We'll assume that the reader is familiar with the concepts of sets, and maps between sets. Some special sets that we'll see a lot are:

- The set of natural numbers: $\mathbb{N} = \{1, 2, 3, \ldots\}$
- The set of integers: $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$
- The set of rational numbers: $\mathbb{Q} \approx \{\frac{a}{b} : a, b \in \mathbb{Z}, b \neq 0\}$ (Note that the set $\mathbb{Q}$ is really a set of *equivalence classes*, so for example, $\frac{1}{2}, \frac{2}{4}, \frac{3}{6}$, etc. all represent the same element of $\mathbb{Q}$.)
- The set of real numbers: $\mathbb{R}$
- The set of complex numbers: $\mathbb{C} = \{a + bi : a, b \in \mathbb{R}, i^2 = -1\}$

Note that $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$.

An *operator* on any given set (A) is a function $A \times A \to A$. Examples of operators are addition + and multiplication * on the number sets mentioned previously.

A set A along with an operator * on A is a *monoid* if it has the following properties:

1. Associativity: For all $a, b, c \in A$, $a*(b*c) = (a*b)*c$
2. Identity: There is some element ($e \in A$) such that for all $a \in A$, $a*e = e*a = a$

The element e is unique, since if we have e_1 and e_2 satisfying the identity property, then $e_1 = e_1*e_2 = e_2$ by that same identity property, and so they turn out to be the same element. The set $\mathbb{N}$ with the usual multiplication operator forms a *monoid*, with 1 playing the role of e.

A monoid is a *group* if it has an additional property:

3. Inverses: For every $a \in A$, there is an $a' \in A$ such that $a * a' = a' * a = e$

Inverses are unique, since if we have a_1' and a_2' as inverses for a, then $a_1' = a_1' * e = a_1' *(a * a_2') = (a_1' * a) * a_2' = e * a_2' = a_2'$, and they turn out to be the same element. Note that e is its own inverse, and depending on the group, there might be other elements that are their own inverse. These are called *involutions*.

The set $\mathbb{Z}$ with the usual addition operator forms a *group* with 0 playing the role of e. For an integer n, its inverse is $-n$, since $n + -n = -n + n = 0$.

A group is said to be *Abelian* if it also has the following property:

4. Commutativity: For every $a, b \in A$, $a * b = b * a$

In particular, $\mathbb{Z}$ with the usual addition operator is an Abelian group.

A set A with two operators, + and *, is a *ring* if it has the following properties:

1. $(A, +)$ is an Abelian group. The identity element of this group is denoted as 0.
2. $(A, *)$ is a monoid. The identity element of this monoid is denoted as 1.
3. The operator * *distributes* over +, that is, for every $a, b, c \in A$: $a*(b + c) = a*b + a*c$ and $(a + b)*c = a*c + b*c$

The set $\mathbb{Z}$ with the usual addition and multiplication is a ring. A ring is said to be *commutative* if it also has the following property:

4. For every $a, b \in A$, $a*b = b*a$

So $\mathbb{Z}$ is a *commutative ring*.

If $(A, +, *)$ is a commutative ring, and $(A\backslash\{0\}, *)$ is an Abelian group, then A is called a *field*. The sets $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ with the usual addition and multiplication are all *fields*, but $\mathbb{Z}$ is not because there are no multiplicative inverses in $\mathbb{Z}$.

