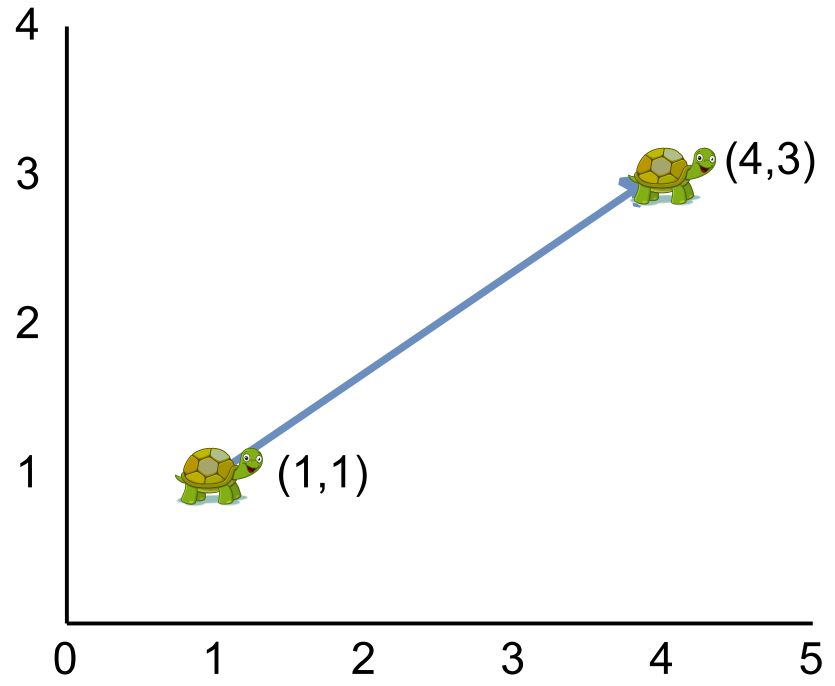
Vectors and logistic regression

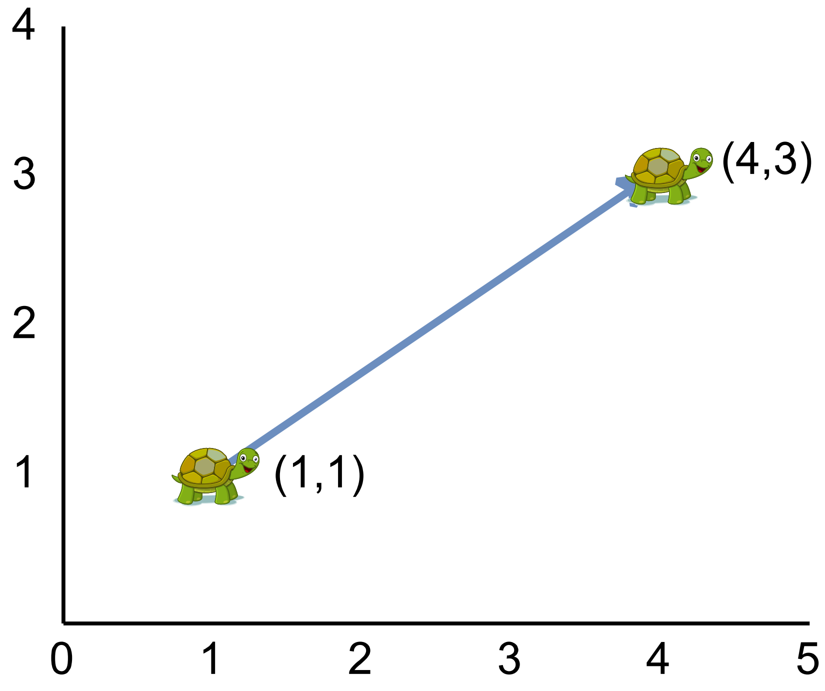
Daniël de Kok

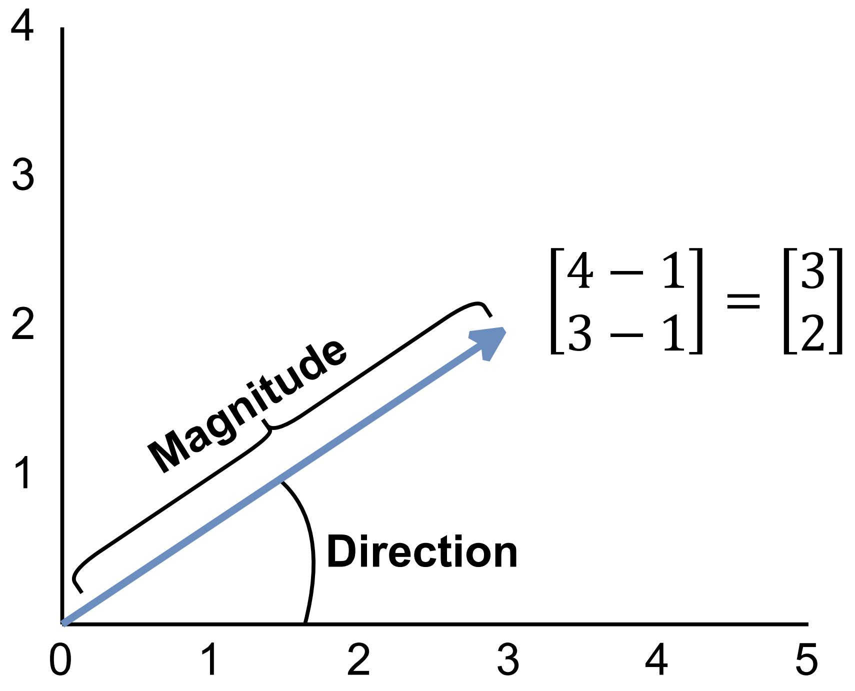
What is a vector?



Goal: describe the *length* and the direction of the movement of the turtle irrespective of its absolute positions.

What is a vector?





Goal: describe the *length* and the direction of the movement of the turtle irrespective of its absolute positions.

This lecture

- Elementary operators
- How do we find the length of a vector?
- How do we find the angle between two vectors?
- Logistic regression

Elementary operators

Notation

- Scalars are named using lowercase letters:

$$a, b, c$$

- Vectors are named using lowercase letters in boldface:

$$\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$$

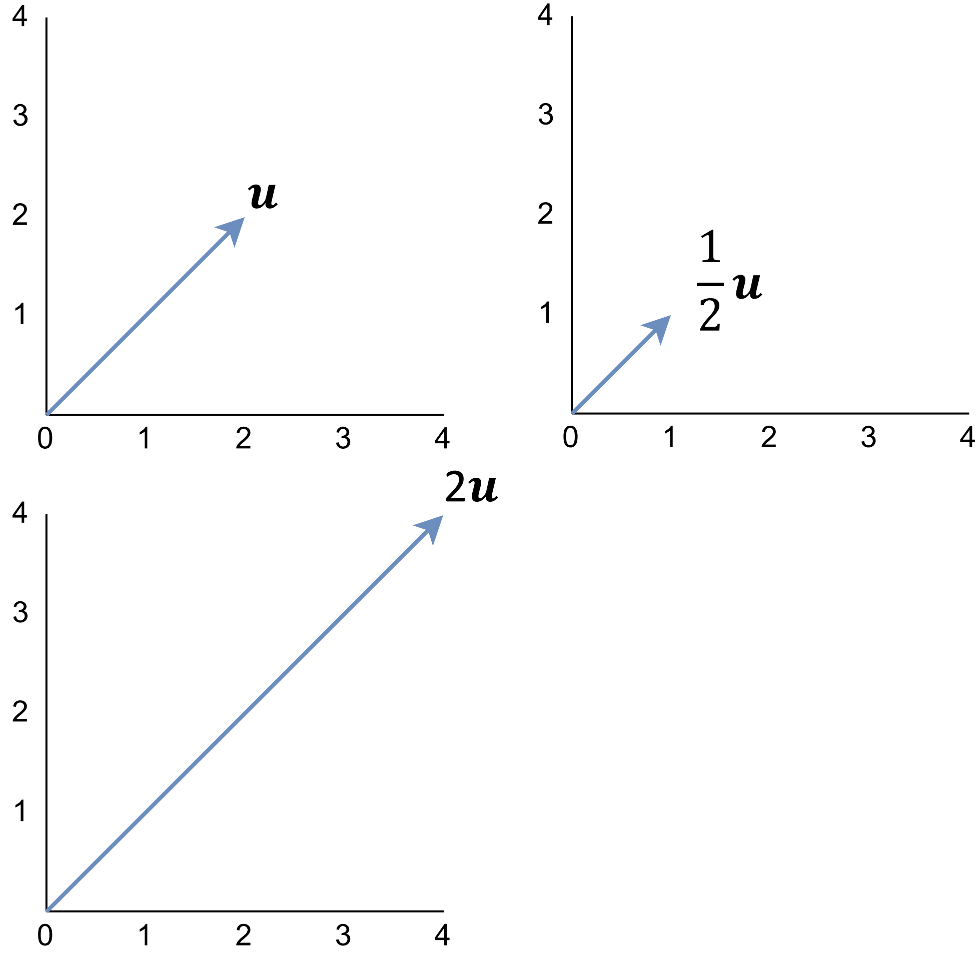
- Vectors are indexed using subscript:

$$u_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}_1 = 3$$

- We denote a vector $\boldsymbol{v}$ in a d-dimensional vector space of real numbers as:

$$\boldsymbol{v} \in \mathbb{R}^d$$

Vector scaling

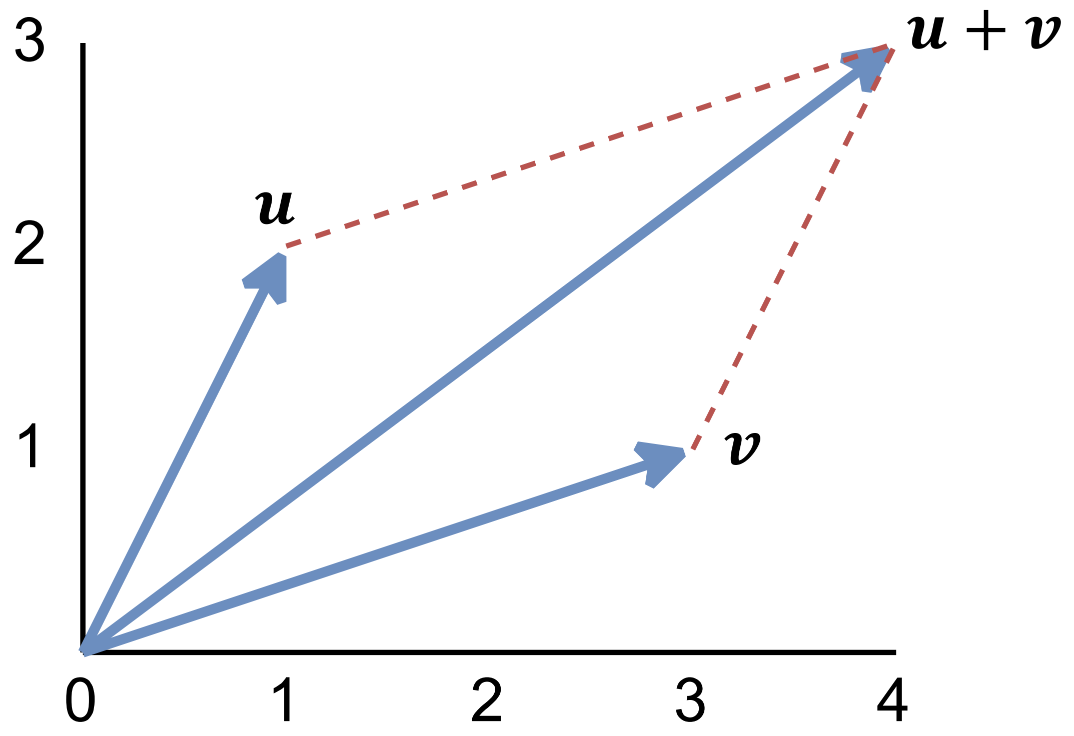


Scaling: each vector $\boldsymbol{u}$ can be scaled with a scalar a,

$$a\boldsymbol{u} = \begin{bmatrix} au_1 \\ \vdots \\ au_n \end{bmatrix}$$

Scaling changes the *length* of the vector, never the *direction*, except when $a = 0$.

