UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Soft Computing Research Group

Mathematical foundations for Machine Learning (*bare essentials*)

Lluís A. Belanche
Computer Science Department
belanche@cs.upc.edu

Big Data Management and Analytics

February 2, 2017

OUTLINE

A matrix $\mathbf{A}$ is a rectangular array of numbers with M rows and N columns ("dimensions") written $\mathbf{A}_{M \times N}$

Example

$$\mathbf{A}_{3\times 2} = \begin{bmatrix} 3 & -4 \\ 5 & 0 \\ 1 & 2 \end{bmatrix}$$

is a 3×2 matrix, since it consists of 3 rows and 2 columns

The (i, j) element of a matrix $\mathbf{A}$ is denoted by a_{ij} and is located in the i-th row and j-th column ($e.g.$, $a_{22} = 0$)

- A matrix $\mathbf{A}$ is called a **diagonal** matrix if the only non-zero elements of $\mathbf{A}$ are in the a_{ii} positions. For example,

$$\mathbf{A} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

  is a diagonal matrix, sometimes written $diag(3, 2)$
- A matrix $\mathbf{A}$ is said to be **symmetric** if $\mathbf{A}^\top = \mathbf{A}$
- A diagonal matrix whose diagonal entries are all 1 is called an **identity** matrix. It is usually denoted by the symbol $\mathbf{I}$

- A matrix $\mathbf{A}$ is said to have an inverse (or to be invertible) if

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}$$

for some matrix $\mathbf{B}$, called the inverse of $\mathbf{A}$ and

$$\mathbf{B} = \mathbf{A}^{-1}$$

Example

$$\mathbf{A} = \begin{bmatrix} 3 & -4 \\ 1 & 2 \end{bmatrix},$$

then

$$\mathbf{A}^{-1} = \begin{bmatrix} 0.2 & 0.4 \\ -0.1 & 0.3 \end{bmatrix}.$$

- A matrix $\mathbf{A}$ is said to be **orthogonal** if $\mathbf{A}^T = \mathbf{A}^{-1}$
- The columns of an orthogonal matrix $\mathbf{A}$ must all be unit vectors and must be mutually orthogonal

Matrix operations

- Matrices are associative and commutative under the addition operation:

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$$

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

- Matrices are associative under the multiplication operation, but not commutative in general

$$\mathbf{A}(\mathbf{B}\mathbf{C}) = (\mathbf{A}\mathbf{B})\mathbf{C}$$

$$\mathbf{A}\mathbf{B} \neq \mathbf{B}\mathbf{A}$$

Matrix multiplication

Let $\mathbf{A}_{M \times N}$ and $\mathbf{B}_{N \times M}$ matrices, so that the product matrices $\mathbf{AB}$ and $\mathbf{BA}$ are both defined:

- If $\mathbf{C}_{M \times M} = \mathbf{AB}$, then

$$[c_{ij}] = \sum_{k=1}^{N} a_{ik} b_{kj}$$

- If $\mathbf{C}_{N \times N} = \mathbf{BA}$, then

$$[c_{ij}] = \sum_{k=1}^{M} b_{ik} a_{kj}$$

Trace of a square matrix $\mathbf{A}_{N \times N}$

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{N} a_{ii}$$

- $\mathrm{Tr}(\mathbf{A} + \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B})$
- $\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^\top)$
- Let $\mathbf{A}_{M \times N}$ and $\mathbf{B}_{N \times M}$ matrices, so that the product matrices $\mathbf{AB}$ and $\mathbf{BA}$ are both defined:

$$\mathrm{Tr}(\mathbf{AB}) = \sum_{i=1}^{M} \left(\sum_{k=1}^{N} a_{ik} b_{ki} \right) = \sum_{k=1}^{N} \left(\sum_{i=1}^{M} b_{ki} a_{ik} \right) = \mathrm{Tr}(\mathbf{BA})$$

Vectors and matrices

In summary ...

