Basics for Statistical Machine Learning

Linear Algebra Basics

Mikaela Keller

IDIAP Research Institute
Martigny, Switzerland
mkeller[at]idiap.ch

www.pascal-network.org

July 2nd, 2007

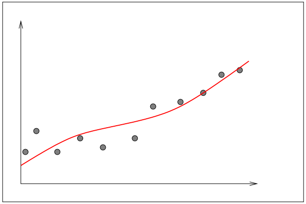
Motivation

Linear Algebra Basics

Motivation

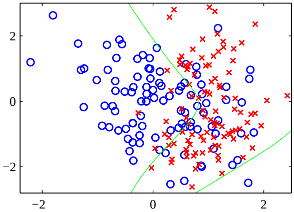
Linear Algebra Basics

Motivation
Concrete Example: Regression



- **Determination of abalone age by:**
- Counting the number of rings in the shell through a microscope $\leftarrow$ time-consuming task.
- Through other measurements: sex, diameter, height, whole weight, shell weight, etc. $\leftarrow$ easy to obtain.
- Regression problem: training examples $= \{($easy measurements, age$)\}$. We want to predict the age of abalone from the easy measurements alone.

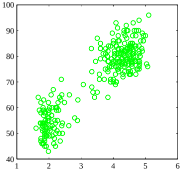
Motivation
Concrete Example: Classification

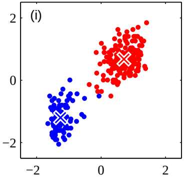


- **Written digits classification:**
- Automatic recognition of postal code from scanned mail.
- Classification problem: training examples = {(image, actual digit)}. We want to predict the correct digit from a new image.

Motivation

Concrete Example: Density Estimation / Clustering





- **Data compression / data visualization / data exploration:**
- Time between two eruptions vs duration of the previous eruption.
- Unsupervised problem: training examples = {(measurement)}. We want to "organize" the information contained in the measurements.

Motivation

- Most of the problems described previously end up reformulated into:
  - curves or surfaces to be discovered,
  - *ie* systems of equations with unknowns to be solved,
  - *ie* matrices manipulation operations.

$\Rightarrow$ **Linear Algebra**.

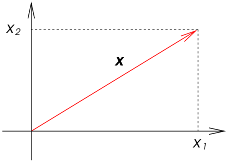
- Diverse sources of uncertainty:
  - limited amount of examples,
  - noise in the measurements,
  - randomness inherent to the observed phenomena, etc.

$\Rightarrow$ **Probability Theory**

Vectors

- Examples **x** are usually represented as vectors of m components:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \quad \mathbf{x}^T = (x_1, \ldots, x_m) \,.$$



- Inner product (*aka* dot product, scalar product):

$$\mathbf{x}^T \mathbf{y} = (x_1, \ldots, x_m) \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = x_1 y_1 + \ldots + x_m y_m.$$

Vectors

- "$\mathbf{x}$ and $\mathbf{y}$ are orthogonal ($\mathbf{x} \perp \mathbf{y}$)" $\quad \Leftrightarrow \quad \mathbf{x}^T\mathbf{y} = 0$.
- The norm (length) of $\mathbf{x}$:

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}}$$

.

- The distance between 2 vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$:

$$d(\mathbf{x}, \mathbf{y})^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x}^T\mathbf{y}$$

Matrices

n Equations with m unknows $x_1, \ldots, x_m$:

$$\left\{ \begin{array}{ccc} a_{11}x_1 + & \ldots & + a_{1m}x_m = b_1 \\ & \vdots & \\ a_{n1}x_1 + & \ldots & + a_{nm}x_m = b_n \end{array} \right. \Leftrightarrow$$

$$\begin{pmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \Leftrightarrow$$

$$\mathbf{A}_{n\times m}\mathbf{x}_{m\times 1} = \mathbf{b}_{n\times 1}.$$

Matrices

n Equations with m unknows $x_1, \ldots, x_m$:

$$\left\{ \begin{array}{ccc} a_{11}x_1 + & \ldots & + a_{1m}x_m = b_1 \\ & \vdots & \\ a_{n1}x_1 + & \ldots & + a_{nm}x_m = b_n \end{array} \right. \Leftrightarrow$$

$$\left[\begin{array}{ccc} (a_{11}, & \ldots, & a_{1m}) \\ & \vdots & \\ (a_{n1}, & \ldots, & a_{nm}) \end{array} \right] \left(\begin{array}{c} x_1 \\ \vdots \\ x_m \end{array} \right) = \left(\begin{array}{c} b_1 \\ \vdots \\ b_n \end{array} \right) \Leftrightarrow$$

$$\mathbf{Ax} = \mathbf{b}.$$

Matrices

n Equations with m unknows $x_1, \dots, x_m$:

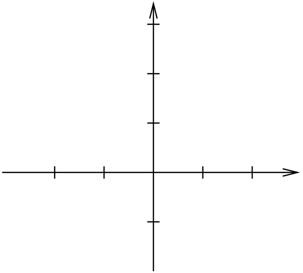
$$\left\{ \begin{array}{ccc} a_{11}x_1 + & \dots & + a_{1m}x_m = b_1 \\ & \vdots & \\ a_{n1}x_1 + & \dots & + a_{nm}x_m = b_n \end{array} \right. \Leftrightarrow$$

$$\begin{bmatrix} (a_{11}, & \dots, & a_{1m}) \\ & \vdots & \\ (a_{n1}, & \dots, & a_{nm}) \end{bmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \Leftrightarrow$$

$$\mathbf{A}\mathbf{x} = \mathbf{b}.$$

2-D Example

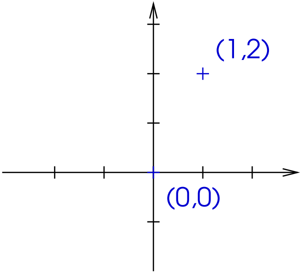
$$\begin{cases} 2x_1 - x_2 = 0 \\ x_1 + 3x_2 = 2 \end{cases}$$



Matrices
Geometrical view

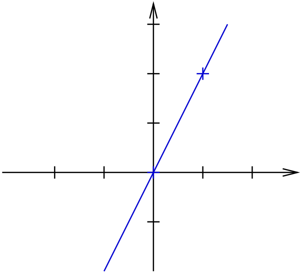
2-D Example

$$\begin{cases} 2x_1 & - x_2 & = 0 \\ x_1 & + 3x_2 & = 2 \end{cases}$$



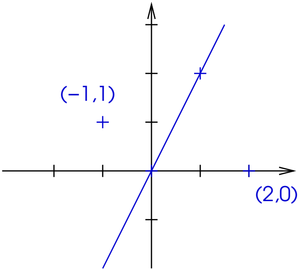
2-D Example

$$\begin{cases} 2x_1 & - x_2 & = 0 \\ x_1 & + 3x_2 & = 2 \end{cases}$$



2-D Example

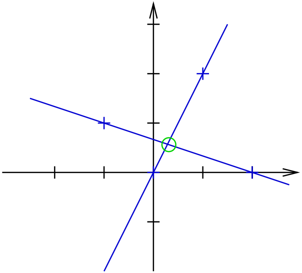
$$\begin{cases} 2x_1 & - x_2 & = 0 \\ x_1 & + 3x_2 & = 2 \end{cases}$$



Matrices
Geometrical view

2-D Example

$$\begin{cases} 2x_1 - x_2 = 0 \\ x_1 + 3x_2 = 2 \end{cases}$$



Matrices

n Equations with m unknows $x_1, \ldots, x_m$:

$$\mathbf{Ax} = \mathbf{b} \Leftrightarrow$$

$$\begin{pmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \Leftrightarrow$$

Matrices

n Equations with *m* unknows $x_1, \ldots, x_m$:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \Leftrightarrow$$

$$\begin{pmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \Leftrightarrow$$

$$x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} + \ldots + x_m \begin{pmatrix} a_{1m} \\ \vdots \\ a_{nm} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}.$$

A real valued matrix $\mathbf{A}_{n \times m}$ is also seen as a linear transformation:

$$\mathbf{A} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$$

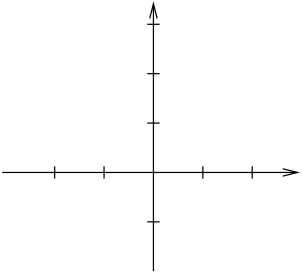
$$\mathbf{x} \longrightarrow \mathbf{A}\mathbf{x}$$

.