Vector addition



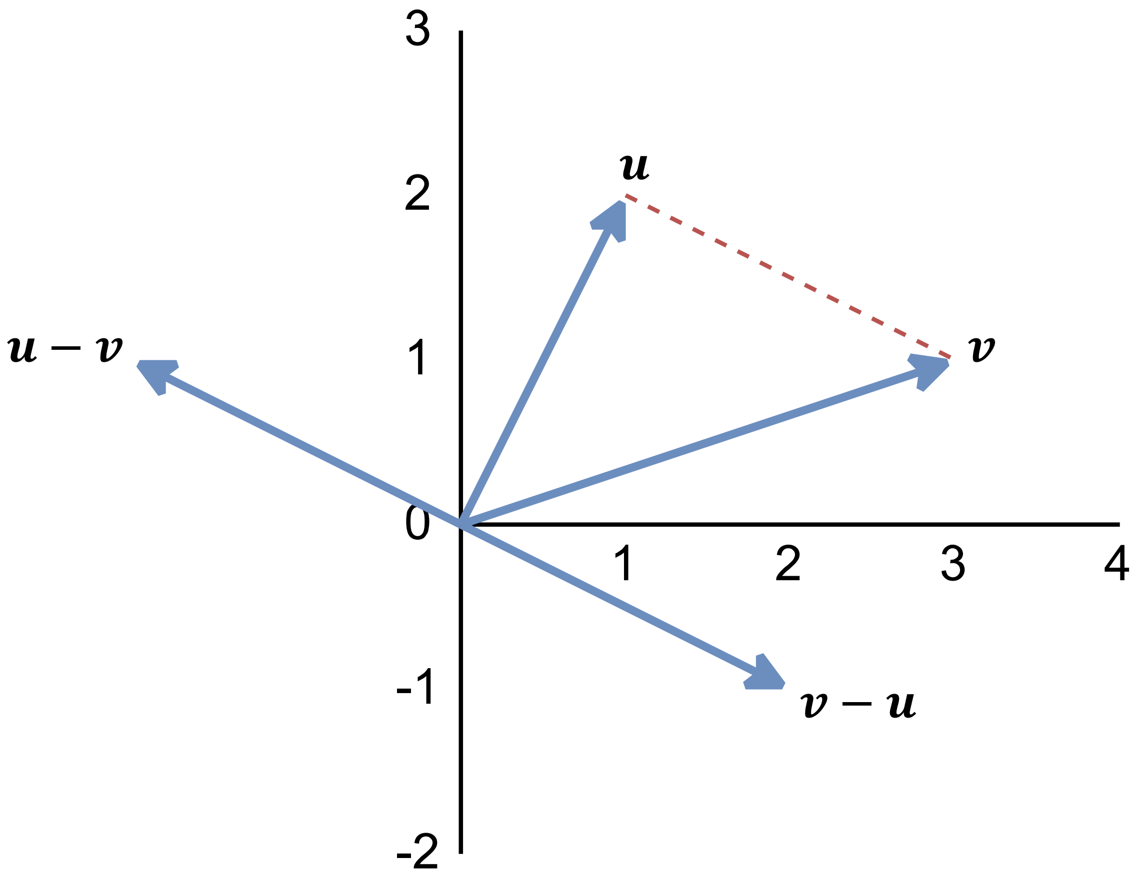
Vector addition: two vectors $\boldsymbol{u}, \boldsymbol{v}$ can be added,

$$\boldsymbol{u} + \boldsymbol{v} = \begin{bmatrix} u_1 + v_1 \\ \vdots \\ u_n + v_n \end{bmatrix}$$

Properties:

- Commutative: $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$
- Associative: $(\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w} = \boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w})$

Vector subtraction



Vector subtraction: two vectors $\boldsymbol{u}, \boldsymbol{v}$ can be subtracted,

$$\boldsymbol{u} - \boldsymbol{v} = \begin{bmatrix} u_1 - v_1 \\ \vdots \\ u_n - v_n \end{bmatrix}$$

Property: $\boldsymbol{u} - \boldsymbol{v} = \boldsymbol{u} + (-\boldsymbol{v})$

In-class assignment

$$\boldsymbol{u} = \begin{bmatrix} 0.5 \\ 1 \\ -2 \end{bmatrix}, \boldsymbol{v} = \begin{bmatrix} -1 \\ 0.5 \\ 1 \end{bmatrix}$$

Calculate:

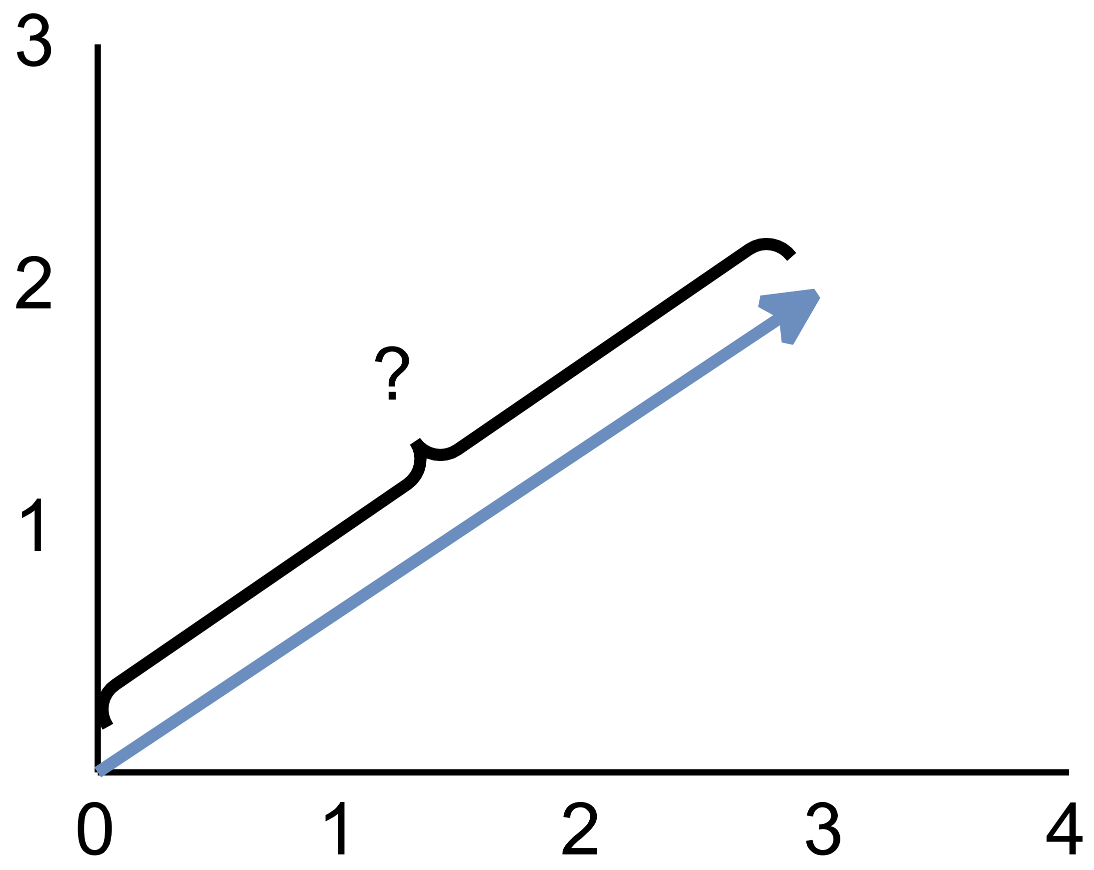
- $2\boldsymbol{u} - \boldsymbol{v}$
- $\frac{1}{2}(\boldsymbol{u} + \boldsymbol{v})$

$$2\boldsymbol{u} - \boldsymbol{v} = \begin{bmatrix} 2 \cdot 0.5 \\ 2 \cdot 1 \\ 2 \cdot -2 \end{bmatrix} - \begin{bmatrix} -1 \\ 0.5 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ -4 \end{bmatrix} - \begin{bmatrix} -1 \\ 0.5 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1.5 \\ -5 \end{bmatrix}$$

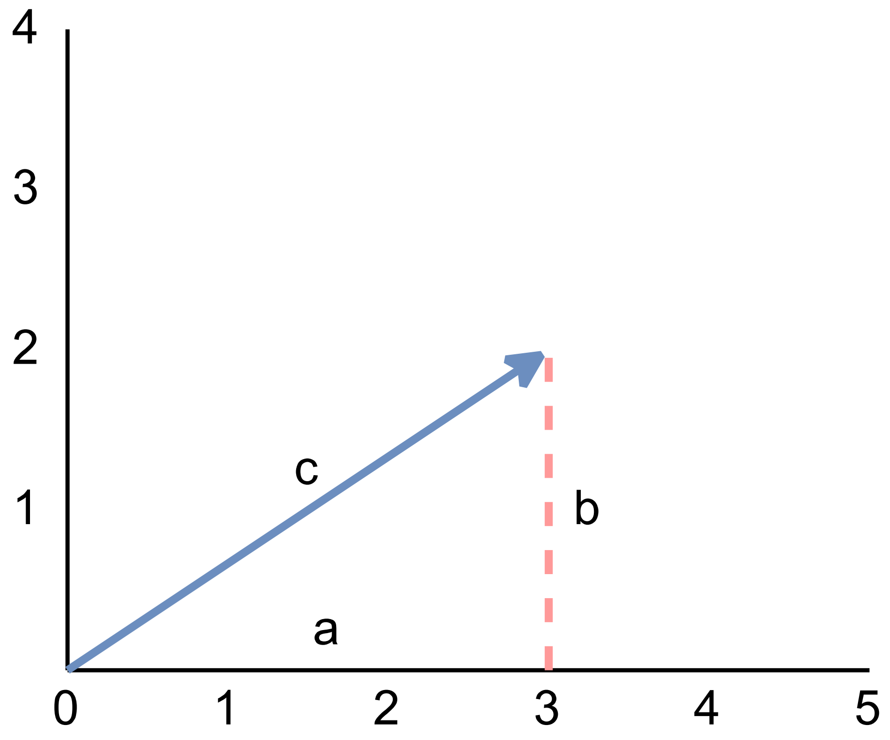
$$\frac{1}{2}(\boldsymbol{u} + \boldsymbol{v}) = \frac{1}{2}\begin{bmatrix} 0.5 + -1 \\ 1 + 0.5 \\ -2 + 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -0.5 \\ 1.5 \\ -1 \end{bmatrix} = \begin{bmatrix} -0.25 \\ 0.75 \\ -0.5 \end{bmatrix}$$

How do we find the length of a vector?

How do we find the length of a vector?



Euclidean length



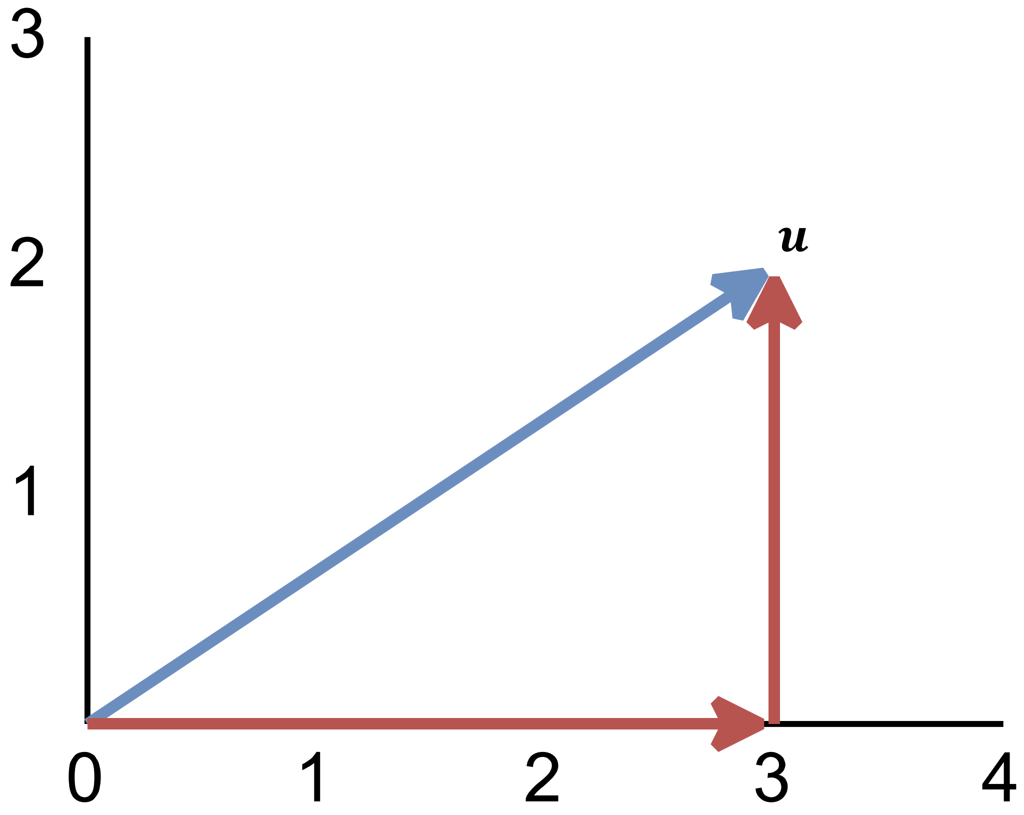
Use the Pythagorean theorem to calculate the vector length:

$$c = \sqrt{a^2 + b^2} = \sqrt{3^2 + 2^2} \approx 3.61$$

Generalization across d dimension for a vector $\boldsymbol{u}$:

$$\sqrt{\sum_{i=1}^{d} u_i^2}$$

Manhattan length



Manhattan length: length by `traveling' along each axis.