- $\mathbf{a} = (a_1, \ldots, a_M)^\top$ is a column vector, a_i is a scalar, $1 \leq i \leq M$
- $\mathbf{A}_{M\times N} = [a_{ij}]$ is a matrix $\mathbf{A} = [\mathbf{a}_1; \ldots; \mathbf{a}_N]$
- Transpose: $\mathbf{A}^\top = [a_{ji}]$; note $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$ holds
- Multiplication: $\mathbf{AB} \neq \mathbf{BA}$, $\mathbf{A}, \mathbf{B}$ must be conformal
- Inverse of a *square* matrix: $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ (note $\mathbf{A}^{-1}$ may not exist)
- $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$

Matrix-vector multiplication

Consider a linear transformation $T\colon \mathbb{R}^N \to \mathbb{R}^M$ satisfying

$$T(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha T(\mathbf{x}) + \beta T(\mathbf{y})$$

Canonical basis

Take a basis $\mathbf{e}_1, \ldots, \mathbf{e}_N$, where $\mathbf{e}_i$ is the (column) N-vector having 0 everywhere except the i-th coordinate that is 1

Assume now that:

$$T(\mathbf{e}_i) = \begin{pmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{Mi} \end{pmatrix}$$

Matrix-vector multiplication

Note that:

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix} = c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_N$$

and therefore

$$T\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix} = c_1 T(\mathbf{e}_1) + \cdots + c_N T(\mathbf{e}_N) = c_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{M1} \end{pmatrix} + \cdots + c_N \begin{pmatrix} a_{1N} \\ a_{2N} \\ \vdots \\ a_{MN} \end{pmatrix}$$

Matrix-vector multiplication

It is useful to "gather" all the involved numbers in a matrix (a rectangular array of the numbers of the transformation T):

$$\mathbf{A}_{M\times N} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{pmatrix}$$

For an arbitrary N-vector $\mathbf{x} = (x_1, \ldots, x_N)^\top$, $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$:

$$T(\mathbf{x}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1N}x_N \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2N}x_N \\ \vdots \\ a_{M1}x_1 + a_{M2}x_2 + \cdots + a_{MN}x_N \end{pmatrix}$$

Matrix-vector multiplication

What happens when we have two linear transformations $T\colon \mathbb{R}^N \to \mathbb{R}^M$ and $S\colon \mathbb{R}^P \to \mathbb{R}^N$?

1. T corresponds to a certain matrix $\mathbf{A}_{M\times N}$
2. S corresponds to a certain matrix $\mathbf{B}_{N\times P}$
3. If we form the composition $T \circ S : \mathbb{R}^P \to \mathbb{R}^M$, then:
    1. $T \circ S$ is again a linear transformation
    2. its corresponding $M \times P$ matrix is $\mathbf{AB}$

Linear Vector Spaces

Let V be a set on which two operations, addition $(+)$ and scalar multiplication, have been defined

Axioms of a Vector Space

1. $\mathbf{u} + \mathbf{v} \in V$ (closure under addition)
2. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ (commutativity)
3. $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$ (associativity)
4. $\exists \mathbf{0} \in V$, called a **zero vector**, s.t. $\mathbf{u}+\mathbf{0}=\mathbf{u}$
5. $\forall \mathbf{u} \in V$, there is a $-\mathbf{u} \in V$ s.t. $\mathbf{u} + (-\mathbf{u})=\mathbf{0}$
6. $\alpha\mathbf{u} \in V$ (closure under scalar multiplication)
7. $\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$ (distributivity)
8. $(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{u}$ (distributivity)
9. $\alpha(\beta\mathbf{u}) = (\alpha\beta)\mathbf{u}$
10. $1\mathbf{u} = \mathbf{u}$

If these axioms hold $\forall \mathbf{u,v,w} \in V$ and $\forall \alpha, \beta \in \mathbb{K}$, then V is called a *vector space* over the field $\mathbb{K}$ (its elements are called vectors)

Linear Vector Spaces

- A vector $\mathbf{v}$ is a **linear combination** of $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k$ if there are scalar coefficients $c_1, c_2, ..., c_k$ s.t. $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_k\mathbf{v}_k = \mathbf{v}$
- A set of vectors $\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_k$ is **linearly dependent** if there are scalars $c_1, c_2, ..., c_k$ (at least one of which is $\neq 0$) s.t. $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + ...c_k\mathbf{v}_k = 0$ (a set of vectors not linearly dependent is **linearly independent**)
- The **rank** of a matrix is the maximum number of linearly independent row (or column) vectors
- A matrix $\mathbf{A}_{M\times N}$ is said to be **full rank** when its rank is $\min(N, M)$
- A square matrix $\mathbf{A}_{N\times N}$ is **non-singular** (it admits $\mathbf{A}^{-1}$) iff it is full rank

Example

Let $\mathbf{u} = (1, 0, 3), \mathbf{v} = (-1, 1, -3), \mathbf{w} = (1, 2, 3)$
$3(1, 0, 3) + 2(-1, 1, -3) - (1, 2, 3) = 0$
$\mathbf{u}, \mathbf{v}$, and $\mathbf{w}$ are linearly dependent, since $3\mathbf{u} + 2\mathbf{v} - \mathbf{w} = 0$

