MACHINE LEARNING

Introduction

Alessandro Moschitti

Department of Computer Science and Information Engineering

University of Trento

Email: moschitti@disi.unitn.it

Course Schedule

- Lectures
  - Tuesday, 14:00-16:00
  - Wednesday, 8:30-10:30
  - Room 107
- Consulting Hours:
  - My office at third floor
  - Thursday at 14:30
  - Sending email is recommended

Lectures

- Introduction to ML
  - Vector spaces
- PAC Learning
  - VC dimension
- Perceptron
  - Vector Space Model
  - Representer Theorem
- Support Vector Machines (SVMs)
  - Hard/Soft Margin (Classification)
  - Regression and ranking

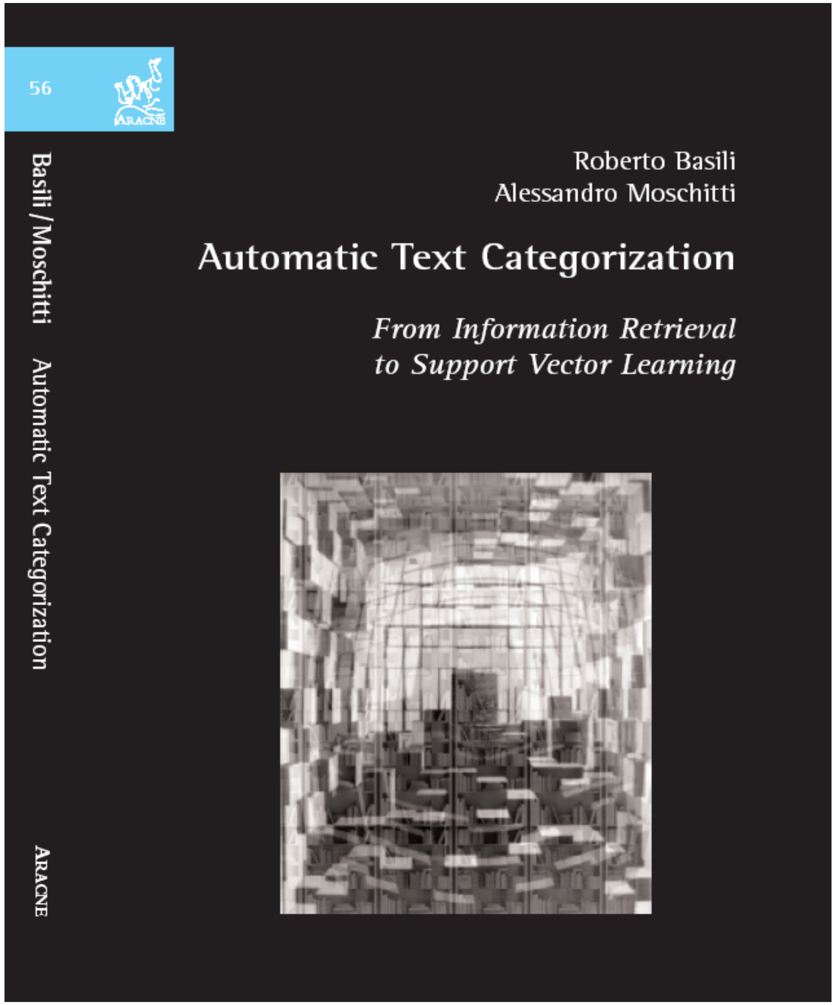
Lectures

- Kernels Methods
  - Theory and Algebraic properties
  - Linear, Polynomial, Gaussian
  - Kernel construction,
- Kernels for structured data
  - Sequence, Tree Kernels
- Structured Output

Lab

- Automated Text Categorization
- Question Classification (Question Answering)

Reference Book + some articles



Today

- Introduction to Machine Learning
- Vector Spaces

Why Learning Functions Automatically?

- Anything is a function
  - From the planet motion
  - To the input/output actions in your computer
- Any problem would be automatically solved