$$\sum_{i=1}^{d} |u_i| = 3 + 2 = 5$$

p-norms

The p-norm is a generalization over all such lengths:

$$\|\boldsymbol{u}\|_p = \left(\sum_{i=1}^{d} |u_i|^p \right)^{\frac{1}{p}}$$

Norm	Common names
$\|\boldsymbol{u}\|_1$	ℓ_1-norm, Manhattan length
$\|\boldsymbol{u}\|_2$	ℓ_2-norm, vector length, Euclidean norm
$\|\boldsymbol{u}\|_\infty$	Infinity norm, maximum norm

p-norm properties

The p-norm has the following properties:

1. Triangle inequality:

$$\|\boldsymbol{u} + \boldsymbol{v}\|_p \leq \|\boldsymbol{u}\|_{\boldsymbol{p}} + \|\boldsymbol{v}\|_{\boldsymbol{p}}$$

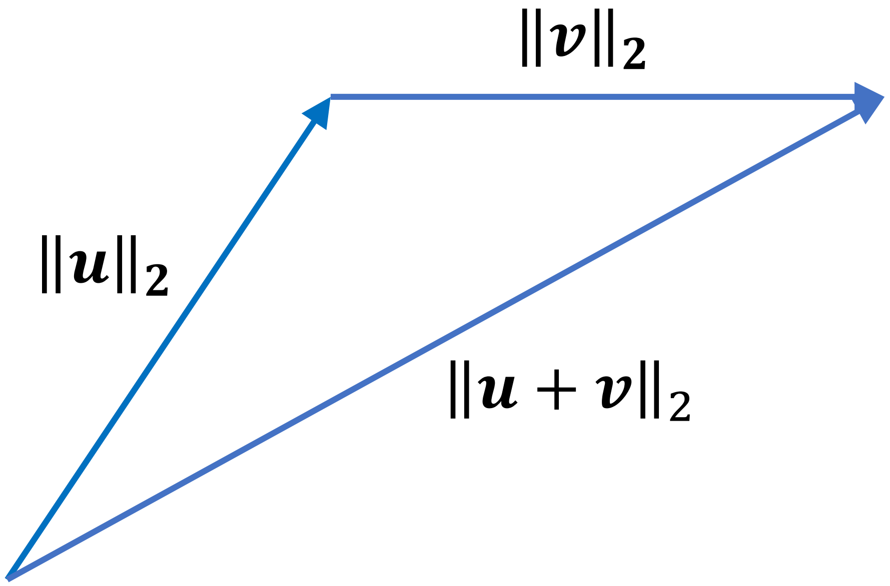
2. Absolutely scalable:

$$\|a\boldsymbol{u}\|_p = |a|\|\boldsymbol{u}\|_p$$

3. For all vectors except 0^d:

$$\|\boldsymbol{u}\|_p > \mathbf{0}$$

Triangle inequality



$$\|\boldsymbol{u}+\boldsymbol{v}\|_2 \leq \|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2$$

In-class assignment

$$\|\boldsymbol{u}\|_p = \left(\sum_{i=1}^{d} |u_i|^p\right)^{\frac{1}{p}} \qquad \boldsymbol{v} = \begin{bmatrix} -1 \\ 0.5 \\ 1 \end{bmatrix}$$

Calculate:

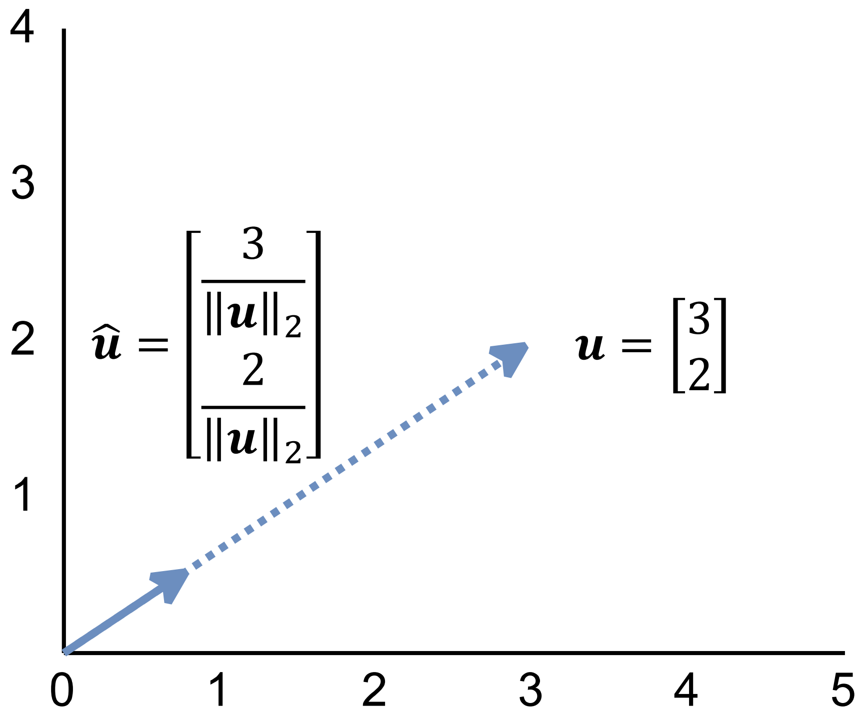
- $\|\boldsymbol{v}\|_1$
- $\|\boldsymbol{v}\|_2$

$$\|\boldsymbol{v}\|_1 = |-1| + |0.5| + |1| = 1 + 0.5 + 1 = 2.5$$

$$\|\boldsymbol{v}\|_2 = \sqrt{-1^2 + 0.5^2 + 1^2} = \sqrt{1 + 0.25 + 1} = \sqrt{2.25} = 1.5$$

Unit vectors

Definition
$\boldsymbol{u}$ is a unit vector iff $\|\boldsymbol{u}\|_2 = 1$

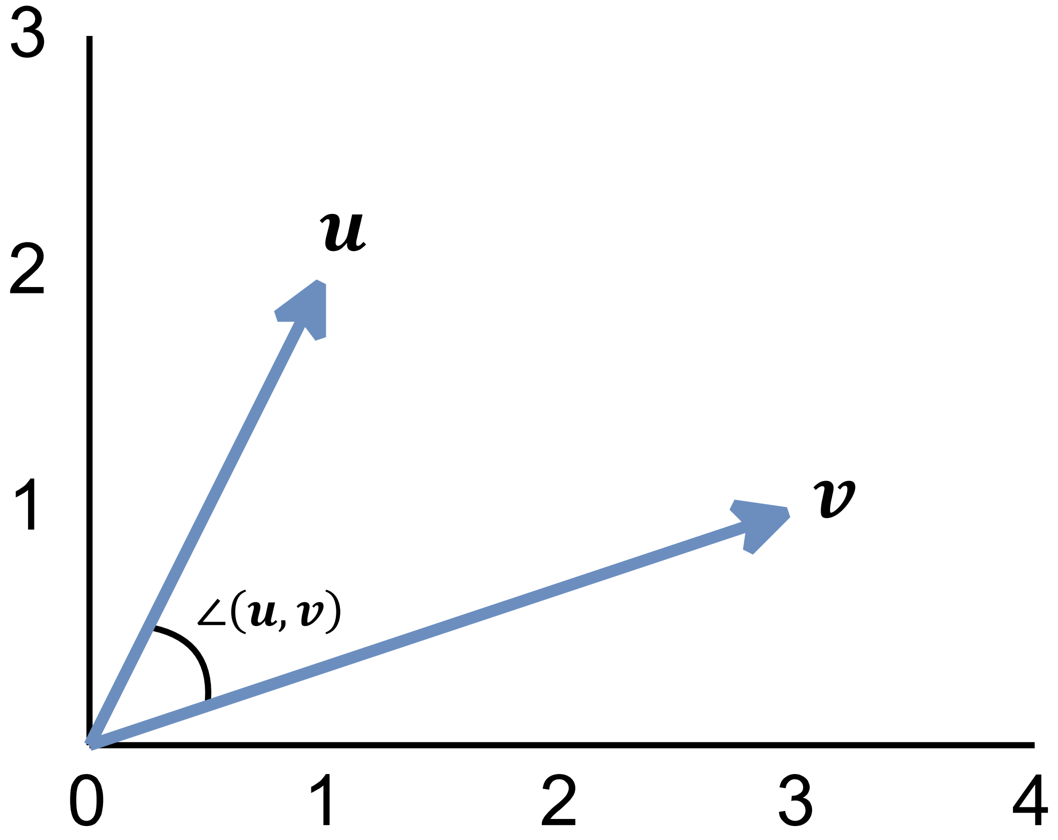


Any vector $\boldsymbol{u}$, except 0^d, can be scaled to a unit vector $\hat{\boldsymbol{u}}$:

$$\hat{\boldsymbol{u}} = \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}$$

How do we find the angle between two vectors?

How do we find the angle between two vectors?



Dot product

Definition: dot product

$$\boldsymbol{u} \cdot \boldsymbol{v} = \sum_{i=1}^{d} u_i v_i$$

Example

$$\boldsymbol{u} = \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix}, \boldsymbol{v} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{u} \cdot \boldsymbol{v} &= (2 \cdot 2) + (0 \cdot -1) + (3 \cdot 1) \\ &= 7 \end{aligned}$$

Note: the *dot product* is also known as the *inner product*.

Dot product properties

- The dot product is commutative:

$$\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{v} \cdot \boldsymbol{u}$$

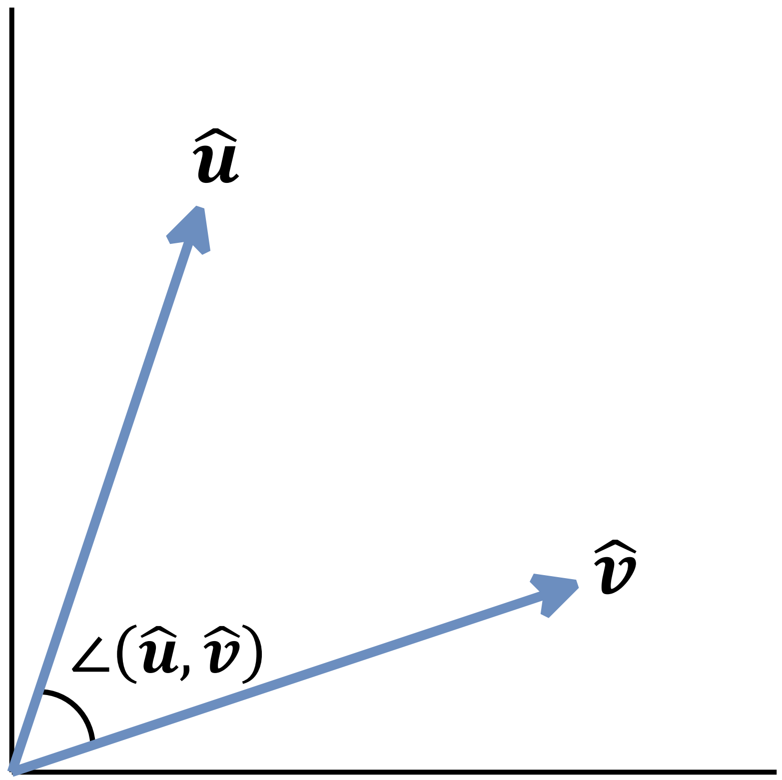
- The dot product is distributive over vector addition:

$$\boldsymbol{u} \cdot (\boldsymbol{v} + \boldsymbol{w}) = \boldsymbol{u} \cdot \boldsymbol{v} + \boldsymbol{u} \cdot \boldsymbol{w}$$

- Scalar multiplication:

$$(a\boldsymbol{u}) \cdot (b\boldsymbol{v}) = ab(\boldsymbol{u} \cdot \boldsymbol{v})$$

Cosine similarity of unit vectors



Definition

$$\cos\left(\angle(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})\right) = \hat{\boldsymbol{u}} \cdot \hat{\boldsymbol{v}} = \sum_{i=1}^{d} \hat{u}_i \cdot \hat{v}_i$$

Cosine similarity of non-unit vectors

A vector $\boldsymbol{u}$ can be normalized to a unit vector with the same direction: $\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}$. Consequently:

$$\begin{aligned}\cos\left(\angle(\boldsymbol{u}, \boldsymbol{v})\right) &= \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} \cdot \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2} \\ &= \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2}\end{aligned}$$