The vector space $\mathbb{R}^N$

- A *point* in the space $\mathbb{R}^N$ may be represented as a *vector*:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{pmatrix} = (x_1, x_2, \cdots, x_N)^\top$$

- Two vectors $\mathbf{x}$ and $\mathbf{y}$ are equal iff $x_i = y_i$ for all $i = 1, \cdots, N$
- A vector $\mathbf{x}$ can be multiplied by a real scalar α to become

$$\alpha\mathbf{x} = [\alpha x_1, \alpha x_2, \cdots, \alpha x_N]^\top$$

- The *sum* of two vectors is defined as

$$\mathbf{x} + \mathbf{y} = [x_1 + y_1, x_2 + y_2, \cdots, x_N + y_N]^\top$$

Note $\mathbf{x} - \mathbf{y} = \mathbf{x} + (-\mathbf{y}) = [x_1 - y_1, x_2 - y_2, \cdots, x_N - y_N]^\top$

Linear Vector Spaces

Span

If $\{v_1, v_2, ..., v_k\}$ is a set of vectors in a vector space V, then the set of all their linear combinations is called the **span**

Basis

Any $\{v_1, v_2, ..., v_k\}$ set of vectors in V whose span is V is called a **basis** for V

The Basis Theorem

If a vector space V has a basis with N vectors, then every basis for V has exactly N vectors

Dimension

A vector space V is called *finite-dimensional* if it has a basis consisting of finitely many vectors N (N is called the **dimension** of V). In any other case it is *infinite-dimensional*

Linear Vector Spaces

- A set of *basis vectors* (or a **basis**) $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N\}$ can be obtained as:

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, \cdots, 0)^\top \\ \mathbf{e}_2 &= (0, 1, \cdots, 0)^\top \\ &\cdots\cdots \\ \mathbf{e}_N &= (0, \cdots, 0, 1)^\top \end{aligned}$$

- Given a set of basis vectors that span the space, any vector $\mathbf{x}$ in the space can be expressed as a linear combination (weighted sum) of these basis vectors:

$$\mathbf{x} = \sum_{i=1}^{N} x_i \mathbf{e}_i$$

with x_i being the coefficient for the i-th basis vector $\mathbf{e}_i$

Linear Vector Spaces

- The **inner product** or **dot product** of two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$ is a scalar defined as:

$$\mathbf{x} \cdot \mathbf{y} := \mathbf{x}^\top \mathbf{y} = (x_1, x_2, \cdots, x_N) \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix} = \sum_{i=1}^{N} x_i y_i = \mathbf{y}^\top \mathbf{x}$$

  With this inner product, $\mathbb{R}^N$ is called a *Euclidean space*. Note $\mathbf{x}^\top \mathbf{e}_i = x_i$

- Two vectors $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal** (meaning perpendicular) if their inner product is zero:

$$\mathbf{x} \cdot \mathbf{y} = 0$$

- The **2-norm** (or length) of a vector $\mathbf{x}$ is defined as

$$||\mathbf{x}|| := \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\sum_{i=1}^{N} x_i^2} \geq 0$$

Linear Vector Spaces

- If $||\mathbf{x}|| = 1$, then the vector $\mathbf{x}$ is a **normalized** (or unit) vector. This can be accomplished as $\mathbf{x}/||\mathbf{x}||$
- The **distance** between two points $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$d(\mathbf{x}, \mathbf{y}) := ||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

- A (metric) distance has the following properties:
  - $d(\mathbf{x}, \mathbf{y}) \geq 0$ and $d(\mathbf{x}, \mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$
  - $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
  - triangle inequality: $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$

Linear Vector Spaces

The **angle** between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$\theta = \cos^{-1}\left(\frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||\ ||\mathbf{y}||}\right)$$

In particular, if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, i.e., $\mathbf{x} \cdot \mathbf{y} = 0$, then the angle between them is $cos^{-1}(0) = \pi/2$ or 90 degrees
Therefore, this inner product can also be obtained as

$$\mathbf{x} \cdot \mathbf{y} = ||\mathbf{x}||\ ||\mathbf{y}|| \cos\theta$$

Cauchy-Schwarz inequality

Taking absolute value on both sides of the above, we get:

$$|\mathbf{x} \cdot \mathbf{y}| = ||\mathbf{x}||\ ||\mathbf{y}|| |\cos\theta| \leq ||\mathbf{x}||\ ||\mathbf{y}||$$

