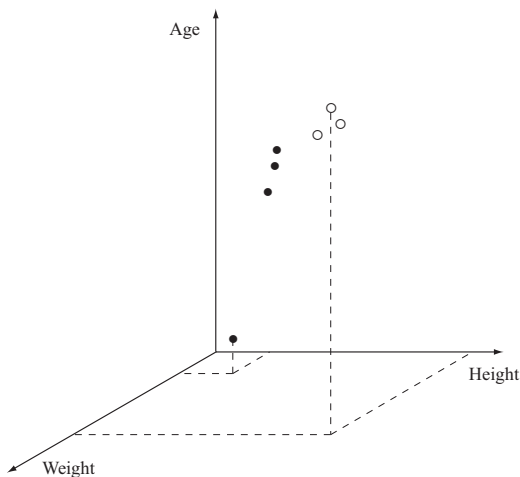# Representing Data

Data with n independent real-valued attributes can be represented in n-dimensional real space.

	Height	Weight	Age	Gender
Jane	25.4	32.7	2.5	F
Amanda	65.2	132.0	36.5	F
Paul	71.7	175.1	25.5	M
Mary	62.6	126.0	31.0	F
Gary	68.2	182.0	42.5	M
Betty	58.5	118.5	21.0	F
John	72.0	195.2	45.2	M

Linear Algebra

Linear algebra allows us to describe structures in n-dimensional spaces very effectively.

The representation of the models in support vector machines draw heavily on concepts such as vector spaces, planes, and norms from linear algebra.

We start with the most basic definition:

Definition: *A directed line segment is called a **vector**. A vector has both a length and a direction. The length is sometimes called the Euclidean norm or magnitude. A vector of length 1 is called a **unit vector**.*

A special case of vector is the position vector:

Definition: *A **position vector** is a vector whose initial point is the origin of some coordinate system and whose terminal point is described by a set of coordinates in the coordinate system.*

Position Vectors

Our data set revisited:

	Height	Weight	Age	Gender
Jane	25.4	32.7	2.5	F
Amanda	65.2	132.0	36.5	F
Paul	71.7	175.1	25.5	M
Mary	62.6	126.0	31.0	F
Gary	68.2	182.0	42.5	M
Betty	58.5	118.5	21.0	F
John	72.0	195.2	45.2	M
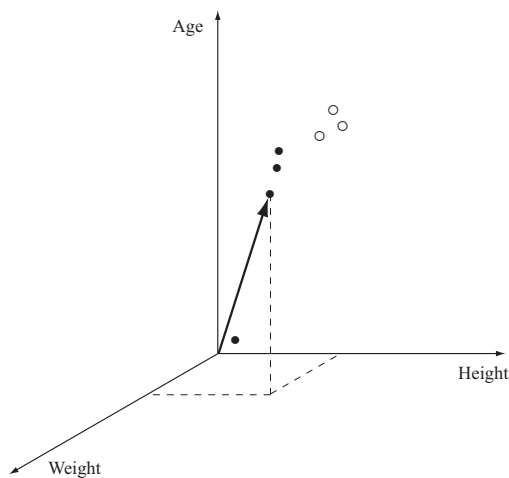
Example: The position vector for Betty starts at $(0, 0, 0)$ and ends at $(58.5, 118.5, 21.0)$

Vector Operations

Vector operations such as *equality*, *addition*, and *scalar multiplication* are *componentwise*.

Example: Let $\overline{a} = (a_1, a_2, a_3)$ and $\overline{b} = (b_1, b_2, b_3)$ be position vectors, then

$$\overline{a} = \overline{b} \text{ iff } a_1 = b_1, a_2 = b_2, a_3 = b_3.$$

This gives rise to identities such as

$$
\begin{aligned}
\overline{a} + \overline{b} &= \overline{b} + \overline{a} && \text{(commutativity)} \\
(\overline{a} + \overline{b}) + \overline{c} &= \overline{a} + (\overline{b} + \overline{c}) && \text{(associativity)} \\
\overline{a} + \overline{0} &= \overline{0} + \overline{a} = \overline{a} && \text{(identity)} \\
\overline{a} + (-\overline{a}) &= \overline{0} && \text{(reciprocal)}
\end{aligned}
$$

Where $\overline{a}, \overline{b}$, and $\overline{c}$ are position vectors. Furthermore, $\overline{0} = (0, 0, \ldots, 0) = 0^n$ is the ***null vector*** and $(-\overline{a})$ is the vector $(-a_1, -a_2, \ldots, -a_n) = -1 \times \overline{a}$.