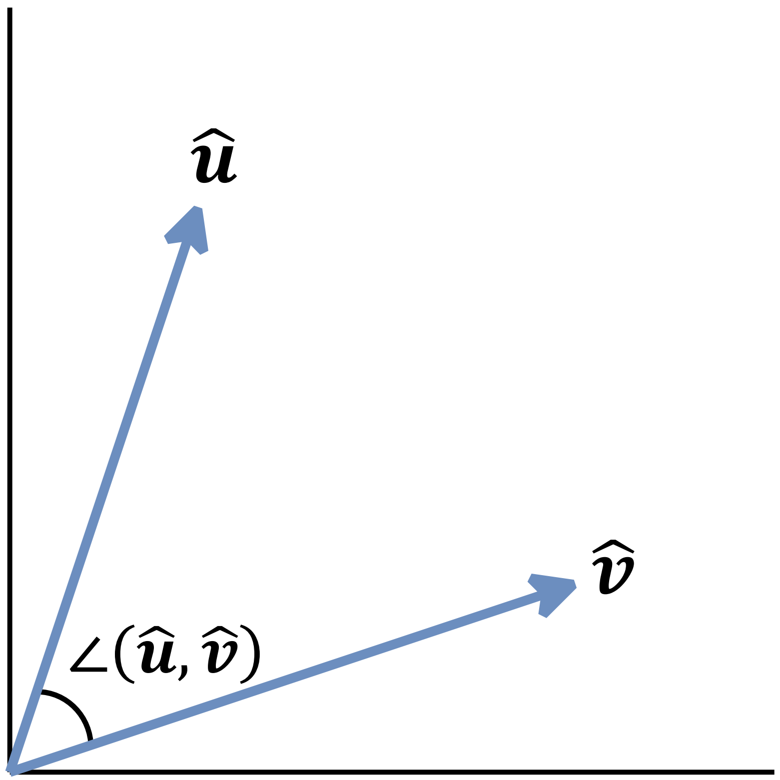
In-class assignment

$$\cos(\angle(\boldsymbol{u},\boldsymbol{v})) = \frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2} \qquad \boldsymbol{u} = \begin{bmatrix} 0.5 \\ 1 \\ -2 \end{bmatrix}, \boldsymbol{v} = \begin{bmatrix} -1 \\ 0.5 \\ 1 \end{bmatrix}$$

Calculate $\cos(\angle(\boldsymbol{u},\boldsymbol{v}))$

$$\begin{aligned}
\|\boldsymbol{u}\|_2 &= \sqrt{0.5^2 + 1^2 + -2^2} = \sqrt{0.25 + 1 + 4} = \sqrt{5.25} \\
\|\boldsymbol{v}\|_2 &= 1.5 \\
\cos(\angle(\boldsymbol{u},\boldsymbol{v})) &= \frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2} \\
&= \frac{0.5\cdot -1 + 1\cdot 0.5 + -2\cdot 1}{1.5\cdot\sqrt{5.25}} \\
&= \frac{-2}{1.5\cdot\sqrt{5.25}} \\
&\approx -0.58
\end{aligned}$$

Why does $\frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2}$ compute $\cos(\angle(\boldsymbol{u}, \boldsymbol{v}))$?



Prerequisites

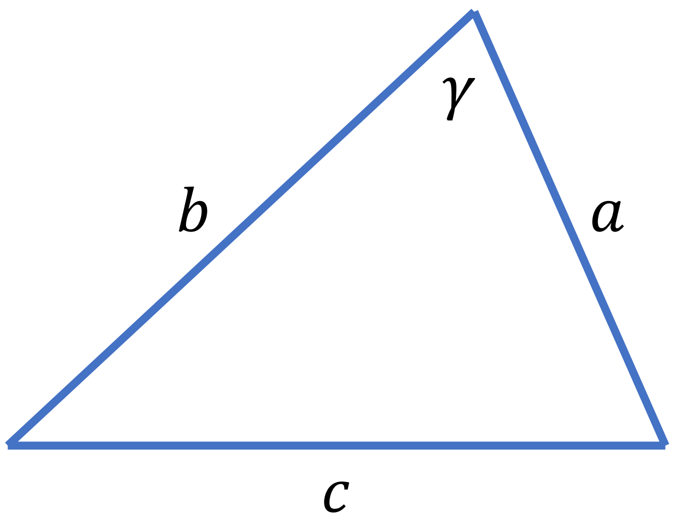
- The ℓ_2 norm can be defined in terms of the dot product:

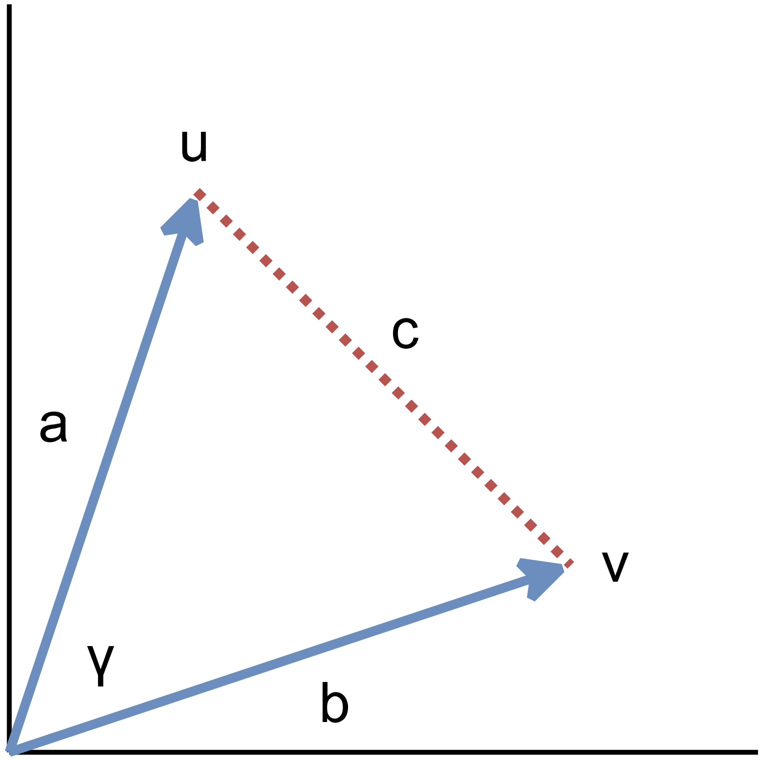
$$\|\boldsymbol{u}\|_2 = \sqrt{\sum_{i=1}^{d} u_i^2} = \sqrt{\boldsymbol{u} \cdot \boldsymbol{u}}$$

- Also observe that:

$$\|\boldsymbol{u}\|_2^2 = \boldsymbol{u} \cdot \boldsymbol{u}$$

Law of cosines





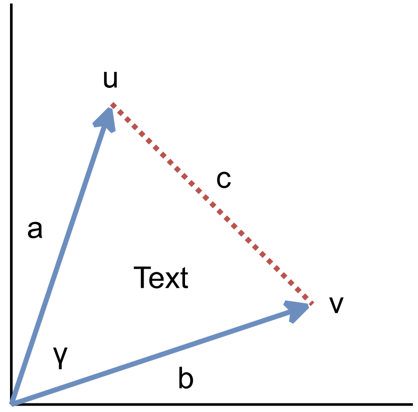
Law of cosines:

$c^2 = a^2 + b^2 - 2ab\cos\gamma$

Law of cosines (applied):

$$a = \|\boldsymbol{u}\|_2$$
$$b = \|\boldsymbol{v}\|_2$$
$$c = \|\boldsymbol{u} - \boldsymbol{v}\|_2$$

Solve for $\cos(\angle(\boldsymbol{u}, \boldsymbol{v}))$

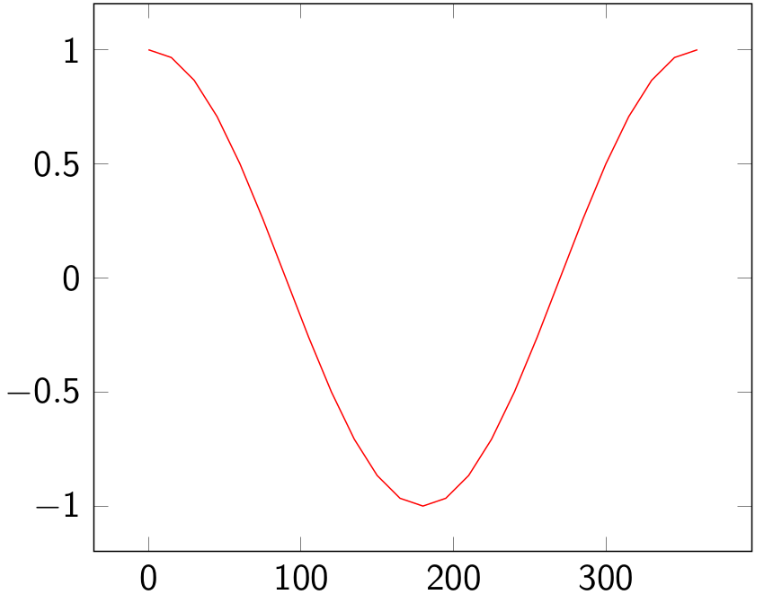


	Step
$c^2 = a^2 + b^2 - 2ab\cos\gamma$	
$\lVert\boldsymbol{u}-\boldsymbol{v}\rVert_2^2 = \lVert\boldsymbol{u}\rVert_2^2 + \lVert\boldsymbol{v}\rVert_2^2 - 2\lVert\boldsymbol{u}\rVert_2\lVert\boldsymbol{v}\rVert_2\cos(\angle(\boldsymbol{u},\boldsymbol{v}))$	
$(\boldsymbol{u}-\boldsymbol{v})\cdot(\boldsymbol{u}-\boldsymbol{v}) = \boldsymbol{u}\cdot\boldsymbol{u} + \boldsymbol{v}\cdot\boldsymbol{v} - 2\lVert\boldsymbol{u}\rVert_2\lVert\boldsymbol{v}\rVert_2\cos(\angle(\boldsymbol{u},\boldsymbol{v}))$	$\lVert\boldsymbol{u}\rVert_2^2 = \boldsymbol{u}\cdot\boldsymbol{u}$
$\boldsymbol{u}\cdot\boldsymbol{u} - 2(\boldsymbol{u}\cdot\boldsymbol{v}) + \boldsymbol{v}\cdot\boldsymbol{v} = \boldsymbol{u}\cdot\boldsymbol{u} + \boldsymbol{v}\cdot\boldsymbol{v} - 2\lVert\boldsymbol{u}\rVert_2\lVert\boldsymbol{v}\rVert_2\cos(\angle(\boldsymbol{u},\boldsymbol{v}))$	Distributivity over vector addition
$-2(\boldsymbol{u}\cdot\boldsymbol{v}) = -2\lVert\boldsymbol{u}\rVert_2\lVert\boldsymbol{v}\rVert_2\cos(\angle(\boldsymbol{u},\boldsymbol{v}))$	Eliminate duplicates on both sides
$\boxed{\dfrac{\boldsymbol{u}\cdot\boldsymbol{v}}{\lVert\boldsymbol{u}\rVert_2\lVert\boldsymbol{v}\rVert_2} = \cos(\angle(\boldsymbol{u},\boldsymbol{v}))}$	Divide both sides by $-2\lVert\boldsymbol{u}\rVert_2\lVert\boldsymbol{v}\rVert_2$

Textbook definition of cosine similarity

Interpretation of cosine similarity

Cosine similarity	Angle (degrees)	Description
$\cos\angle(\boldsymbol{u},\boldsymbol{v}) = 1$	0	Same direction
$\cos\angle(\boldsymbol{u},\boldsymbol{v}) = 0$	90	Orthogonal
$\cos\angle(\boldsymbol{u},\boldsymbol{v}) = -1$	180	Opposite



Dot product of unnormalized vectors

Question: how should we interpret the dot product of *unnormalized* vectors?

Dot product	Angle (degrees)
$\boldsymbol{u} \cdot \boldsymbol{v} > 0$	< 90
$\boldsymbol{u} \cdot \boldsymbol{v} = 0$	90
$\boldsymbol{u} \cdot \boldsymbol{v} < 0$	> 90

As we will see, the dot product is a very useful similarity function.