Squaring both sides, we get $|\mathbf{x} \cdot \mathbf{y}|^2 \leq ||\mathbf{x}||^2\ ||\mathbf{y}||^2$ or

$$|\mathbf{x} \cdot \mathbf{y}| \leq \sqrt{(\mathbf{x} \cdot \mathbf{x})(\mathbf{y} \cdot \mathbf{y})}$$

Inner Product Spaces

An inner product on a real vector space V is a real-valued function $\langle \mathbf{x}, \mathbf{y} \rangle$ of two vectors $\mathbf{x}, \mathbf{y} \in V$, satisfying the following conditions:

- **Positive definite**:
$$\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$$
and
$$\langle \mathbf{x}, \mathbf{x} \rangle = 0 \quad \text{iff} \quad \mathbf{x} = \mathbf{0}$$
- **Symmetry:**
$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$$
- **linearity:**
$$\langle \alpha \mathbf{x} + \beta \mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle$$

A vector space with an inner product defined on it is called an *inner product space*. Example: $\mathbb{R}^N$ with $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$

Some useful derivatives

Some students may not have seen or remember *partial derivatives*. For example, let $f(x, y) = xy^2$; then

$$\frac{\partial f}{\partial x} = y^2$$
$$\frac{\partial f}{\partial x} = 2xy$$

- The ∂ symbol means treating all other variables as if they were constants for the differentiation
- To express all partial derivatives of a function:

$$\nabla_{\mathbf{x}} f := \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \cdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Some useful derivatives

Let $\mathbf{a}, \mathbf{x}$ be two N-vectors; then:

$$\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial ||\mathbf{x}||^2}{\partial \mathbf{x}} = 2\mathbf{x}$$

Some useful derivatives

Let $\mathbf{A}_{M \times N}$ be a matrix not depending on an N-vector $\mathbf{x}$ and an M-vector $\mathbf{y}$; then:

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}$$

$$\frac{\partial \mathbf{y}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{y}^\top \mathbf{A}$$

$$\frac{\partial \mathbf{y}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{y}} = \mathbf{x}^\top \mathbf{A}^\top$$

Some useful derivatives

For the special case in which $\mathbf{x} = \mathbf{y}$ and $N = M$, we get a **quadratic form**

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i a_{ij} x_j$$

$$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$$

What do we get for the special case where $\mathbf{A}$ is a symmetric matrix? $2\mathbf{A}\mathbf{x}$

Some useful derivatives

For the special case in which $\mathbf{x} = \mathbf{y}$ and $N = M$, we get a **quadratic form**

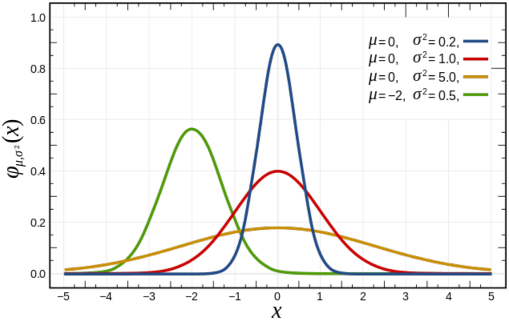
$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i a_{ij} x_j$$

$$\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$$

What do we get for the special case where $\mathbf{A}$ is a symmetric matrix? $2\mathbf{A}\mathbf{x}$

A continuous random variable X is **normally distributed**, written $X \sim \mathcal{N}(x; \mu, \sigma^2)$, when its pdf is:

- $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$
- $\mathbb{E}(X) = \mu$, $Var(X) = \sigma^2$



A continuous d-variate random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ is **normally distributed**, written $\mathbf{X} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$, when its joint pdf is:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

where $\boldsymbol{\mu}$ is the *mean vector* and $\Sigma_{d \times d} = (\sigma_{ij}^2)$ is the (real symmetric and p.d.) *covariance matrix*.

