Machine Learning for Computational Linguistics

A refresher on linear algebra

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

April 14, 2016

Frequently asked questions

- The course is worth 9 ECTS.
- Term project/paper deadline will extend to semester break, but you should start working on your projects during during the semester.
- Please check the course web page (`http://coltekin.net/cagri/courses/ml/`) for reading material, slides, and assignments.

A few example (supervised) machine learning tasks

Input	Output
Email messages	spam or not
Product reviews	positive/neutral/negative
Books/blog posts/tweets	age of the author
Images of digits	the digit
Images of scenes	objects/people in the image
Music (audio) files	genre of the music
People/companies	credit risk/reliability
Sentences	syntactic representation
Questions	answers

A few example (supervised) machine learning tasks

Input				Output
x_1	x_2	x_3	...	y
30	0	0.10	...	18
60	1	1.20	...	45
20	1	−1.20	...	65
90	0	0.00	...	23
...	...	...	...	...

A few example (supervised) machine learning tasks

Input				Output
x_1	x_2	x_3	...	y
30	0	0.10	...	N
60	1	1.20	...	P
20	1	−1.20	...	N
90	0	0.00	...	P
...	...	...	...	...

Machine learning as function approximation

- We assume that data we observe is generated by an unknown functions

$$y = f(x_1, x_2, x_3, \ldots)$$

- During training we want to estimate the function f
- Once we have an estimate of f, $\hat{f}$, we use it to predict y, given an input

$$\hat{y} = \hat{f}(x_1, x_2, x_3, \ldots)$$

How do we approximate f?

- We assume that f comes from a class of functions F. For example,

$$F(x) = w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots$$

  where w_1, w_2, w_3 are parameters
- The approximation, or learning, is finding an optimum set of weights

Linear algebra

Linear algebra is the field of mathematics that studies vectors and matrices.

- A vector is an ordered sequence of numbers

$$\boldsymbol{v} = (6, 17)$$

- A matrix is a rectangular arrangement of numbers

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

- Most common application of linear algebra includes solving a set of linear equations

$$\begin{array}{rcrcr} 2x_1 & + & x_2 & = & 6 \\ x_1 & + & 4x_2 & = & 17 \end{array}$$

Why study linear algebra?

Remember our input matrix:

Input				Output
x_1	x_2	x_3	...	y
30	0	0.10	...	18
60	1	1.20	...	45
20	1	−1.20	...	65
90	0	0.00	...	23
...	...	...	...	...

Why study linear algebra?

Remember our input matrix:

Input				Output
x_1	x_2	x_3	...	y
30	0	0.10	...	18
60	1	1.20	...	45
20	1	−1.20	...	65
90	0	0.00	...	23
...	...	...	...	...

You should now be seeing vectors and matrices here.

Why study linear algebra?

In machine learning,

- We typically represent input, output, parameters as vectors or matrices.
- Some insights from linear algebra is helpful in understanding ML methods
- It makes notation concise and manageable
- In programming, many machine learning libraries make use of vector and matrices explicitly
- 'Vectorized' operations may run much faster on GPUs

Vectors: some notation

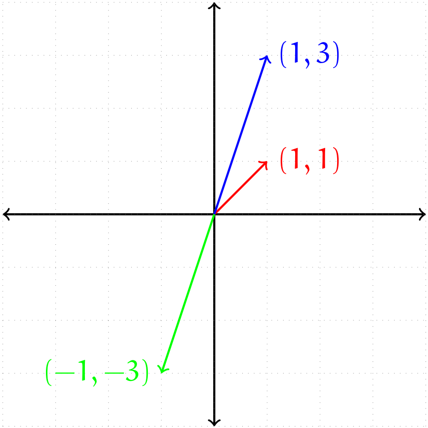
- Typical notation for vectors include

$$\boldsymbol{v} = \vec{v} = (v_1, v_2, v_3) = \langle v_1, v_2, v_3 \rangle = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

- A vector of n real numbers $\boldsymbol{v} = (v_1, v_2, \ldots v_n)$ is said to be in vector space $\mathbb{R}^n$ ($\boldsymbol{v} \in \mathbb{R}^n$).