A field $(F, +, \times)$, an Abelian group $(V, +)$, and an operator $*: F \times V \to V$ are called a vector space if they satisfy the following properties:

1. For all $v \in V$, $1*v = v$
2. For all $a, b \in F$, and all $v \in V$, $(a + b) * v = a*v + b*v$
3. For all $a, b \in F$, and all $v \in V$, $(a \times b) * v = a * (b * v)$
4. For all $a \in F$, and all $u, v \in V$, $a * (u + v) = a*u + a*v$

The elements of F are called *scalars*, and the elements of V are called *vectors.* Note that we can use + to denote addition in both F and V without ambiguity because we can never add vectors and scalars. The identity element of V is denoted as 0, which is called the *zero vector*. We say that *V is a vector space over F.*

From the properties above, it follows that for $a \in F$ and $v \in V$, $a*v = 0$ if and only if $a = 0$ or $v = 0$:

If $a=0$, then $a*v = 0*v = 0*v + 0 = 0*v + (0*v + -(0*v)) = (0*v + 0*v) + -(0*v) = (0 + 0)*v + -(0*v) = 0*v + -(0*v) = 0$, so $0*v = 0$ for all $v \in V$.

Similarly, if $v = 0$, then $a*v = a*0 = a*0 + 0 = a*0 + (a*0 + -(a*0)) = (a*0 + a*0) + -(a*0) = a*(0 + 0) + -(a*0) = a*0 + -(a*0) = 0$, so $a*0 = 0$ for all $a \in F$.

Finally, if $a*v = 0$ and $a \neq 0$, then let a' be the multiplicative inverse of a in F, and $v = 1*v = (a' \times a)*v = a' * (a * v) = a' * 0 = 0$ since $a *0 = 0$ for all $a \in F$ as previously shown. So, if $a*v = 0$ then either $a=0$ or $v=0$.

If V is a vector space over $\mathbb{R}$, an operator $\cdot : V \times V \rightarrow \mathbb{R}$ is an inner product on V if it has the following properties:

1. Positive Definiteness: $v \cdot v \geq 0$ for all $v \in V$ and $v \cdot v = 0$ if and only if $v=0$
2. Symmetry: $u \cdot v = v \cdot u$ for all $u, v \in V$
3. Linearity: $(a* (u+v)) \cdot w = a \times (u \cdot w + v \cdot w)$ for all $a \in F$ and all $u, v, w \in V$

For $d \in \mathbb{N}$, Let $\mathbb{R}^d = \{(x_1, x_2, ..., x_d) : x_1, x_2, ..., x_d \in \mathbb{R}\}$, so $0 = (0, 0, ..., 0) \in \mathbb{R}^d$.

For $u, v \in \mathbb{R}^d$, $u = (u_1, u_2, ..., u_d)$, $v = (v_1, v_2, ..., v_d)$, define:

$$u+v = (u_1+v_1, u_2+v_2, ..., u_d+v_d)$$

and for $a \in \mathbb{R}$, $u \in \mathbb{R}^d$ define:

$$a*u = a*(u_1, u_2, ..., u_d) = (a \times u_1, a \times u_2, ..., a \times u_d)$$

Then $\mathbb{R}^d$ is a vector space over $\mathbb{R}$.

Define the operator $\cdot : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ as $u \cdot v = u_1 \times v_1 + u_2 \times v_2 + ... + u_d \times v_d$. This is an inner product on $\mathbb{R}^d$, also called the *dot product*.

Most machine learning discussions involve vector spaces $\mathbb{R}^d$ over $\mathbb{R}$ as defined here.

For a vector space V over a field F, a set of non-zero vectors $\{v_1, v_2, ..., v_n\} \subset V$ is said to be *linearly independent* if for arbitrary $a_1, a_2, ..., a_n \in F$:

$a_1 * v_1 + a_2 * v_2 + ... + a_n * v_n = 0$ if and only if $a_1 = a_2 = ... = a_n = 0$

The *dimension* of a vector space is the size of the largest linearly independent set of vectors that can be found in it. For example, the dimension of $\mathbb{R}^d$ is d.

If we can find n linearly independent vectors in V for every $n \in \mathbb{N}$, then V is said to be of infinite dimension or infinite-dimensional. Examples of infinite-dimensional vector spaces crop up in machine learning with support vector machines.