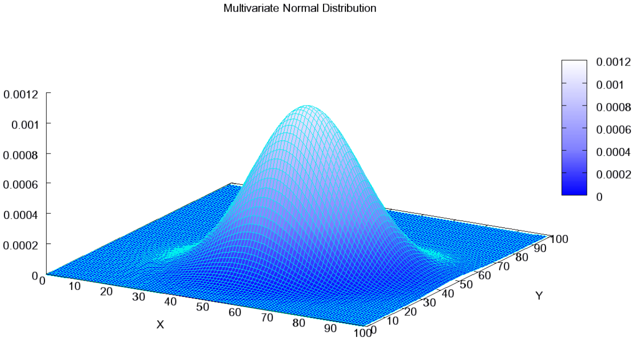
- $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ and $\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] = \Sigma.$
- $CoVar[X_i, X_j] = \sigma_{ij}^2$ and $Var[X_i] = \sigma_{ii}^2$

---

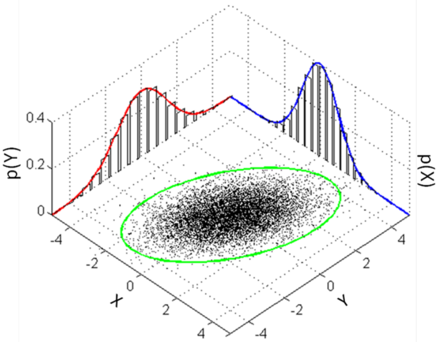
if $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, then X_i, X_j are independent $\iff$ $CoVar[X_i, X_j] = 0$

(in general, only the left-to-right implication holds)

The Gaussian Distribution ($d = 2$)

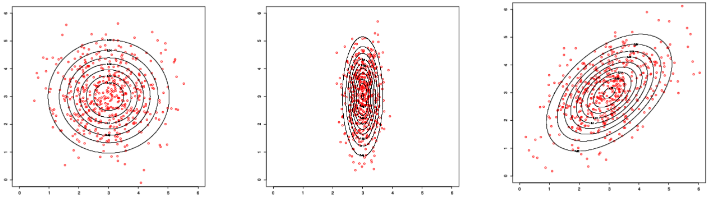


The Gaussian Distribution ($d = 2$)



Observations from a bivariate normal distribution, a contour ellipsoid, the two marginal distributions, and their histograms (images from the Wikipedia)

The Gaussian Distribution ($d = 2$)



$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

- The principal directions (a.k.a. PCs) of the hyperellipsoids are given by the *eigenvectors* $\mathbf{u}_i$ of Σ, which satisfy $\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i$.
- The lengths of the hyperellipsoids along these axes are proportional to $\sqrt{\lambda_i}$ (note $\lambda_i > 0$)

Conceptual view

- What is behind the choice of a **multivariate Gaussian**?

  Examples from a class are noisy versions of an ideal class member (a *prototype*):

  - Prototype: modeled by the mean vector
  - Noise: modeled by the covariance matrix

- The quantity

$$d(\mathbf{x}) := \sqrt{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

  is called the **Mahalanobis distance** for $\mathbf{x}$

- Very important! the number of parameters is $\frac{d(d+1)}{2} + d$

Eigenvalues and eigenvectors of an $N \times N$ matrix $\mathbf{A}$

Eigenvector

A vector $\mathbf{x}$ in $\mathbb{R}^N$ is called an **eigenvector** of the matrix $\mathbf{A}$ if $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{A}\mathbf{x}$ is a scalar multiple of $\mathbf{x}$, that is, if there is a scalar λ (called an *eigenvalue*) s.t. $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$

Theorem: λ is an eigenvalue of $\mathbf{A}$ if and only if

$$\det(\lambda\,\mathbf{I} - \mathbf{A}) = 0$$

Characteristic polynomial

If $\mathbf{A}$ is an $N \times N$ matrix, the expression $\det(\lambda\,\mathbf{I} - \mathbf{A})$ defines a polynomial of degree N in λ, called the *characteristic polynomial* of $\mathbf{A}$ and denoted by $p_{\mathbf{A}}(\lambda)$

Some useful facts:

- $\mathbf{A}$ has N eigenvalues (though some may be repeated). If an eigenvalue λ is repeated k times, we say it has *algebraic multiplicity* k
- Eigenvalues can be real or complex-valued; however, if $\mathbf{A}$ is symmetric, then its eigenvalues are all real, and it will have N linearly independent eigenvectors
- If $\mathbf{A}$ is a triangular matrix, the eigenvalues appear along its diagonal
- The sum of the eigenvalues is equal to $\mathrm{Tr}(\mathbf{A})$
- The product of the eigenvalues is equal to $\det(\mathbf{A})$. Thus, $\mathbf{A}^{-1}$ exists iff all of its eigenvalues are nonzero

Positive definiteness and quadratic forms

Suppose $\mathbf{A}$ is a real symmetric $N \times N$ matrix and $\mathbf{x}$ is a vector of length N; then

$$\mathbf{x}^\top \mathbf{A}\mathbf{x}$$

is a quadratic form: a quadratic polynomial in the elements of $\mathbf{x}$

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix},$$

then

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = 2x_1^2 + 6x_1x_2 + 4x_2^2$$

Positive definiteness and quadratic forms

A symmetric matrix $\mathbf{A}$ is said to be **positive definite** (p.d.) if its quadratic forms in $\mathbf{x}$ are all positive when $\mathbf{x} \neq \mathbf{0}$.
In other words, $\mathbf{A}$ is p.d. if and only if

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$$

whenever $\mathbf{x} \neq \mathbf{0}$

Example (cont.)