Geometric interpretation of vectors

- Vectors are objects with a magnitude and a direction
- Geometrically, they are represented by arrows from the origin



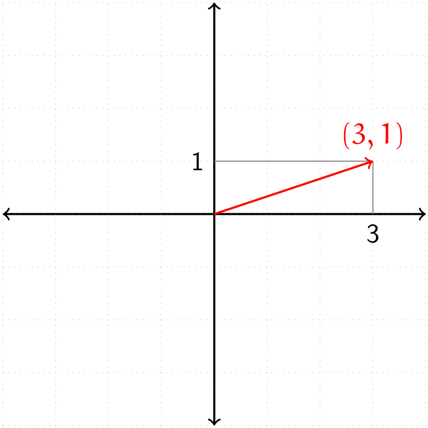
Vector norms

- Euclidian norm, or L2 (or L_2) norm is the most commonly used norm For $\boldsymbol{v} = (v_1, v_2)$,

  $$\|v\|_2 = \sqrt{v_1^2 + v_2^2}$$

  $$\|(3, 1)\|_2 = \sqrt{3^2 + 1^2} = 3.16$$

  L2 norm is often written without a subscript: $\|\boldsymbol{v}\|$



Vector norms

- Euclidian norm, or L2 (or L_2) norm is the most commonly used norm For $\boldsymbol{v} = (v_1, v_2)$,

$$\|v\|_2 = \sqrt{v_1^2 + v_2^2}$$

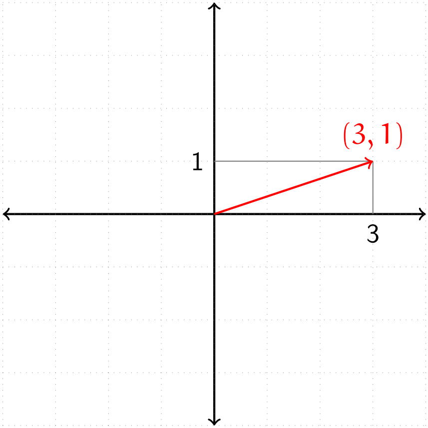
$$\|(3,1)\|_2 = \sqrt{3^2 + 1^2} = 3.16$$

L2 norm is often written without a subscript: $\|\boldsymbol{v}\|$

- Another norm often used in machine learning is L1 norm

$$\|v\|_1 = |v_1| + |v_2|$$

$$\|(3,1)\|_1 = |3| + |1| = 4$$

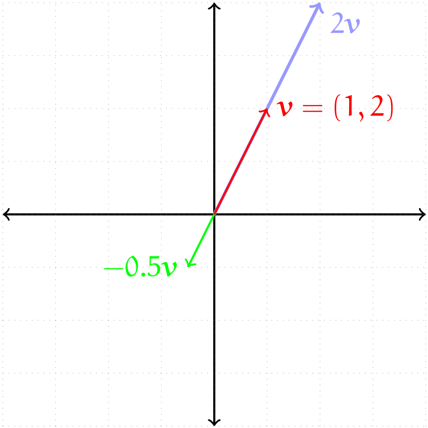


Multiplying a vector with a scalar

- For a vector $\boldsymbol{v} = (v_1, v_2)$ and a scalar α,

$$\alpha v = (\alpha v_1, \alpha v_2)$$

- multiplying with a scalar 'scales' the vector



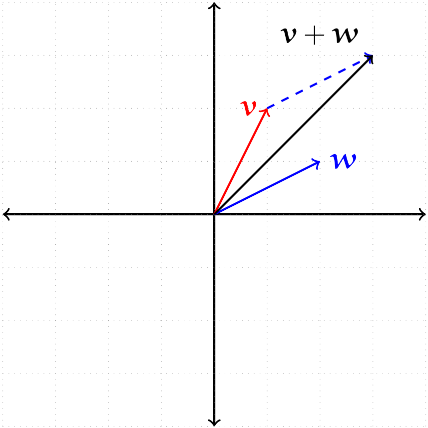
Vector addition and subtraction

- For vectors $\boldsymbol{v} = (v_1, v_2)$ and $\boldsymbol{w} = (w_1, w_2)$ and