If V is a vector space of dimension d over a field F and $\{v_1, v_2, ..., v_d\} \subset V$ is linearly independent, we call that set a *basis* for V, and the vectors in the set are called *basis vectors*. Any $u \in V$ can be written as a linear combination of basis vectors, which means we can find $a_1, a_2, ..., a_d \in F$ such that:

$$u = a_1 * v_1 + a_2 * v_2 + ... + a_d * v_d$$

A basis is not unique. Any set of d linearly independent vectors will do. For example, in $\mathbb{R}^d$, we normally use the following set as a basis:

$$\{e_1 = (1, 0, 0, ..., 0), e_2 = (0, 1, 0, ..., 0), e_3 = (0, 0, 1, ..., 0), ..., e_d = (0, 0, 0, ..., 1)\}$$

We denote this as the *standard basis* of $\mathbb{R}^d$. It then becomes clear that every vector $v = (v_1, v_2, ..., v_d)$ in $\mathbb{R}^d$ can be written in the following way:

$$v = v_1 * e_1 + v_2 * e_2 + ... + v_d * e_d$$

A *real-valued matrix* is an array of real numbers, arranged in rows and columns. An $n \times m$ matrix A has n rows and m columns, and it is written as:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

Where each $a_{ij} \in \mathbb{R}$. We denote the set of all real-valued $n \times m$ matrices as $\mathbb{R}^{n \times m}$
We can add two $n \times m$ matrices by adding them component-wise. So, if

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}$$

and

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{bmatrix}$$

then,

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1m} + b_{1m} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2m} + b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \cdots & a_{nm} + b_{nm} \end{bmatrix}$$

Note that $\mathbb{R}^{n \times m}$ with this addition operator forms an Abelian group, with the identity element being the matrix *O*ϵ $\mathbb{R}^{n \times m}$ – the matrix where all entries are zero. In fact, $\mathbb{R}^{n \times m}$ is an (*n*×*m)*-dimensional vector space over $\mathbb{R}$, and for *x* ∈ $\mathbb{R}$ and *A* ∈ $\mathbb{R}^{n \times m}$, we define:

$$x * A = \begin{bmatrix} x \times a_{11} & x \times a_{12} & \cdots & x \times a_{1m} \\ x \times a_{21} & x \times a_{22} & \dots & x \times a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x \times a_{n1} & x \times a_{n2} & \dots & x \times a_{nm} \end{bmatrix}$$

Note that an *n*×*m* matrix *A* has *n* rows, each of which can be viewed as a vector in $\mathbb{R}^m$. These are the *row vectors* of *A*. The matrix also has *m* columns, each of which can be viewed as a vector in $\mathbb{R}^n$. These are called the *column vectors* of *A*.

Since the row vectors of an *n*×*m* matrix *A* are in $\mathbb{R}^m$, we can define an operator $\mathbb{R}^{n \times m} \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ as follows: for *A* ∈ $\mathbb{R}^{n \times m}$ let (a_1, a_2, ..., a_n) denote the row vectors of *A*. Then for any c define:

*A*x = (a_1·x, a_2·x, ..., a_n·x)

a·x then denotes the dot product in $\mathbb{R}^m$

Similarly, since the column vectors in A are in $\mathbb{R}^n$, we can define an operator $\mathbb{R}^n \times \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^m$ as follows: for $A \in \mathbb{R}^{n\times m}$ let $(a_1, a_2, \ldots, a_m)$ denote the column vectors of A. Then for any $x \in \mathbb{R}^n$ define:

$$xA = (x\cdot a_1, x\cdot a_2, \ldots, x\cdot a_m)$$

$x\cdot a$ then denotes the dot product in $\mathbb{R}^n$

This can be extended to define an operator $\mathbb{R}^{n\times d} \times \mathbb{R}^{d\times m} \rightarrow \mathbb{R}^{n\times m}$ as follows: for $A \in \mathbb{R}^{n\times d}$, let $(a_1, a_2, \ldots, a_n)$ denote the row vectors of A. Note that these are vectors in $\mathbb{R}^d$. For $B \in \mathbb{R}^{d\times m}$, let $(b_1, b_2, \ldots, b_m)$ denote the column vectors of B. Note that these are also vectors in $\mathbb{R}^d$. So, we can use the dot product in $\mathbb{R}^d$ to define:

$$AB = \begin{bmatrix} a_1 \cdot b_1 & a_1 \cdot b_2 & \ldots & a_1 \cdot b_m \\ a_2 \cdot b_1 & a_2 \cdot b_2 & \cdots & a_2 \cdot b_m \\ \vdots & \vdots & \ddots & \vdots \\ a_n \cdot b_1 & a_n \cdot b_2 & \cdots & a_n \cdot b_m \end{bmatrix}$$

In the special case where $n = d = m$, this defines a multiplication operator on $\mathbb{R}^{n\times n}$, and in fact $\mathbb{R}^{n\times n}$ with this multiplication operator is a monoid, with the identity element being the matrix:

$$I = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

In this matrix, diagonal entries are 1 and all other entries are zero. Combining this with the matrix addition operator defined previously, $\mathbb{R}^{n\times n}$ is a ring. Note that in general, $AB \neq BA$, so this is an example of a non-commutative ring.

The transpose of an $n\times m$ matrix A, written as A^T, is the $m\times n$ matrix whose row vectors are the column vectors of A, and whose column vectors are the row vectors of A. So,

If A is the matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix},$$

then its transpose is:

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}$$

Furthermore, A^TA will be an $m \times m$ matrix, and AA^T will be an $n \times n$ matrix. A matrix $A \in \mathbb{R}^{n\times n}$ is *symmetric* if $A = A^T$, which means that its column vectors and row vectors are the same.

If $A, B \in \mathbb{R}^{n\times n}$, then:

- $(A^T)^T = A$
- $(A + B)^T = A^T + B^T$
- $(AB)^T = B^TA^T$

Previously, we defined an operator $\mathbb{R}^{n\times m} \times \mathbb{R}^m \to \mathbb{R}^n$ given by Ax = (a_1·x, a_2·x, ..., a_n·x) where the a_i are the row vectors of $A \in \mathbb{R}^{n\times m}$ and x $\in \mathbb{R}^m$. This means that if we are given a specific $A \in \mathbb{R}^{n\times m}$, then A can be viewed as a map A: $\mathbb{R}^m \to \mathbb{R}^n$ so that for x $\in \mathbb{R}^m$, A(x) = Ax.

Let F be a field, and let U, V be vector spaces over F. A function L:$U \to V$ is a linear map, also called a linear transformation, if for all x, y $\in U$ and all a, $b \in F$:

$$L(a*x + b*y) = a*L(x) + b*L(y)$$

Let 0_U denote the zero vector in U and 0_V denote the zero vector in V. Then:

$$L(0_U) = L(0*0_U) = 0*L(0_U) = 0_V$$

This is true since $L(0_U) \in V$ by definition of L, and 0*v = 0_V for all v $\in V$ as shown previously.

If $A \in \mathbb{R}^{n\times m}$, then the map A: $\mathbb{R}^m \to \mathbb{R}^n$ discussed previously is a linear map. Conversely, given a linear map L: $\mathbb{R}^m \to \mathbb{R}^n$, there will a matrix $A \in \mathbb{R}^{n\times m}$ such that L(x) = Ax. The proof of this is a bit lengthy, but if you want to try to prove this yourself, look at how L acts on the standard basis of $\mathbb{R}^m$, and show that the i^{th} component of $L(e_j)$ is the element a_{ij} of the matrix A you're looking for.

This means that there is a one-to-one correspondence between the matrices in $\mathbb{R}^{n\times m}$ and the set of linear maps from $\mathbb{R}^m$ to $\mathbb{R}^n$. In the special case where $n=m$, the matrices in $\mathbb{R}^{n\times n}$ correspond one-to-one with the linear maps from $\mathbb{R}^n$ to itself.