More Identities

$$
\begin{aligned}
q(\overline{a} + \overline{b}) &= q\overline{a} + q\overline{b} && \text{(distributivity)} \\
(p + q)\overline{a} &= p\overline{a} + q\overline{a} && \text{(distributivity)} \\
p(q\overline{a}) &= (pq)\overline{a} && \text{(associativity)} \\
1\overline{a} &= \overline{a} && \text{(identity)} \\
0\overline{a} &= \overline{0} && \\
q\overline{0} &= \overline{0} && \\
(-1)\overline{a} &= -\overline{a} &&
\end{aligned}
$$

Vector Spaces

Informally we say that a vector space is a collection of vector sthat can be added and scaled. More formally,

Definition: *A non-empty set V of vectors in $\mathbb{R}^n$ is called a **(real)** **vector space** if vector addition and scalar multiplication are defined and closed over this set and satisfy the following axioms for all $\overline{a}, \overline{b}, \overline{c} \in V$ and $p, q \in \mathbb{R}$.*
Addition:

$$
\begin{aligned}
\overline{a} + \overline{b} &= \overline{b} + \overline{a} \\
(\overline{a} + \overline{b}) + \overline{c} &= \overline{a} + (\overline{b} + \overline{c}) \\
\overline{a} + \overline{0} &= \overline{a} \\
\overline{a} + (-\overline{a}) &= \overline{0}
\end{aligned}
$$

Multiplication:

$$
\begin{aligned}
q(\overline{a} + \overline{b}) &= q\overline{a} + q\overline{b} \\
(p + q)\overline{a} &= p\overline{a} + q\overline{a} \\
p(q\overline{a}) &= (pq)\overline{a} \\
1\overline{a} &= \overline{a}
\end{aligned}
$$

Linear Combinations

We can use addition and multiplication in a vector space to construct new vectors from given ones.

Example: Let $\overline{a}_1, \overline{a}_2, \ldots, \overline{a}_m$ be vectors in some vector space V, then an expression of the form

$$\sum_{i=1}^{m} q_i \overline{a}_i = q_1 \overline{a}_1 + \ldots + q_m \overline{a}_m,$$

where $q_i \in \mathbb{R}$, is called a *linear combination* and the closure property of vector spaces guarantees that $\sum_{i=1}^{m} q_i \overline{a}_i \in V$.

Linear Combinations

An important example of linear combinations is the formal notion of dimensionality of a data set.

Example: Let the unit vectors $\bar{\imath} = (1, 0, 0)$, $\bar{\jmath} = (0, 1, 0)$, and $\bar{k} = (0, 0, 1)$ be *linearly independent* vectors, then we can view them as the *canonical basis* of $\mathbb{R}^3$, that is, any vector in $\mathbb{R}^3$ can be represented as a linear combination of these three vectors.

Example: Consider the position vector for Amanda. We can rewrite it as,

$$(65.2, 132.0, 36.5) = 65.2(1, 0, 0) + 132.0(0, 1, 0) + 36.5(0, 0, 1).$$

Observation: The real space $\mathbb{R}^n$ can always be considered a vector space.

The Dot Product

Definition: *Given two vectors $\overline{a} = (a_1, \ldots, a_n)$ and $\overline{b} = (b_1, \ldots, b_n)$ in an n-dimensional vector space V, then we define the **dot product** as the operation*

$$\overline{a} \bullet \overline{b} = a_1 b_1 + \ldots + a_n b_n \tag{1}$$

or

$$\overline{a} \bullet \overline{b} = |\overline{a}||\overline{b}| \cos \gamma, \tag{2}$$

where $|\overline{a}| = \sqrt{a_1^2 + \ldots + a_n^2}$ is the length of vector $\overline{a}$ and γ is the angle between the two position vectors.

The following identities hold for dot products. Let $\overline{a}, \overline{b}, \overline{c} \in V$ and $p, q \in \mathbb{R}$, then

$$
\begin{aligned}
(p\overline{a} + q\overline{b}) \bullet \overline{c} &= p\overline{a} \bullet \overline{c} + q\overline{b} \bullet \overline{c} && \text{(linearity)} \\
\overline{a} \bullet \overline{b} &= \overline{b} \bullet \overline{a} && \text{(symmetry)} \\
\overline{a} \bullet \overline{a} &\geq 0, \text{ and} \\
\overline{a} \bullet \overline{a} &= 0 \text{ iff } \overline{a} = \overline{0} && \text{(positive-definiteness)}
\end{aligned}
$$

The Dot Product

We also have the following interesting identities that are not as fundamental as the previous set but very useful:

$$|\overline{a}| = \sqrt{\overline{a} \bullet \overline{a}},$$

$$\cos\gamma = \frac{\overline{a} \bullet \overline{a}}{|\overline{a}||\overline{b}|}.$$

So what does this mean? What is the intuition behind the dot product?