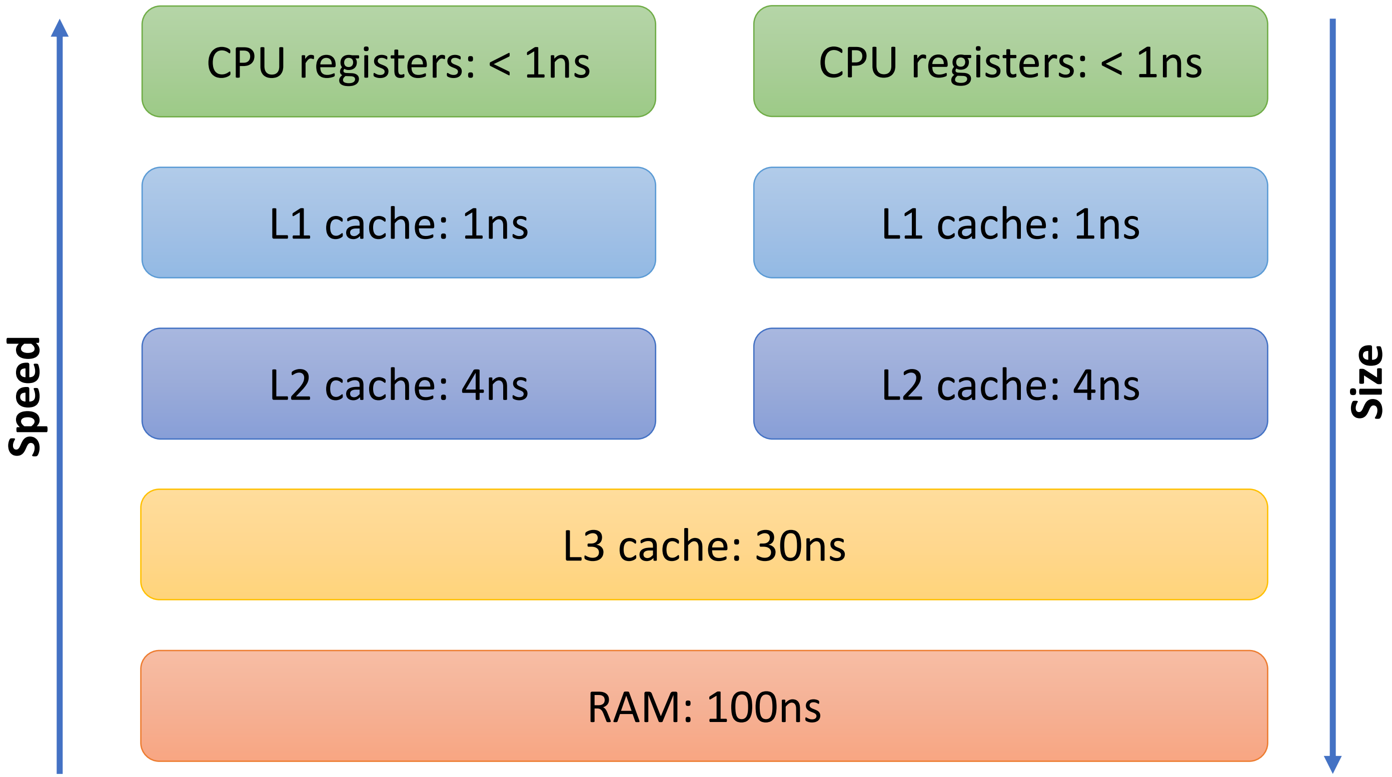
How is the dot product computed in hardware?

Naïve Python implementation:

```
def dot(u, v):
  return sum([ui * vi for (ui, vi) in zip(u, v)])
```

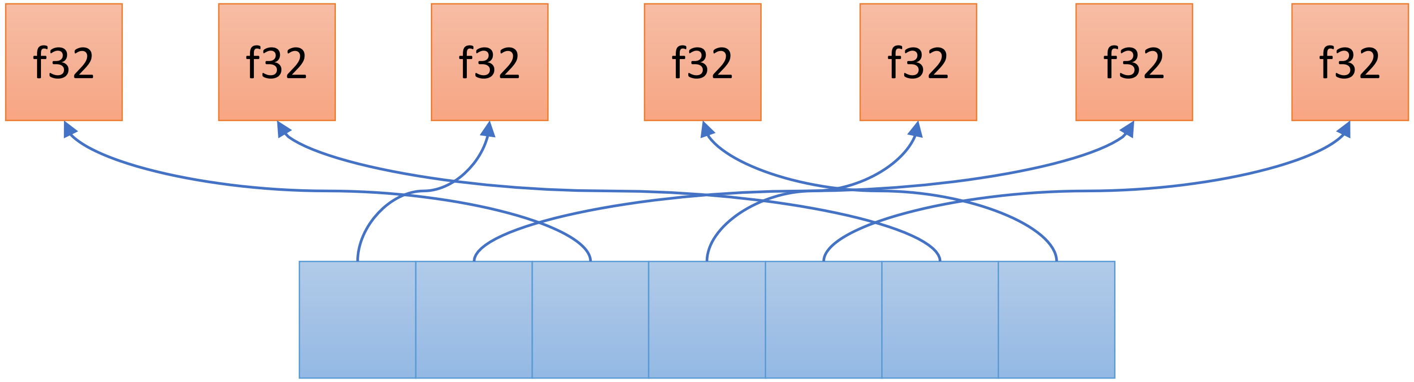
This is excessively slow!

Memory hierarchy



Contiguous vs non contiguous memory

Java `List<Float>`, Python lists:



Java `float[]`, numpy/PyTorch array:

Example timings

What	Time (ns)	Floats per clock cycle	Speedup compared to *boxed (shuffled)*
Unboxed	342,079	0.33	11.57
Unboxed (shuffled)	341,971	0.33	11.58
Boxed	1,133,880	0.10	3.49
Boxed (shuffled)	3,958,834	0.03	1.00

- Running times of computing the dot product of two vectors:
  - 500,000 components
  - single-precision floating point numbers
  - Rust + LLVM
  - AMD Ryzen 3700X
- Shuffling makes memory non-contiguous in boxed arrays

Single Instruction, Multiple Data

Regular CPU multiplication: $0.5 \times 2.0 = 1.0$

SIMD multiplication:

$$\begin{bmatrix} 0.5 \\ 1.0 \\ -1.0 \\ -1.0 \end{bmatrix} \times \begin{bmatrix} 2.0 \\ 1.0 \\ 1.5 \\ -2.0 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1.0 \\ -1.5 \\ 2 \end{bmatrix}$$

Example

```
let mut sums = _mm_setzero_ps();

while u.len() >= 4 {
    let ux4 = _mm_loadu_ps(&u[0] as *const f32);
    let vx4 = _mm_loadu_ps(&v[0] as *const f32);

    sums = _mm_add_ps(_mm_mul_ps(ux4, vx4), sums);

    u = &u[4..];
    v = &v[4..];
}

sse_add(sums) + dot_unvectorized(u, v)
```

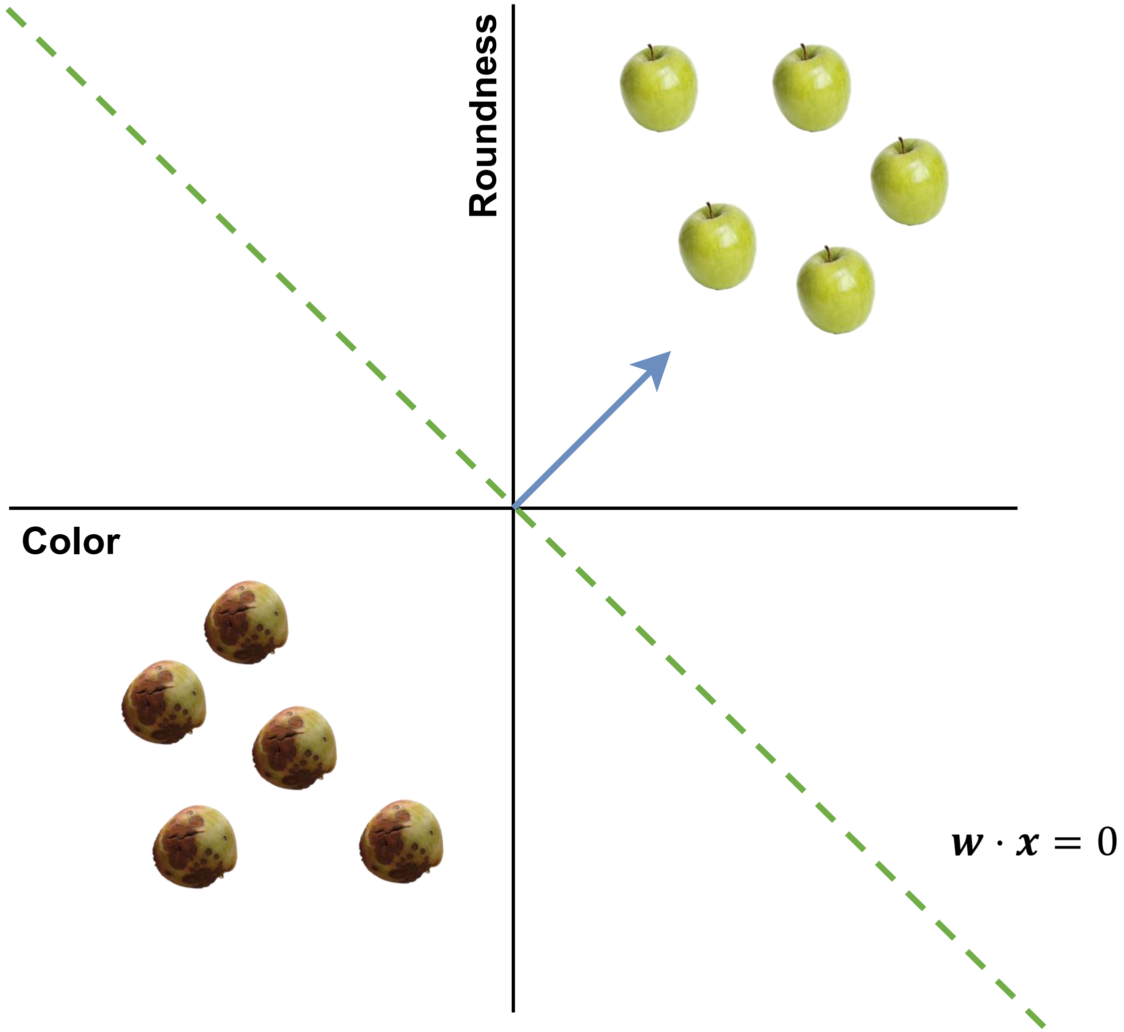
Dot product with SIMD

What	Time (ns)	Float pairs per clock cycle	Speedup compared to *scalar*
Scalar	339	0.34	1.00
SSE	81	1.44	4.19
AVX	38	3.06	8.92
AVX + FMA	34	3.42	9.97

- Running times of computing the dot product of two vectors:
  - 512 vector components
  - single-precision floating point numbers
  - Rust + LLVM
  - AMD Ryzen 3700X
- AVX + FMA DP is 10x faster than scalar DP for 512-component vectors

Logistic regression

Linear binary classifier



Goal: separate two classes. **Here:** good apples and bad apples.

Input: instances as vectors. **Here:** $\begin{bmatrix} \text{color} \\ \text{roundness} \end{bmatrix}$

Classifier: vector $\boldsymbol{w}$ pointing towards positive instances. **Here:** good apples

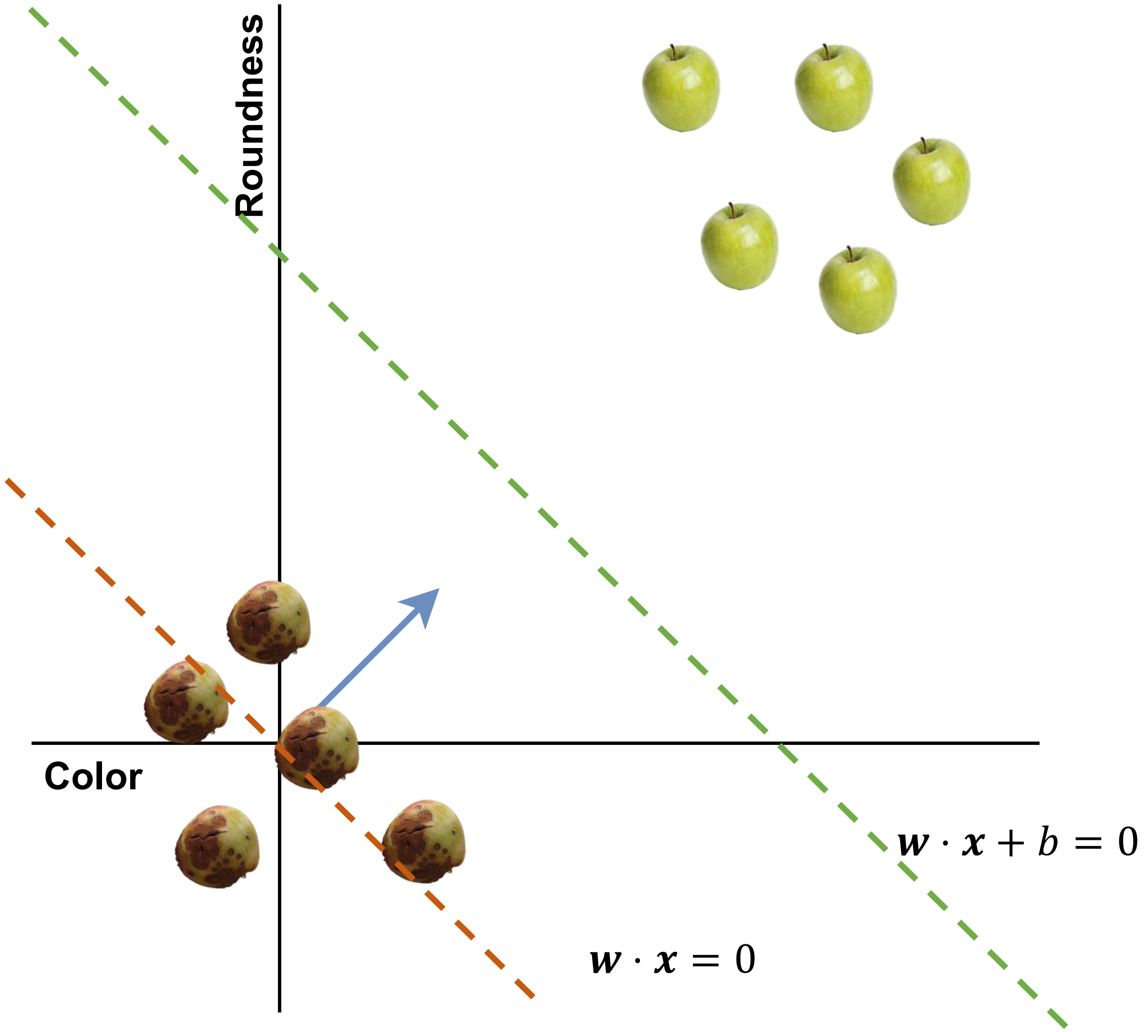
How: given an apple represented as $\boldsymbol{x}^{(i)}$,

$$y\left(\boldsymbol{x}^{(i)}\right) = \begin{cases} 1, & \boldsymbol{w} \cdot \boldsymbol{x}^{(i)} \geq 0 \\ 0, & \boldsymbol{w} \cdot \boldsymbol{x}^{(i)} < 0 \end{cases}$$

Decision boundary: $\boldsymbol{w} \cdot \boldsymbol{x} = 0$

Alternatively: $y\left(\boldsymbol{x}^{(i)}\right) = \operatorname{sign}\left(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)}\right)$

Linear binary classifier (bias)



Problem: in many classification scenarios a good decision boundary does not cross the origin.