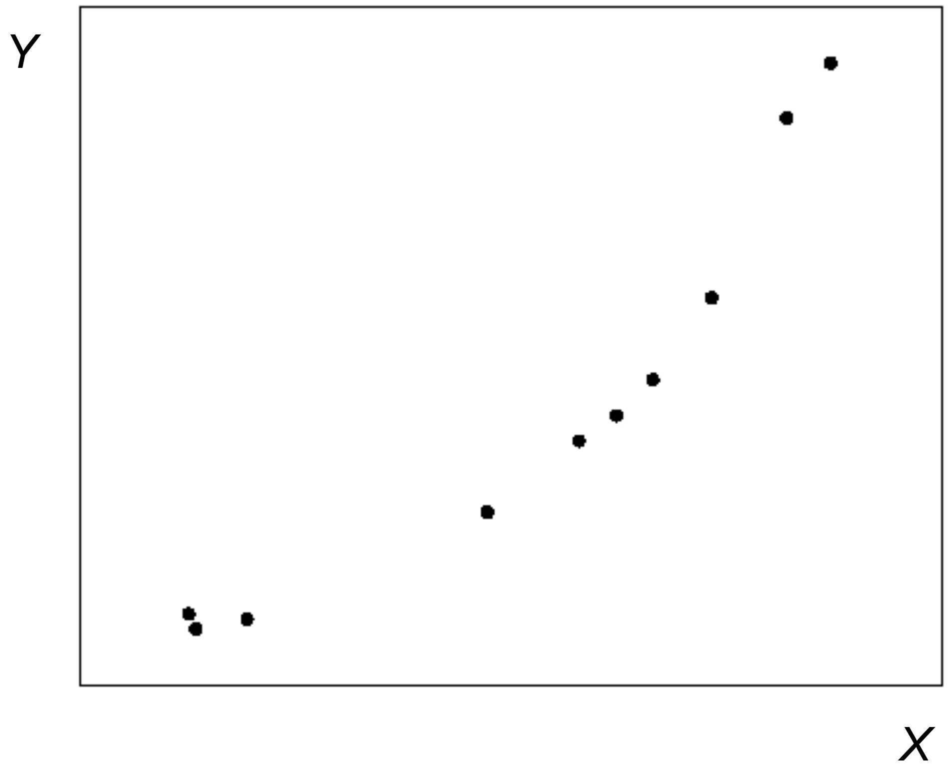
More concretely

- Given the user requirement (input/output relations) we write programs
- Different cases typically handled with *if-then* applied to input variables
- What happens when
  - millions of variables are present and/or
  - values are not reliable (e.g. noisy data)
- Machine learning writes the *program* (rules) for you

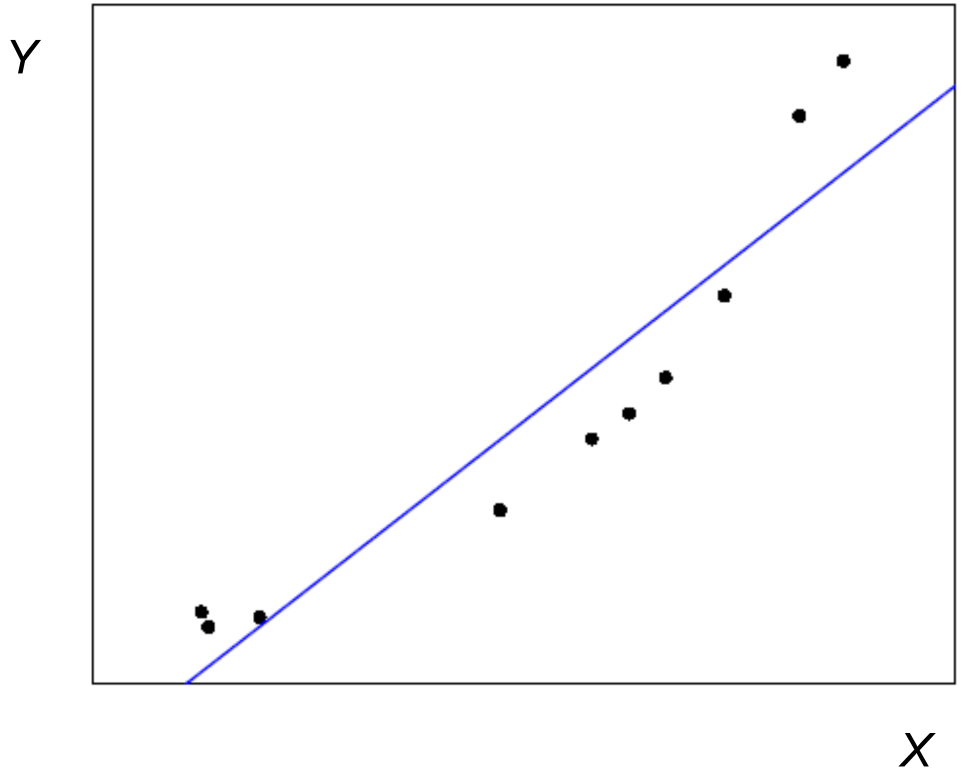
What is Statistical Learning?