> The dot product is a ***measure of similarity***.

Consider the second equation above with $k = |\overline{a}||\overline{b}|$, then $\overline{a} \bullet \overline{b} = k\cos\gamma$ and we see that the value of the dot product is proportional to the cosine of the angle between them: $0° \mapsto k$, $90° \mapsto 0$, $180° \mapsto -k$, *etc.*
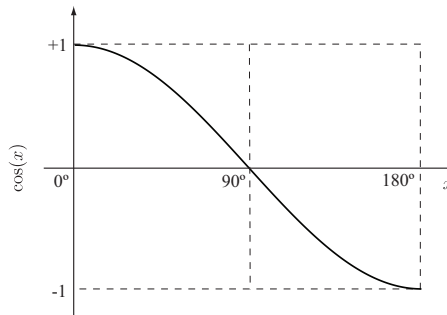
Dot Product Spaces

Definition: *A **dot product space** is a vector space where dot products are defined.*

That is, given any two vectors in the vector space we can measure their similarity.

Definition: *Two non-zero length vectors are **orthogonal** if and only if their dot product is zero (maximum dissimilarity).*

Observation: The real space $\mathbb{R}^n$ can always be considered a dot product space.

Lines & Dot Products

We can represent functions using dot products, all of the following identities describe the same set of points:

$$
\begin{aligned}
f(x) &= -mx \\
y &= -mx \\
mx + y &= 0 \\
w_1 x + w_2 y &= 0, \text{ where } w_1 = m, w_2 = 1 \\
\overline{w} \bullet \overline{x} &= 0, \text{ where } \overline{w} = (w_1, w_2) \text{ and } \overline{x} = (x, y)
\end{aligned}
$$

Note: an advantage of the dot product notation of the last identity is that dimensionality is implicit rather than explicit as in the other identities allowing us to describe planes and hyperplanes with very compact notation.

Note: the vectors $\overline{w}$ and $\overline{x}$ are orthogonal, and since $\overline{x}$ describes a line (hyperplane) we call $\overline{w}$ the *normal vector* of the line (hyperplane).