Observe: the larger the dot product (negative or positive), the further an instance is removed from the boundary.

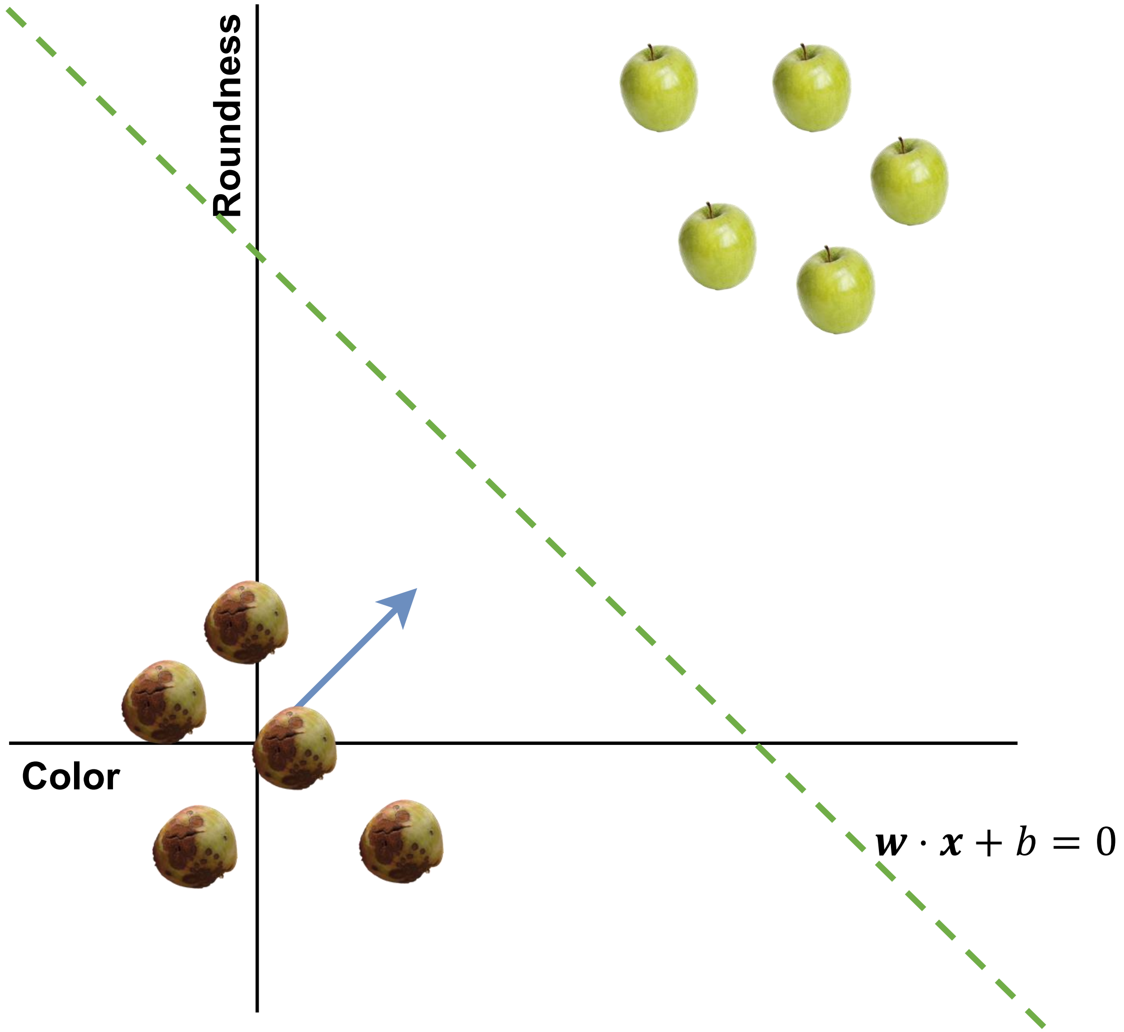
Solution: add a bias term:

- Negative bias: move the boundary towards the positive class.
- Positive bias: move the boundary towards the negative class.

Decision boundary: $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$

Classifier: $y\left(\boldsymbol{x}^{(i)}\right) = \operatorname{sign}\left(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b\right)$

Linear binary classifier (bias)



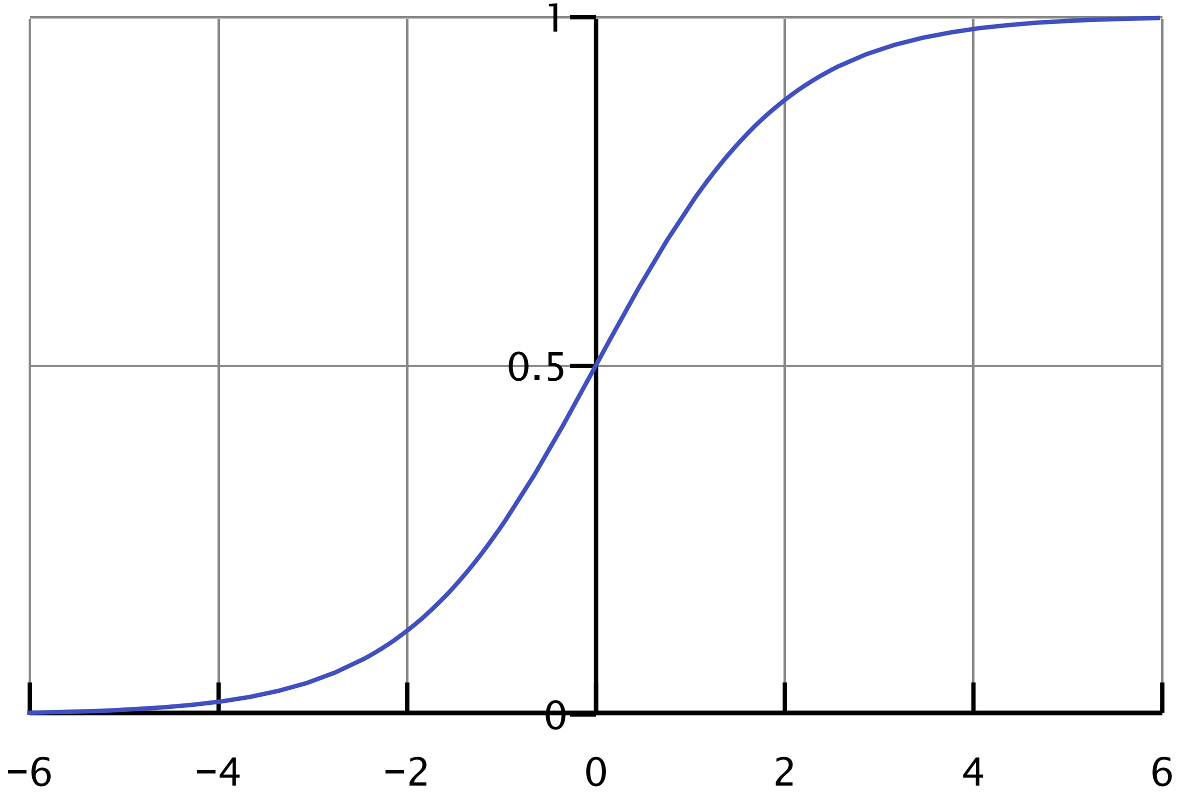
Problem: $y(\boldsymbol{x}^{(i)}) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b)$ only predicts a class, unclear how much confidence should be put into the prediction.

Idea: modify the model such that we can get a probability estimation $p(1|\boldsymbol{x}^{(i)})$ from the model.

Desiderata: modify the model such that we can get a probability estimation $p(1|\boldsymbol{x})$ from the model:

- $p(1|\boldsymbol{x}^{(i)}) = 0.5$ when $\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b = 0$
- $p(1|\boldsymbol{x}^{(i)}) > 0.5$ when $\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b > 0$
- $p(1|\boldsymbol{x}^{(i)}) < 0.5$ when $\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b < 0$

Logistic function



Definition: logistic function

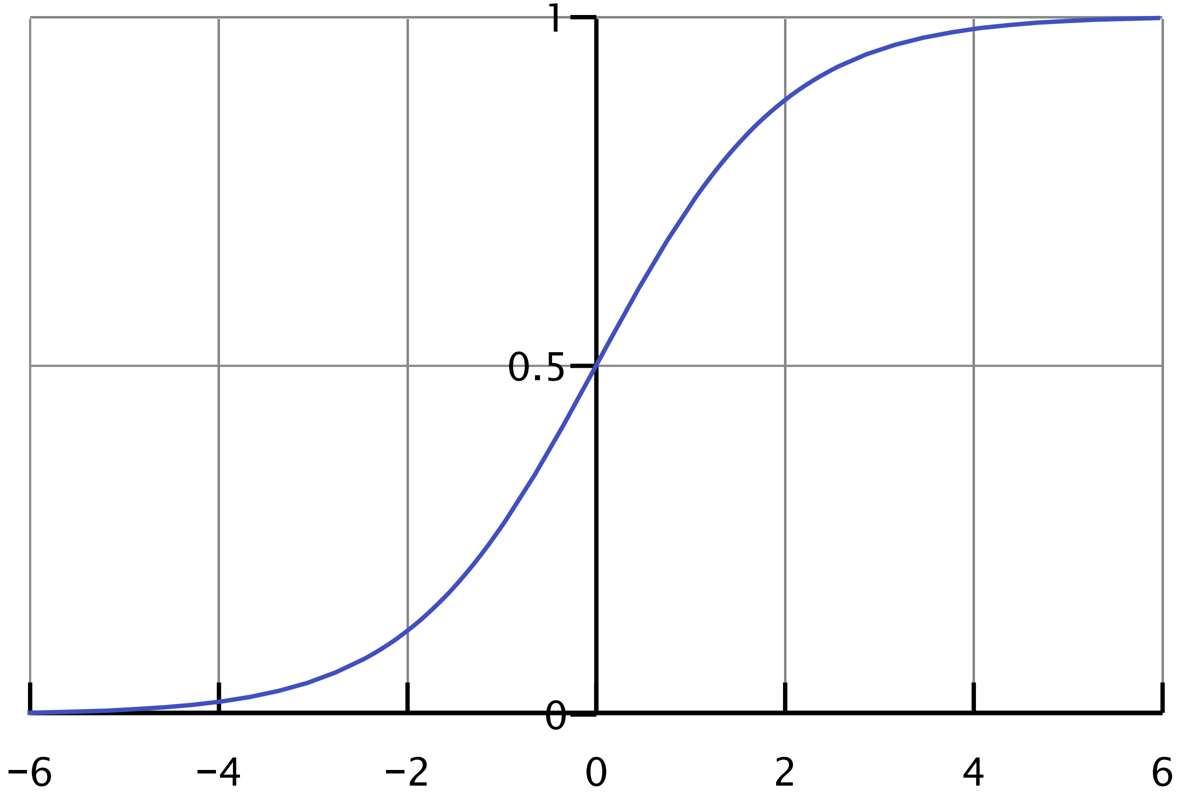
$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

$\sigma(a)$ is a **squashing function**: clips extreme values in $(0,1)$.

Note: lies between two horizontal asymptotes, never reaches 0 or 1.

Fulfills the stated desiderata.