Matrices
Alternate geometrical view

2-D Example

$$\begin{cases} 2x_1 & -x_2 & = 0 \\ x_1 & +3x_2 & = 2 \end{cases} \Leftrightarrow x_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

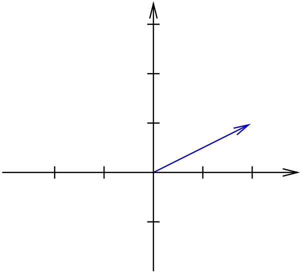


Matrices

Alternate geometrical view

2-D Example

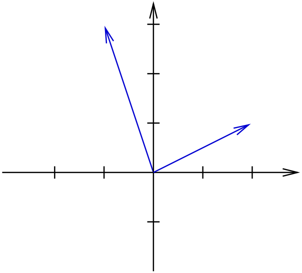
$$\begin{cases} 2x_1 & -x_2 & = 0 \\ x_1 & +3x_2 & = 2 \end{cases} \Leftrightarrow x_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$



Matrices
Alternate geometrical view

2-D Example

$$\left\{\begin{array}{ccc} 2x_1 & -x_2 & = 0 \\ x_1 & +3x_2 & = 2 \end{array}\right. \Leftrightarrow x_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

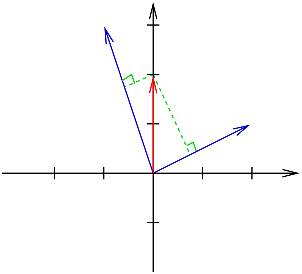


Matrices

Alternate geometrical view

2-D Example

$$\left\{ \begin{array}{rrl} 2x_1 & -x_2 & = \color{red}{0} \\ x_1 & +3x_2 & = \color{red}{2} \end{array} \right. \Leftrightarrow x_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} -1 \\ 3 \end{pmatrix} = \begin{pmatrix} \color{red}{0} \\ \color{red}{2} \end{pmatrix}.$$

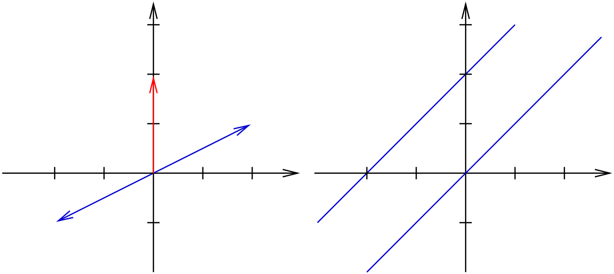


Matrices

Alternate geometrical view (No solution)

2-D Example

$$\begin{cases} 2x_1 & -2x_2 & = 0 \\ x_1 & -x_2 & = 2 \end{cases} \Leftrightarrow x_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$



Determinant

Recursive Definition: Let A be a square matrix ($m \times m$),

$$det(A) = \begin{vmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mm} \end{vmatrix} = \sum_{j=1}^{m} (-1)^{1+j} a_{1j} det(M_{1j}),$$

where M_{ij} is A without its line i and its column j and $det(m) = m$ for m scalar.

Example:

$$det(A) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}(a_{22}a_{33} - a_{32}a_{23}) + a_{12}(a_{21}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{31}a_{22})$$

Inverses

- Definition: A square matrix $\mathbf{A}_{m\times m}$ is called non-singular or invertible if there exists a matrix $\mathbf{B}_{m\times m}$ such that:

$$\mathbf{AB} = \mathbf{I}_m = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} = \mathbf{BA}.$$

  If such B exists it is called the inverse of $\mathbf{A}$ and noted $\mathbf{A}^{-1}$.

