb{R}$ If $n = 2$ a hypersurface is also simply called surface. If $n = 1$ a hyper-surface is simply a continuous real valued function of one variable.

The graph of both, an n-dimensional curve and an n-dimensional hyper-surface, is a subset of $\mathbb{R}^{n+1}$. A graphical illustration of such a graph therefore requires $n \leq 2$. However, unlike surfaces which are typically illustrated as in Figure 1 , a curve $c : I \to \mathbb{R}^n$ is usually not illustrated through

$$\text{graph}\, c = \{(t,y) \in R \times R^n \mid y = c(t) \wedge t \in I\}$$

Instead curves are typically visualized by skipping the independent variable t from the graph, i.e. by visualizing the image of the curve

$$c(I) = \{y \in R^n \mid y = c(t) \wedge t \in I\}$$

which means that we are projecting the graph of the curve onto the gray plane that is spanned from the dependent variables. As this kind of plot saves one dimension we can also visualize curves for which the range Z is a subset of $\mathbb{R}^3$.

TODO:
pt3d bugfixes

Example 7

The graph of the function

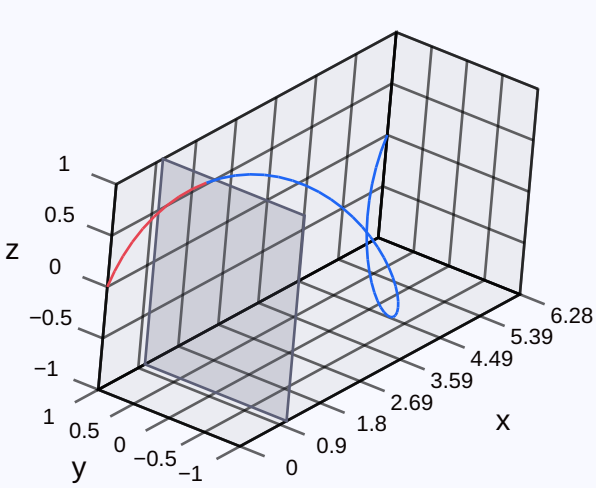
$$f : \begin{cases} \mathbb{R} \to \mathbb{R}^2 \\ t \mapsto (\cos t, \sin t) \end{cases}$$

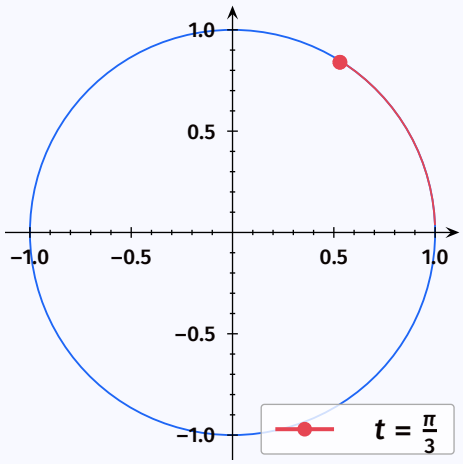
is

$$\begin{aligned} \text{graph}(f) &= \{(t, x, y) \in \mathbb{R}^3 \mid (x, y) = f(t) \wedge t \in \mathbb{R}\} \\ &= \{(t, \cos t, \sin t) \in \mathbb{R}^3 \mid t \in \mathbb{R}\} \end{aligned}$$

and its image is the set

$$\begin{aligned} f(\mathbb{R}) &= \{f(t) \in \mathbb{R}^2 \mid t \in \mathbb{R}\} \\ &= \{(\cos t, \sin t) \in \mathbb{R}^2 \mid t \in \mathbb{R}\} \end{aligned}$$





Unlike for image processing, curves are of little importance in machine learning. On the contrary, hypersurfaces belong to the most important functions in machine learning. This is because, to a large extend, machine learning can be thought of being a branch of statistics. In statistics the most important class of functions are probability distributions, which, from a technical perspective, are functions mapping subsets of $\boldsymbol{R^n}$ into $\boldsymbol{R+}$.

Example 8

TODO:
scatter plot + distribution 3d (p22)