Logistic regression classifier



Definition: logistic regression

$$p\left(1|\boldsymbol{x}^{(i)}\right) = \frac{1}{1+e^{-a}}$$

$$a = \boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b$$

$$p\left(0|\boldsymbol{x}^{(i)}\right) = 1 - p(1|\boldsymbol{x}^{(i)})$$

Example

Definition: logistic regression

$$p\left(1|\boldsymbol{x}^i\right) = \frac{1}{1+e^{-a}}$$
$$a = \boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b$$

Example model

$$\boldsymbol{w} = \begin{bmatrix} -4 \\ 3 \end{bmatrix}$$
$$b = 0$$

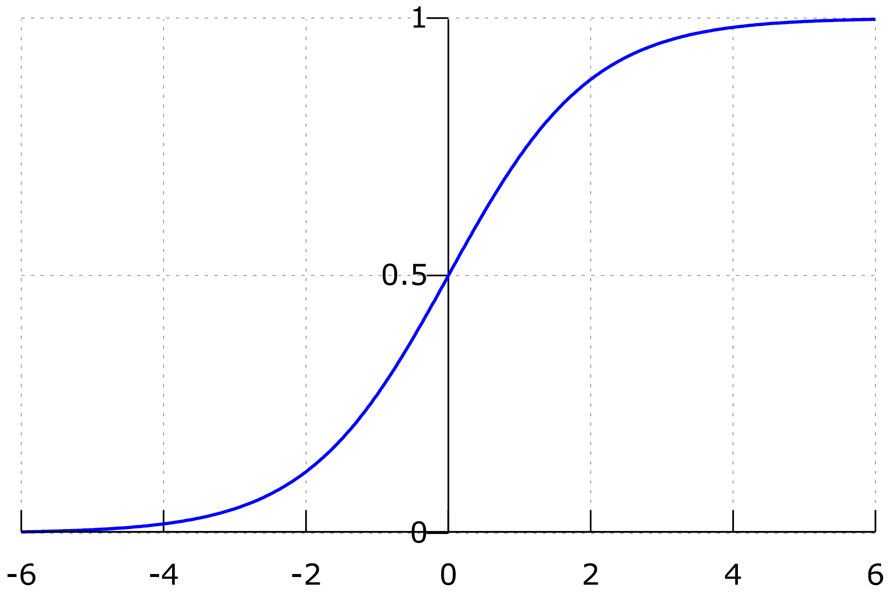
Instance	Prediction
$\boldsymbol{x}^{(1)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$	$a = -4 \cdot 1 + 3 \cdot 3 = 5$ $p\left(1\middle|\boldsymbol{x}^{(1)}\right) = \frac{1}{1+e^{-5}} \approx 0.9933$
$\boldsymbol{x}^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$a = -4 \cdot 1 + 3 \cdot 1 = -1$ $p\left(1\middle|\boldsymbol{x}^{(2)}\right) = \frac{1}{1+e^{--1}} \approx 0.2689$
$\boldsymbol{x}^{(3)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$	$a = -4 \cdot 2 + 3 \cdot 1 = -5$ $p\left(1\middle|\boldsymbol{x}^{(3)}\right) = \frac{1}{1+e^{--5}} \approx 0.0067$

Reformulation $p(0|\boldsymbol{x}^{(i)})$

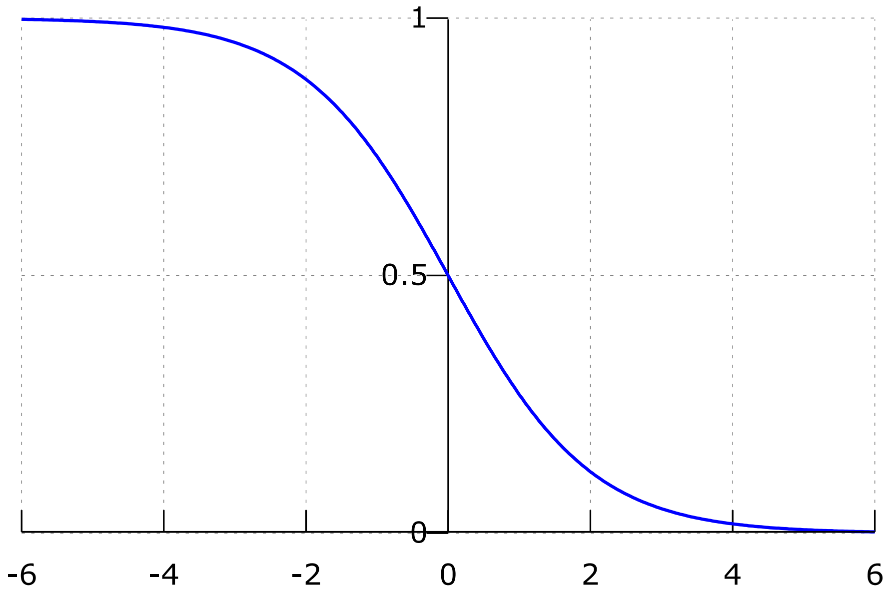
	Step
$p(0\|\boldsymbol{x}^{(i)}) = 1 - \sigma(a), a = \boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b$	
$p(0\|\boldsymbol{x}^{(i)}) = 1 - \frac{1}{1+e^{-a}}$	Expand
$p(0\|\boldsymbol{x}^{(i)}) = \frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}$	Rewrite 1
$p(0\|\boldsymbol{x}^{(i)}) = \frac{e^{-a}}{1+e^{-a}}$	Subtract
$p(0\|\boldsymbol{x}^{(i)}) = \frac{1}{\frac{1}{e^{-a}}+1}$	Divide enumerator and denominator by e^{-a}
$p(0\|\boldsymbol{x}^{(i)}) = \frac{1}{1\boxed{+e^{a}}} = \sigma(-a)$	Apply $x^{-n} = \frac{1}{x^n}$

Note that we are just flipping the sign.

$\sigma(a)$ and $\sigma(-a)$



$\sigma(a)$



$\sigma(-a)$

Why is logistic regression a linear classifier?

$$p\left(1|\boldsymbol{x}^{(i)}\right) = \frac{1}{1 + e^{-(\boldsymbol{w}\cdot\boldsymbol{x}^{(i)}+b)}}$$

	Step
$\frac{1}{1 + e^{-(\boldsymbol{w}\cdot\boldsymbol{x}^{(i)}+b)}} = \frac{1}{2}$	Decision boundary: $p\left(y\middle|\boldsymbol{x}^{(i)}\right) = \frac{1}{2}$
$1 + e^{-(\boldsymbol{w}\cdot\boldsymbol{x}^{(i)}+b)} = 2$	Simplify
$e^{-(\boldsymbol{w}\cdot\boldsymbol{x}^{(i)}+b)} = 1$	Subtract 1 from both sides
$-\left(\boldsymbol{w}\cdot\boldsymbol{x}^{(i)} + b\right) = 0$	Apply *log* to both sides
$\boxed{\boldsymbol{w}\cdot\boldsymbol{x}^{(i)} + b = 0}$	Multiply both sides by -1

Linear decision boundary

Model likelihood

If the possible classes are $Y = \{0,1\}$, we need to optimize the model parameters such that:

- $p\left(1 \middle| \boldsymbol{x}^{(i)}\right) = 1$ and $p\left(0 \middle| \boldsymbol{x}^{(i)}\right) = 0$ iff $y^{(i)} = 1$
- $p\left(1 \middle| \boldsymbol{x}^{(i)}\right) = 0$ and $p\left(0 \middle| \boldsymbol{x}^{(i)}\right) = 1$ iff $y^{(i)} = 0$

This is done by maximizing the likelihood:

$$L = \prod_{i=1}^{n} p(y^{(i)} | \boldsymbol{x}^{(i)})$$

Example

Instance	Prediction
$\boldsymbol{x}^{(1)} = \begin{bmatrix}1\\3\end{bmatrix}, y^{(1)} = 1$	$a = -4 \cdot 1 + 3 \cdot 3 = 5$ $p\left(y^{(1)} \middle\| \boldsymbol{x}^{(1)}\right) = \frac{1}{1 + e^{-5}} \approx 0.9933$
$\boldsymbol{x}^{(2)} = \begin{bmatrix}1\\1\end{bmatrix}, y^{(2)} = 1$	$a = -4 \cdot 1 + 3 \cdot 1 = -1$ $p\left(y^{(2)} \middle\| \boldsymbol{x}^{(2)}\right) = \frac{1}{1 + e^{--1}} \approx 0.2689$
$\boldsymbol{x}^{(3)} = \begin{bmatrix}2\\1\end{bmatrix}, y^{(3)} = 0$	$a = -4 \cdot 2 + 3 \cdot 1 = -5$ $p\left(y^{(3)} \middle\| \boldsymbol{x}^{(3)}\right) = 1 - \frac{1}{1 + e^{--5}} \approx 0.9933$

$$L = 0.9933 \cdot 0.2689 \cdot 0.9933 = 0.2653$$

Negative log-likelihood

First, remember that most algorithms focus on minimization:

$$NL = -\prod_{i=1}^{n} p(y^{(i)}|\boldsymbol{x}^{(i)})$$

Multiplying a lot of small numbers can lead to underflow:

$$NLL = -\log \prod_{i=1}^{n} p(y^{(i)}|\boldsymbol{x}^{(i)})$$

$\log(ab) = \log(a) + \log(b)$:

$$NLL = \sum_{i=1}^{n} -\log p(y^{(i)}|\boldsymbol{x}^{(i)})$$

Derivative of the objective function

- Let’s break this up in two steps:
  - Find the derivative of the logistic function.
  - Find the derivative of the full objective function.

Relevant derivative rules

Function	Derivative	Comment
$f(x) = c$	$f'(x) = 0$	
$f(x) = ag(x)$	$f'(x) = ag'(x)$	
$f(x) = x^n$	$f'(x) = nx^{n-1}$	Therefore: $f(x) = x, f'(x) = 1$
$f(x) = e^x$	$f'(x) = e^x$	
$f(x) = \log x$	$f'(x) = \frac{1}{x}$	
$h(x) = \frac{f(x)}{g(x)}$	$h'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{\left(g(x)\right)^2}$	Quotient rule
$h(x) = f\left(g(x)\right)$	$h'(x) = f'\left(g(x)\right)g'(x)$	Chain rule

Example

	What
$f(x) = \dfrac{\log(x)}{x^3}$	
$f'(x) = \dfrac{[\log(x)]' x^3 - \log(x)\,[x^3]'}{(x^3)^2}$	Quotient rule: $\left[\dfrac{\mathrm{f(x)}}{\mathrm{g(x)}}\right]' = \dfrac{f'(x)g(x) - f(x)g'(x)}{\left(g(x)\right)^2}$
$f'(x) = \dfrac{\frac{1}{x} x^3 - \log(x)\, 3x^2}{x^6}$	$[\log(x)]' = \dfrac{1}{x}, [x^n]' = nx^{n-1}$
$f'(x) = \dfrac{x^2 - \log(x)\, 3x^2}{x^6}$	
$f'(x) = \dfrac{1 - 3\log(x)}{x^4}$	Divide numerator and denominator by x^2.

Exercise: find the derivative

Function	Derivative	Comment
$f(x) = c$	$f'(x) = 0$	
$f(x) = ag(x)$	$f'(x) = ag'(x)$	
$f(x) = x^n$	$f'(x) = nx^{n-1}$	Therefore: $f(x) = x, f'(x) = 1$
$f(x) = e^x$	$f'(x) = e^x$	
$f(x) = \log x$	$f'(x) = \frac{1}{x}$	
$h(x) = \frac{f(x)}{g(x)}$	$h'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{\left(g(x)\right)^2}$	Quotient rule
$h(x) = f\left(g(x)\right)$	$h'(x) = f'\left(g(x)\right)g'(x)$	Chain rule

Given $f(x) = \log(e^x + x^2)$, find $f'(x)$

Example: find the derivative

	What
$f(x) = \log(e^x + x^2)$	
$f'(x) = [\log(e^x + x^2)]'$	
$f'(x) = \log'(e^x + x^2)[e^x + x^2]'$	Chain rule: $\big(f(g(x))\big)' = f'\big(g(x)\big) \cdot g'(x)$
$f'(x) = \dfrac{1}{e^x + x^2}(e^x + 2x)$	
$f'(x) = \dfrac{e^x + 2x}{e^x + x^2}$	