Completing the square, $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 2(x_1 + 1.5x_2)^2 - 0.5x_2^2$.
The first term can be 0 by taking $x_1 = -1.5x_2$ for any real x_2, say $x_2 = 1$. This means that at $\mathbf{x} = (-1.5 \;\; 1)^\top$,

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = -0.5(1)^2 = -0.5 < 0$$

Therefore, $\mathbf{A}$ is not p.d.

Gaussian distribution and p.d.

Positive definiteness

For a Gaussian distribution to be well-defined, Σ has to be real symmetric and positive definite (p.d.): for all non-null vectors $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^{\mathrm{T}}\Sigma\mathbf{x} > 0$ must hold true

Examples: are these matrices p.d.?

$$a. \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad b. \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

$$c. \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix} \qquad d. \begin{pmatrix} 1 & 4 \\ \frac{1}{2} & 1 \end{pmatrix}$$

a. YES; b. YES
c. YES; d. NO

Gaussian distribution and p.d.

Positive definiteness

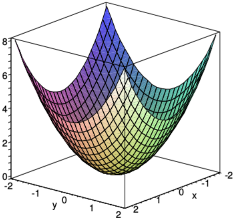
For a Gaussian distribution to be well-defined, Σ has to be real symmetric and positive definite (p.d.): for all non-null vectors $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^{\mathrm{T}}\Sigma\mathbf{x} > 0$ must hold true

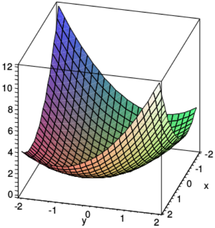
Examples: are these matrices p.d.?

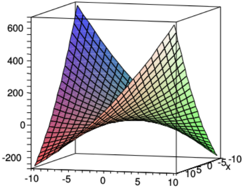
$$a. \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad b. \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

$$c. \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix} \qquad d. \begin{pmatrix} 1 & 4 \\ \frac{1}{2} & 1 \end{pmatrix}$$

a. YES; b. YES
c. YES; d. NO







$a.\ x_1^2 + x_2^2$ $b.\ x_1^2 + x_1 x_2 + x_2^2$ $d.\ x_1^2 + \frac{9}{2} x_1 x_2 + x_2^2$

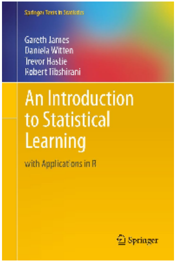
On data pre-processing

Each problem requires a different approach in what concerns data cleaning and preparation. This pre-process is very important because it can have a deep impact on performance; it can easily take you a significant part of the time.

1. treatment of lost values (missing values)
2. treatment of anomalous values (outliers)
3. treatment of incoherent or incorrect values
4. coding of non-continuous or non-ordered variables
5. possible elimination of irrelevant or redundant variables (feature selection)
6. creation of new variables that can be useful (feature extraction)
7. normalization of the variables (*e.g.* standardization)
8. transformation of the variables (*e.g.* correction of serious skewness and/or kurtosis)

Recommended reading: introductory

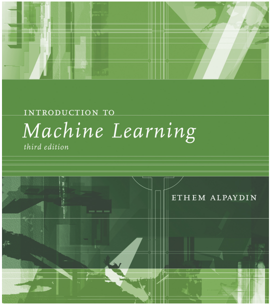
- A free online version of *An Introduction to Statistical Learning, with Applications in R* by James, Witten, Hastie and Tibshirani (Springer, 2013) is available from January 2014
- Springer has agreed –no need to worry about copyright. However, you may not distribute printed versions of the pdf



`http://www-bcf.usc.edu/~gareth/ISL/`

Recommended reading: intermediate

- *Introduction to Machine Learning*, by E. Alpaydin (The MIT Press, 2009)
- There are several editions (the latest, the better)



https://mitpress.mit.edu/books/introduction-machine-learning-0