  $$\boldsymbol{v} + \boldsymbol{w} = (v_1 + w_1, v_2 + w_2)$$

  $$(1, 2) + (2, 1) = (3, 3)$$

- $\boldsymbol{v} - \boldsymbol{w} = \boldsymbol{v} + (-\boldsymbol{w})$



Dot product

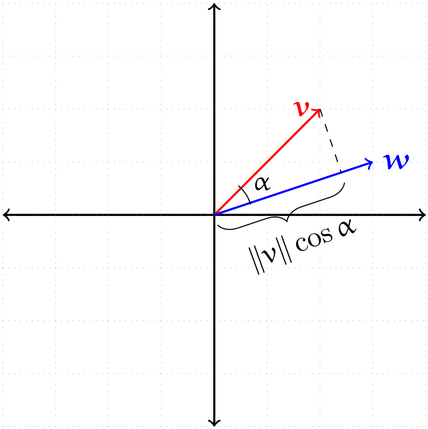
- For vectors $\boldsymbol{w} = (w_1, w_2)$ and $\boldsymbol{v} = (v_1, v_2)$,

$$\boldsymbol{wv} = w_1 v_1 + w_2 v_2$$

or,

$$\boldsymbol{wv} = \|w\| \|v\| \cos \alpha$$

- The dot product of orthogonal vectors is 0
- $\|\boldsymbol{w}\| = \boldsymbol{ww}$
- Dot product is often used as a similarity measure between two vectors.



Cosine similarity

- Cosine of the angle between two vectors

$$\cos \alpha = \frac{\boldsymbol{v}\boldsymbol{w}}{\|\boldsymbol{v}\| \|\boldsymbol{w}\|}$$

  is often used as another similarity metric, called cosine similarity
- The cosine similarity related to dot product, but ignores the magnitudes of the vectors
- For unit vectors (vectors of length 1) cosine similarity is equal to dot product

Matrices

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \dots & a_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & a_{m,3} & \dots & a_{m,n} \end{bmatrix}$$

- We can think of matrices as collection of row or column vectors
- A matrix with n rows and m columns is in $\mathbb{R}^{n \times m}$

Transpose of a matrix

Transpose of a $n \times m$ matrix is a $m \times n$ matrix whose rows are the columns of the original matrix.
Transpose of a matrix $\boldsymbol{A}$ is denoted with $\boldsymbol{A}^\mathsf{T}$.

$$\text{If } \boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix}, \boldsymbol{A}^\mathsf{T} = \begin{bmatrix} a & c & e \\ b & d & f \end{bmatrix}.$$

Multiplying a matrix with a scalar

Similar to vectors, each element is multiplied by the scalar.

$$2 \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 2 \times 2 & 2 \times 1 \\ 2 \times 1 & 2 \times 4 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 8 \end{bmatrix}$$

Matrix addition and subtraction

Each element is added to (or subtracted from) the corresponding element

$$\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + \dots a_{1k}b_{k1}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{12} = a_{11}b_{12} + a_{12}b_{22} + \dots a_{1k}b_{k2}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{1m} = a_{11}b_{1m} + a_{12}b_{2m} + \dots a_{1k}b_{km}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{21} = a_{21}b_{11} + a_{22}b_{22} + \dots a_{2k}b_{k1}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{22} = a_{21}b_{12} + a_{22}b_{22} + \dots a_{2k}b_{k2}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{2m} = a_{21}b_{1m} + a_{22}b_{2m} + \dots a_{2k}b_{km}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{n1} = a_{n1}b_{11} + a_{n2}b_{22} + \dots a_{nk}b_{k1}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{n2} = a_{n1}b_{12} + a_{n2}b_{22} + \dots a_{nk}b_{k2}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{nm} = a_{n1}b_{1m} + a_{n2}b_{2m} + \dots a_{nk}b_{km}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{pmatrix}$$

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots a_{ik}b_{kj}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{pmatrix}$$

Dot product as matrix multiplication

In machine learning literature, dot product of two vectors are often written as

$$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{v}$$

For example, $\boldsymbol{w} = (2, 2)$ and $\boldsymbol{v} = (2, -2)$,

$$\begin{bmatrix} 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = 2 \times 2 + 2 \times \; -2 = 4 - 4 = 0$$

Although, this notation is somewhat sloppy, since the result of matrix multiplication is in fact not a scalar.