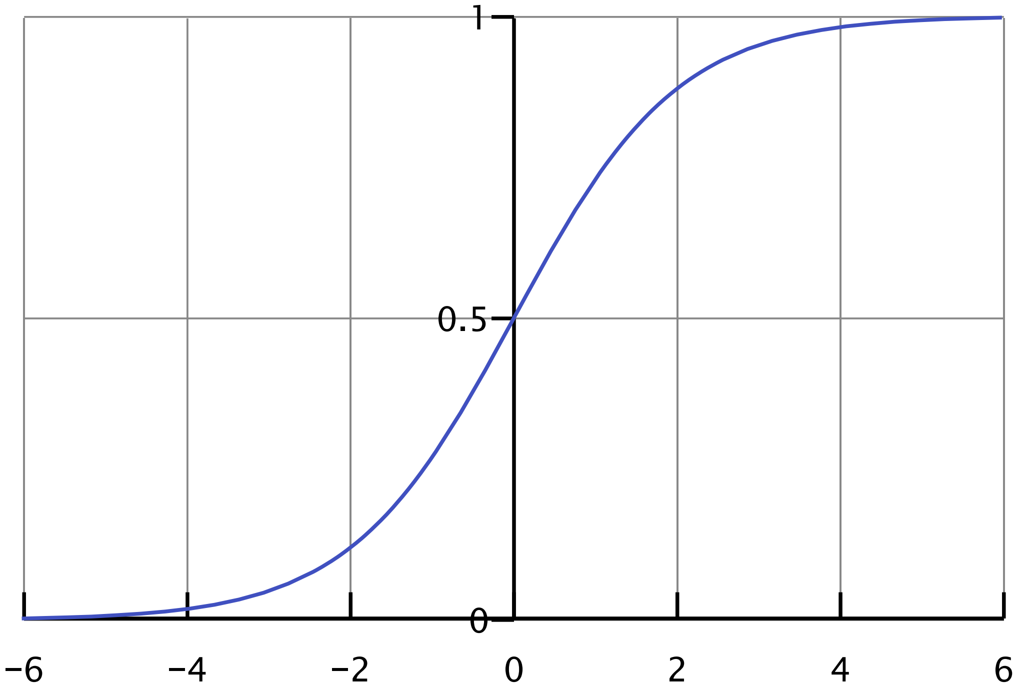
Derivative of $\sigma(a)$

	Step
$\sigma(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a}$	Multiply the numerator and denominator by e^a.
$\sigma'(a) = \left[\frac{e^a}{1+e^a}\right]'$	
$\sigma'(a) = \frac{[e^a]'(1+e^a) - e^a[1+e^a]'}{(1+e^a)^2}$	Quotient rule: $\left[\frac{\mathrm{f(x)}}{\mathrm{g(x)}}\right]' = \frac{f'(x)g(x)-f(x)g'(x)}{\left(g(x)\right)^2}$
$\sigma'(a) = \frac{e^a(1+e^a) - e^a e^a}{(1+e^a)^2}$	Apply $[e^x]' = e^x$
$\sigma'(a) = \frac{e^a + e^a e^a - e^a e^a}{(1+e^a)^2}$	Distributive property
$\sigma'(a) = \frac{e^a}{(1+e^a)^2}$	Subtract

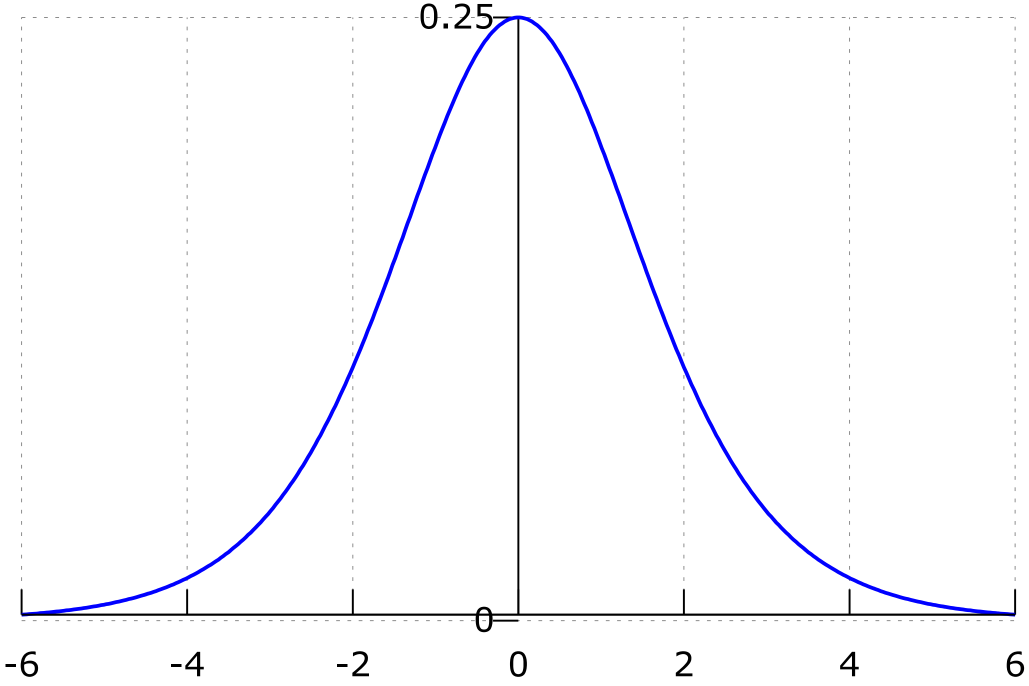
Derivative of $\sigma(a)$, continued

	Step
$\sigma'(a) = \frac{e^a}{(1+e^a)^2}$	Continue from the previous step
$\sigma'(a) = \frac{e^a}{1+e^a} \cdot \frac{1}{1+e^a}$	
$\sigma'(a) = \frac{1}{1+e^{-a}} \cdot \frac{1}{1+e^a}$	Divide the numerator and denominator of the first term by e^a
$\sigma'(a) = \sigma(a) \cdot \left(1 - \sigma(a)\right)$	

$\sigma(a)$ gradient



$$\sigma(a)$$



$$\frac{d}{dx}\sigma(a)$$

Derivative of NLL

Find the derivative of the objective function:

$$NLL = \sum_{i=1}^{n} -\log p(y^{(i)}|\boldsymbol{x}^{(i)})$$

First we will simplify our problem at bit:

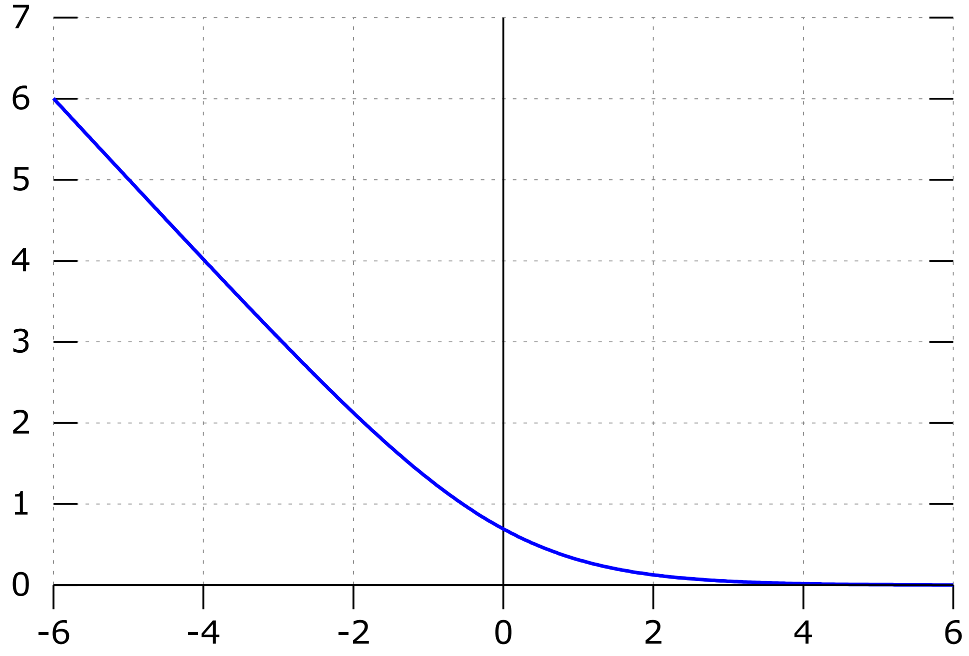
- Find the derivative of $-\log p(y^i = 1|x^{(i)})$
- Pretend that a is a scalar.

We will then later remove these simplifications.

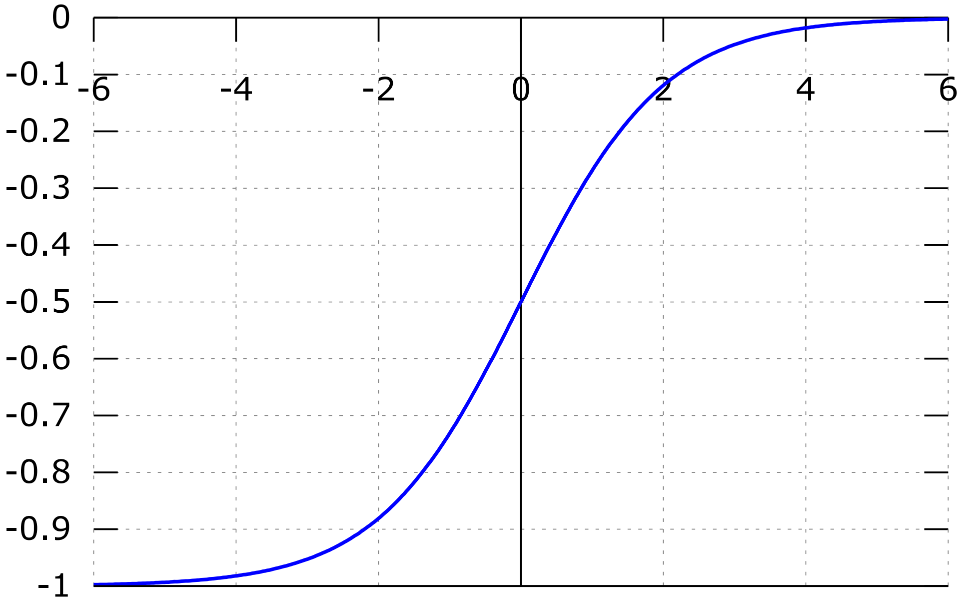
Derivative of $-\log p(1|x^{(i)})$

$-\log p(y^{(i)} = 1\|\boldsymbol{x}^{(i)}) = -\log(\sigma(a))$	
$\left[-\log p\left(y^{(i)} = 1\middle\|\boldsymbol{x}^{(i)}\right)\right]' = -[\log(\sigma(a))]'$	
$\left[-\log p\left(y^{(i)} = 1\middle\|\boldsymbol{x}^{(i)}\right)\right]' = (\log' \circ \sigma)(a)[\sigma(a)]'$	Chain rule: $\left(f\left(g(x)\right)\right)' = f'\left(g(x)\right) \cdot g'(x)$
$\left[-\log p\left(y^{(i)} = 1\middle\|\boldsymbol{x}^{(i)}\right)\right]' = -\frac{1}{\sigma(a)}\sigma(a) \cdot \left(1 - \sigma(a)\right)$	Derivative of log: $[\log(x)]' = \frac{1}{x}$
$\left[-\log p\left(y^{(i)} = 1\middle\|\boldsymbol{x}^{(i)}\right)\right]' = -(1 - \sigma(a))$	

Objective function and derivative



$$-\log p(y^{(i)} = 1|\boldsymbol{x}^{(i)})$$



$$\frac{\mathrm{d}}{\mathrm{da}}(-\log p(y^{(i)} = 1|\boldsymbol{x}^{(i)}))$$

Partial derivative

We have found the derivative:

$$\left[-\log p\left(y^{(i)} = 1 \middle| \boldsymbol{x}^{(i)}\right)\right]' = -\left(1 - \sigma(a)\right)$$

However, we want the derivative with respect to a particular weight w_i. This is the so-called **partial derivative**.

Next steps:

- Find the partial derivative the dot product with respect to w_i.
- Combine with the objective derivative that we have found.

Partial derivative (dot product)

Remember: $\boldsymbol{w} \cdot \boldsymbol{x} = w_1 x_1 + \cdots w_j x_j + \cdots w_n x_n$

$$[\boldsymbol{w} \cdot \boldsymbol{x}]_{x_j} = x_j$$

Partial derivative $-\log p(y^{(i)} = 1|x^{(i)})$

$\left[-\log p\left(y^{(i)} = 1 \middle| \boldsymbol{x}^{(i)}\right)\right]_{w_j} = -\left[\log \sigma\left(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b\right)\right]_{w_i}$	
$\left[-\log p\left(y^{(i)} = 1 \middle| \boldsymbol{x}^{(i)}\right)\right]_{w_j} = -\left(1 - \sigma(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b)\right) x_j^{(i)}$	Chain rule: $\left(f\left(g(x)\right)\right)' = f'\left(g(x)\right) \cdot g'(x)$

Partial derivative for both classes

$$\left[-\log p\left(y^{(i)} = 1 \middle| \boldsymbol{x}^{(i)}\right)\right]_{w_j} = -\left(1 - \sigma(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b)\right)x_j^{(i)}$$

$$\left[-\log p\left(y^{(i)} = 0 \middle| \boldsymbol{x}^{(i)}\right)\right]_{w_j} = -\left(-\sigma(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b)\right)x_j^{(i)}$$

Combine:

$$\begin{aligned}\left[-\log p\left(y^{(i)} \middle| \boldsymbol{x}^{(i)}\right)\right]_{w_j} &= -\left(y^{(i)} - \sigma\left(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b\right)\right) x_j^{(i)} \\ &= \left(\sigma\left(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b\right) - y^{(i)}\right)x_j^{(i)}\end{aligned}$$

Question: what is $\left[-\log p\left(y^{(i)} \middle| \boldsymbol{x}^{(i)}\right)\right]_b$?

The end