Lines & Dot Products

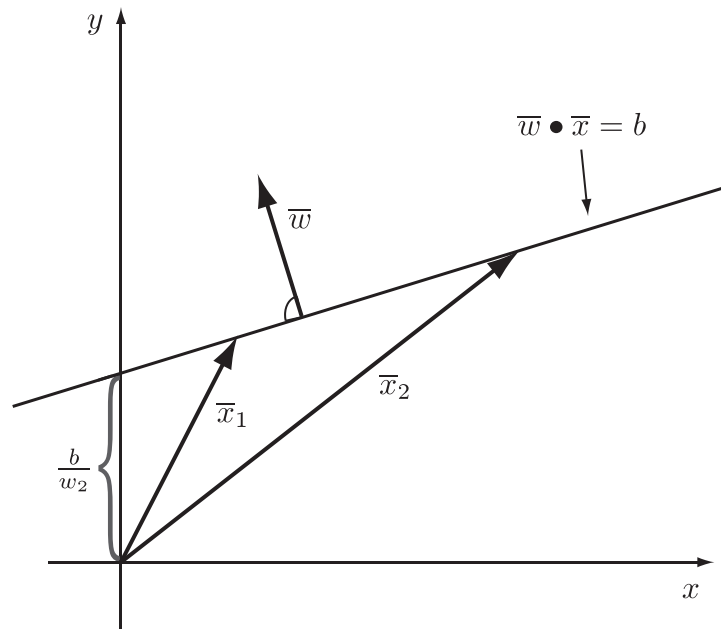$$\overline{w} \bullet \overline{x} = mx + y = 0$$
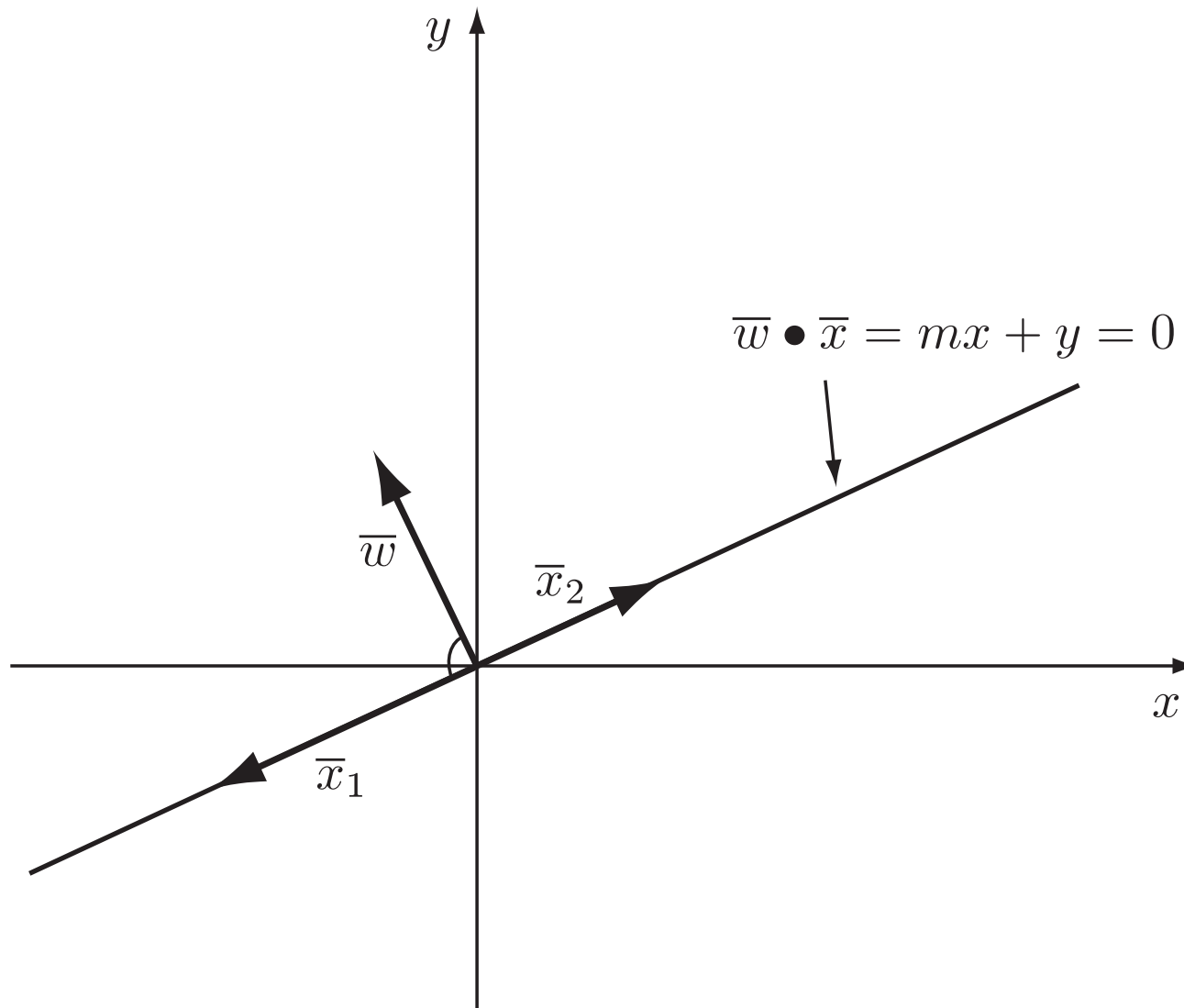
Lines & Dot Products

We can generalize the above expression a little bit by admitting lines that do not have to go through the origin of the coordinate system:

$$\overline{w} \bullet \overline{x} = b$$

where $\overline{w} = (w_1, w_2)$ and $\overline{x} = (x, y)$.

The constant $\frac{b}{w_2}$ is called the the *y-intercept*.

Hyperplanes

We can generalize this even further by admitting arbitrary dimensions $n > 2$ such that $\overline{w} = (w_1, \ldots, w_n)$ and $\overline{x} = (x_1, \ldots, x_n)$.

The dot product notation itself does not change

$$\boxed{\overline{w} \bullet \overline{x} = b}$$

This equation defines a ***hyperplane*** in n-dimensional space.

Hyperplanes

It is difficult to draw hyperplanes, for $n = 3$ hyperplanes degenerate into the well known 3-dimensional plane.

Example: Let $\overline{w} = (w_1, w_2, w_3)$ and $\overline{x} = (x, y, z)$, then