Identity matrix

- A square matrix in which all the elements of the principal diagonal are ones and all other elements are zeros, is called identity matrix and often denoted $\mathbf{I}$.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Multiplying a matrix with the identity matrix does not change the original matrix.

$$\mathbf{I}\boldsymbol{A} = \boldsymbol{A}$$

Matrix multiplication as transformation

- Multiplying a vector with a matrix transforms the vector
- Some exmaples for transformaton to/from $\mathbb{R}^2$
  - Identity: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
  - 90-dgrees rotation: $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$

    In general: $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$
  - Shear: $\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$
  - Stretch along y-axis $\begin{bmatrix} 1 & 0 \\ 0 & k \end{bmatrix}$

Matrix-vector representation of a set of linear equations

Our earlier example set of linear equations

$$\begin{array}{rcrcr} 2x_1 & + & x_2 & = & 6 \\ x_1 & + & 4x_2 & = & 17 \end{array}$$

can be written as:

$$\underbrace{\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}}_{\mathbf{W}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} 6 \\ 17 \end{bmatrix}}_{\mathbf{b}}$$

One can solve the above equation using *Gaussian elimination* (we will not cover it today).

Inverse of a matrix

Inverse of a square matrix W is defined denoted W^{-1}, and defined as

$$W^{-1}W = \mathbf{I}$$

The inverse can be used to solve equation in our previous example:

$$\begin{aligned} W\mathbf{x} &= \mathbf{b} \\ W^{-1}W\mathbf{x} &= W^{-1}\mathbf{b} \\ \mathbf{I}\mathbf{x} &= W^{-1}\mathbf{b} \\ \mathbf{x} &= W^{-1}\mathbf{b} \end{aligned}$$

Determinant of a matrix

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

The above formula generalizes to higher dimensional matrices through a recursive definition, but you are unlikely to calculate it by hand. Some properties:

- A matrix is invertible if it has a non-zero determinant
- A system of linear equations has a uniqe solution if the coefficient matrix has a non-zero determinant
- Geometric interpretation of determinant is the (signed) changed in the volume of a unit (hyper)cube caused by transformation caused by the matrix

Eigen values and eigen vectors of a matrix

An eigen vector of a matrix $\boldsymbol{A}$ is such that

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$$

where λ is a scalar called eigenvalue.

- Eigen values an eigen vectors have many applications from communication theory to quantum mechanics
- A better known example (and close to home) is Google's PageRank algorighm
- We will return to them while discussing PCA

Summary & next week

- See bibliography at the end of the slides if you want a ‘more complete’ refresher/introduction
- Next week we will do a similar excursion to probability theory

Further reading

A classic reference book in the field is Strang (2009). Shifrin and Adams (2011) and Farin and Hansford (2014) are textbooks with a more practical/graphical orientation. Cherney, Denton, and Waldron (2013) and Beezer (2014) are two textbooks that are freely available!

Beezer, Robert A. (2014). *A First Course in Linear Algebra*. version 3.40. Congruent Press. ISBN: 9780984417551.

Cherney, David, Tom Denton, and Andrew Waldron (2013). *Linear algebra*. math.ucdavis.edu. URL: https://www.math.ucdavis.edu/~linear/.

Farin, Gerald E. and Dianne Hansford (2014). *Practical linear algebra: a geometry toolbox*. Third edition. CRC Press. ISBN: 9781466579569,1466579560,978-1-4665-7958-3,1466579587,9781466579590,1466579595,9781482211283,1482211289.

Shifrin, Theodore and Malcolm R Adams (2011). *Linear Algebra. A Geometric Approach*. 2nd. W. H. Freeman. ISBN: 1429215216, 978-1429215213.

Strang, Gilbert (2009). *Introduction to Linear Algebra, Fourth Edition*. 4th ed. Wellesley Cambridge Press. ISBN: 9780980232714.