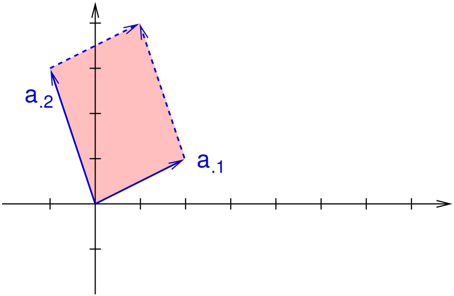
- "$\mathbf{A}$ is invertible" $\Leftrightarrow det(\mathbf{A}) \neq 0 \Leftrightarrow$ "$\mathbf{Ax} = 0$ iff $\mathbf{x} = 0$".
- If $\mathbf{A}$ (square) is invertible, the solution of the system $\mathbf{Ax} = \mathbf{b}$ is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Determinants and Inverses

Geometrical view

2-D Example

$$|det(\mathbf{A})| = |\begin{vmatrix} 2 & -1 \\ 1 & 3 \end{vmatrix}| = |2 \cdot 3 - 1 \cdot (-1)|$$

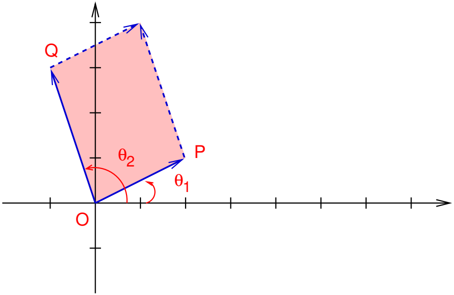


Determinants and Inverses

Geometrical view

2-D Example

$$|det(\mathbf{A})| = |\begin{vmatrix} 2 & -1 \\ 1 & 3 \end{vmatrix}| = OP.OQ.\sin(\theta_2 - \theta_1).$$



Matrices

- If $\mathbf{A}$ is rectangular and $\mathbf{A}^T\mathbf{A}$ is invertible, the solution of the system $\mathbf{Ax} = \mathbf{b}$ is $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.
- $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is called the pseudo-inverse of $\mathbf{A}$.
- Let $\mathbf{X}_{n\times m} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$ be a collection of examples.
- The Gram matrix of this collection is:

$$\mathbf{G} = \mathbf{XX}^T = \begin{pmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \dots & \mathbf{x}_1^T\mathbf{x}_n \\ \vdots & \ddots & \vdots \\ \mathbf{x}_n^T\mathbf{x}_1 & \dots & \mathbf{x}_n^T\mathbf{x}_n \end{pmatrix}.$$

- A real valued squared matrix $\mathbf{A}$ is said to be positive semidefinite if for any vector $\mathbf{z}$: $\mathbf{z}^T\mathbf{Az} \geq 0$.
- Gram matrices are positive semidefinite matrices.

Matrix Diagonalization

- An eigenvector $\mathbf{u}$ of $\mathbf{A}$ (square matrix) is a solution ($\neq 0$) of the equation: $\mathbf{A}u = \lambda\mathbf{u} \Leftrightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = 0$, for a particular λ called the associated eigenvalue.
- Eigenvalues are solution of the characteristic polynomial: $det(\mathbf{A} - \lambda\mathbf{I}) = 0$.
- If $\mathbf{A}_{n\times n}$ is real valued and symmetric then:
  - all eigenvalues $\lambda_1, \dots, \lambda_n$ are real valued and
  - we can find n eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_n$ such that $\mathbf{u}_i \perp \mathbf{u}_j$ and $\|\mathbf{u}_j\| = 1$, *ie* a new basis for $\mathbb{R}^n$.
- If $\mathbf{P} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$, then $\mathbf{A}$ can be rewritten as:

$$\mathbf{A} = \mathbf{P} \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix} \mathbf{P}^T.$$

- "$\mathbf{A}$ positive semidefinite" $\Leftrightarrow \lambda_i \geq 0$ for all i.

Singular Value Decomposition

- The Singular Value Decomposition is a generalization of matrix diagonalization for rectangular matrices.
- Any real valued matrix $\mathbf{M}_{n \times m}$ can be rewritten as:

$$\mathbf{M} = \mathbf{U}_{n \times n} \boldsymbol{\Sigma}_{n \times m} \mathbf{V}^T_{m \times m}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\sigma_{ij} = 0$ unless $i = j$.

Acknowledgement

- Sources of inspiration:
- Linear Algebra: Gilbert Strang MIT course and "Elementary Linear Algebra" Keith Matthews (both on the web).
- Some of the motivating figures: Christopher M. Bishop's book "Pattern Recognition and Machine Learning".