To compose linear maps on $\mathbb{R}^n$, we simply multiply the corresponding matrices: if $L_1: \mathbb{R}^n \to \mathbb{R}^n$ with $L_1(\mathrm{x}) = A\mathrm{x}$, and $L_2: \mathbb{R}^n \to \mathbb{R}^n$ with $L_2(\mathrm{x}) = B\mathrm{x}$, then $L_1(L_2(\mathrm{x})) = L_1(B\mathrm{x}) = A(B\mathrm{x}) = (AB)\mathrm{x}$.

Some of these maps will have an *inverse*, which means we are able to find an $L^{-1}: \mathbb{R}^n \to \mathbb{R}^n$ such that $L^{-1}(L(\mathrm{x})) = L(L^{-1}(\mathrm{x})) = \mathrm{x}$. Also, the identity matrix I satisfies $\mathrm{Ix}=\mathrm{x}$, so if A is the matrix corresponding to L, and A' is the matrix corresponding to L^{-1}, then $AA' = A'A = \mathrm{I}$. We call A' the *inverse* of A and write it as A^{-1}. We say that A is *invertible* if A^{-1} exists. The set of all invertible matrices in $\mathbb{R}^{n\times n}$ form a group under matrix multiplication with the identity element of the group being the matrix I. Note that this group is not Abelian, since in general $AB \neq BA$ even when the matrices are invertible.

For $A \in \mathbb{R}^{n\times n}$, define the trace of A, written tr$A$, as:

$$\mathrm{tr}A = a_{11} + a_{22} + \ldots + a_{nn}$$

The trace of a square matrix is the sum of its diagonal elements. The trace has the following properties for $A, B \in \mathbb{R}^{n\times n}$:

- $\mathrm{tr}A = \mathrm{tr}A^T$
- $\mathrm{tr}\,(A+B) = \mathrm{tr}A + \mathrm{tr}B$
- $\mathrm{tr}AB = \mathrm{tr}BA$

More generally, for $A_1, A_2, \ldots, A_{k-1}, A_k \in \mathbb{R}^{n\times n}$:

- $\mathrm{tr}A_1A_2\ldots A_{k-1}A_k = \mathrm{tr}A_kA_1A_2\ldots A_{k-1} = \mathrm{tr}A_2\ldots A_{k-1}A_k\,A_1$

Many other properties can be derived from these, for example:

- $\mathrm{tr}AB = \mathrm{tr}(AB)^T = \mathrm{tr}B^TA^T = \mathrm{tr}A^TB^T$
- $\mathrm{tr}B^{-1}AB = \mathrm{tr}B^{-1}(AB) = \mathrm{tr}(AB)B^{-1} = \mathrm{tr}A(B\,B^{-1}) = \mathrm{tr}A\mathrm{I} = \mathrm{tr}A$

And so on. Traces are also important within vector calculus.

Consider the standard basis (e_1, e_2, ..., e_n) in $\mathbb{R}^n$. Viewed geometrically, these span a unit *n*-cube, so for *n*=2, (e_1, e_2) span a unit square; for *n*=3, (e_1, e_2, e_3) span a unit cube, etc.

For $A \in \mathbb{R}^{n \times n}$, let ($a_1$, a_2, ..., a_n) denote the column vectors of *A*, so $Ae_k = a_k$ for $1 \le k \le n$. That means *A* maps the unit *n*-cube to some *parallelotope* spanned by (a_1, a_2, ..., a_n). Thus, for *n*=2, (a_1, a_2) span a parallelogram, for *n=3*, (a_1, a_2, a_3) span a *parallelepiped*, etc.

The determinant of *A*, written det*A* or $|A|$, is the *signed volume* of the *parallelotope* spanned by the column vectors of *A*. If $|A|>0$, then *A* preserves the orientation of vectors, and if $|A|<0$, then *A* reverses the orientation of vectors.

If $|A| = 0$, then the region has no *n*-dimensional volume, and so the region has fewer than *n* dimensions. This means that the linear transformation cannot be inverted, and so A^{-1} does not exist, *i.e.*, *A* is not an invertible matrix. In this case, we say that *A* is *singular*. If *A* is invertible, it is *non-singular*.

Determinants have the following properties. For $A, B \in \mathbb{R}^{n \times n}$:

- *A* is non-singular if and only if $|A| \neq 0$
- $|I| = 1$ (since it spans a unit *n*-cube)
- If any row or column vector of *A* is the zero vector, then $|A|=0$
- If the row vectors of *A* are not linearly independent, then $|A|=0$
- If the column vectors of *A* are not linearly independent, then $|A|=0$
- $|A| = |A^T|$
- $|AB| = |A|\ |B|$
- If $|A| \neq 0$, then $|A^{-1}| = |A|^{-1}$

Given $A \in \mathbb{R}^{n \times n}$, define $M_{ij} \in \mathbb{R}^{(n-1)\times(n-1)}$ to be the matrix obtained by removing the i^{th} row vector and j^{th} column vector from *A*. The determinants $|M_{ij}|$ are called the *minors* of *A*. Let $C_{ij} = (-1)^{i+j}\,|M_{ij}|$; these are called the *cofactors* of A. We use these to compute $|A|$:

- If $A \in \mathbb{R}^{1 \times 1}$, then $A = [a_{11}]$ and $|A| = a_{11} \in \mathbb{R}$
- Otherwise, pick any row vector $a_i = (a_{i1}, a_{i2}, ..., a_{in})$ in *A*, and compute:

$$|A| = \sum_{j=1}^{n} a_{ij}\, C_{ij}$$

To compute each cofactor C_{ij}, we must compute the determinant $|M_{ij}|$, which we do recursively. Note that each M_{ij} is of a lower dimension than the previous one, so we'll eventually hit the $A \in \mathbb{R}^{1\times1}$ case.
1. What have you accomplished since your last status update?

2. What are you working on today?

3. Are there any obstacles impeding your progress?

4. What's something you're grateful for today?
Once we've computed $|A|$, and we find that $|A| \neq 0$, we can use it to compute A^{-1}. Given $A \in \mathbb{R}^{n\times n}$, define:

$$C = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \dots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \dots & C_{nn} \end{bmatrix},$$

Where the C_{ij} are the cofactors of A as previously defined, then:

$$A^{-1} = \frac{1}{|A|} C^T$$

There are more efficient ways to compute determinants and inverses. For any application where computing determinants or inverses of matrices is required, it is easiest to use existing "off-the-shelf" linear algebra packages rather than writing code from scratch.

How to Build a Supervised Learning Algorithm

We discussed different types of learning algorithms in a previous article. With a supervised learning algorithm, the example data set provides an input and output value for each data point:

$$D = \{(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \ldots, (\mathrm{x}_N, y_N)\}$$

In the hypothesis set (H) for this learning problem, we'll use *linear models*. We will pick $\mathrm{w} = (w_0, w_1, \ldots, w_d)$ and define:

$$h_{\mathrm{w}}(\mathrm{x}) = w_0 + w_1x_1 + w_2x_2 + \ldots + w_dx_d$$

This is a *linear combination* of the data points (x_i) that comprise x, hence the name *linear models.* Our set H is the set of all such functions. By convention, we'll write each $\mathrm{x} = (x_1, x_2, \ldots, x_d)$ as $(1, x_1, x_2, \ldots, x_d)$, in other words, we'll insert an $x_0 = 1$ in the first component. This allows us to write h_{w} as an inner product:

$$h_{\mathrm{w}}(\mathrm{x}) = \mathrm{w} \cdot \mathrm{x}$$

We also need a way to measure how accurate h_{w} is. Since we have a y_n for each x_n, one way to measure our accuracy is to compute the difference between $h_{\mathrm{w}}(\mathrm{x}_n)$ and y_n for each point within our known data set. We call this an *error function* because it measures the error in h_{w} on D. We can denote this function as E_{w} and define the function as:

$$E_{\mathrm{w}} = \sum_{n-1}^{N}(\mathrm{w} \cdot \mathrm{x}_n - y_n)$$

However, it's more convenient to define E_{w} in terms of h_{w} as follows:

$$E_{\mathrm{w}} = \frac{1}{2}\sum_{n-1}^{N}(\mathrm{w} \cdot \mathrm{x}_n - y_n)^2 = \frac{1}{2}\sum_{n-1}^{N}(h_{\mathrm{w}}(\mathrm{x}_n) - y_n)^2$$

This will help simplify later calculations.

Now that we've defined our hypothesis set H, the task of our learning algorithm will be to find an h_{w} that minimizes the value of E_{w}. Note that E_{w} is a function of several variables, and from how we've defined it, it's differentiable everywhere. This allows us to find a minimum value for it by computing its *gradient*, ∇E_{w}, and solving $\nabla E_{\mathrm{w}} = 0$.

We can do this analytically using some linear algebra. Define an $N \times (d+1)$ matrix with X to be the matrix whose rows are the x values from our data set, so:

$$X = \begin{pmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \dots \\ \mathrm{x}_N \end{pmatrix}$$

Where each x_n is $(1\ x_{n,1}\ x_{n,2}\ \dots\ x_{n,d})$, then for $\mathrm{w} = (w_0, w_1, \dots, w_d)$,

$$X\mathrm{w} = \begin{pmatrix} \mathrm{w}\cdot\mathrm{x}_1 \\ \mathrm{w}\cdot\mathrm{x}_2 \\ \dots \\ \mathrm{w}\cdot\mathrm{x}_N \end{pmatrix} = \begin{pmatrix} \mathrm{h_w}(\mathrm{x}_1) \\ \mathrm{h_w}(\mathrm{x}_2) \\ \dots \\ \mathrm{h_w}(\mathrm{x}_N) \end{pmatrix}$$

if we also write our output values as $\mathrm{y} = (y_1, y_2, \dots, y_N)$, then:

$$X\mathrm{w} - \mathrm{y} = \begin{pmatrix} \mathrm{h_w}(\mathrm{x}_1) - y_1 \\ \mathrm{h_w}(\mathrm{x}_2) - y_2 \\ \dots \\ \mathrm{h_w}(\mathrm{x}_N) - y_N \end{pmatrix}$$

Note that this is a vector, and we can take the inner product of this vector with itself:

$$(X\mathrm{w} - \mathrm{y}) \cdot (X\mathrm{w} - \mathrm{y}) = \sum_{n-1}^{N} (h_\mathrm{w}(\mathrm{x}_n) - y_n)^2$$

And we almost have our error function from before – we just need to divide by 2:

$$E_\mathrm{w} = \frac{1}{2}(X\mathrm{w} - \mathrm{y}) \cdot (X\mathrm{w} - \mathrm{y}) = \frac{1}{2}\sum_{n-1}^{N} (h_\mathrm{w}(\mathrm{x}_n) - y_n)^2$$

We'll omit the lengthy and tedious calculation of ∇E_w, and go straight to the punch line:

$$\nabla E_\mathrm{w} = X^T X\mathrm{w} - X^T\mathrm{y}$$

Setting this to zero, we solve for w:

$$X^T X\mathrm{w} = X^T\mathrm{y}$$

And we find that:

$$w = (X^T X)^{-1} X^T y$$

As long as the matrix $X^T X$ has a non-zero determinant, we will have an exact value for w, and our final hypothesis will be the function g(x) = w·x with w computed as above. Note that g is entirely dependent on the data in our training set. Also, computing w could be an issue if we have a very large training set.