- Statistical Methods – Algorithms that learn relations in the data from examples
- Simple relations are expressed by pairs of variables: $\langle x_1,y_1\rangle$, $\langle x_2,y_2\rangle,\ldots,\langle x_n,y_n\rangle$
- Learning *f* such that evaluate y^* given a new value x^*, i.e. $\langle x^*, f(x^*)\rangle = \langle x^*, y^*\rangle$

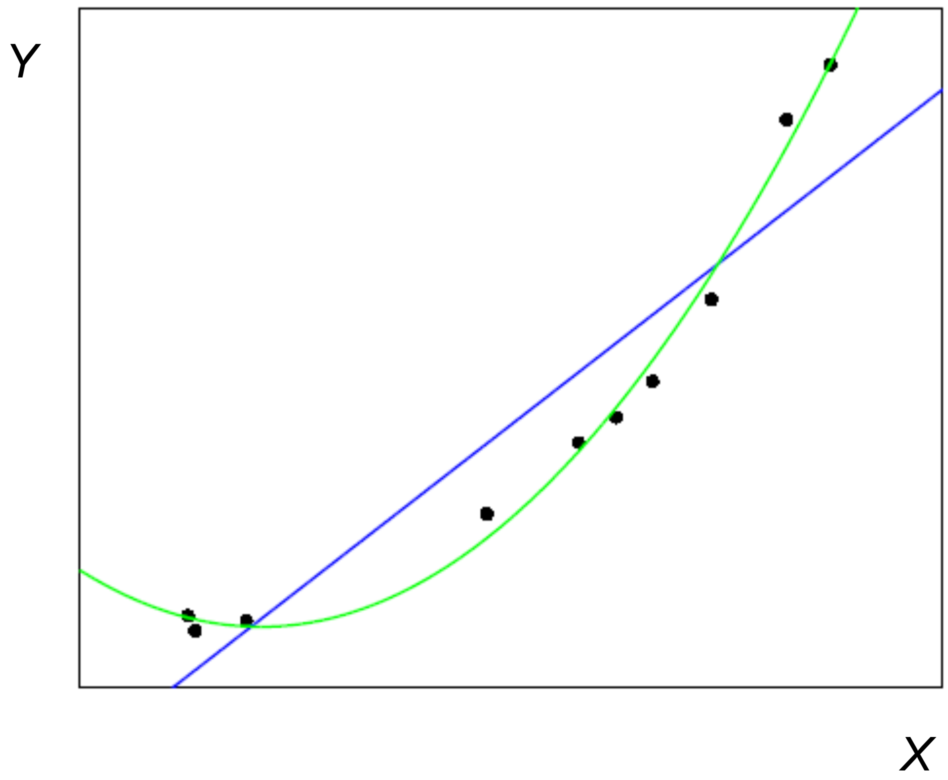
You have already tackled the learning problem



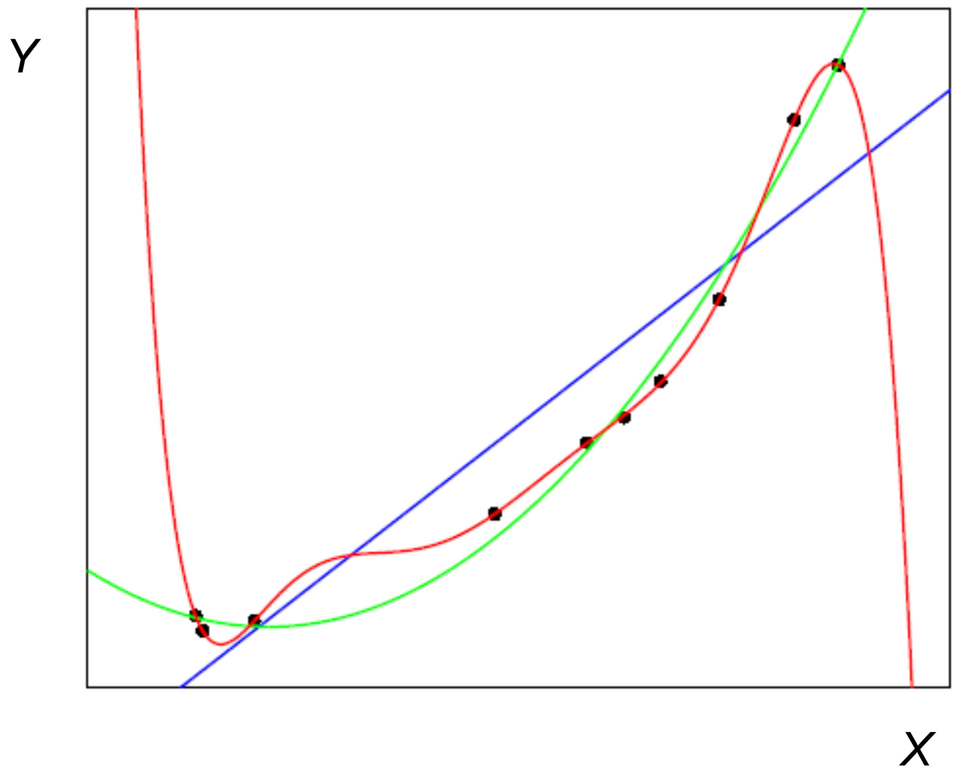
Linear Regression



Degree 2



Degree



Machine Learning Problems

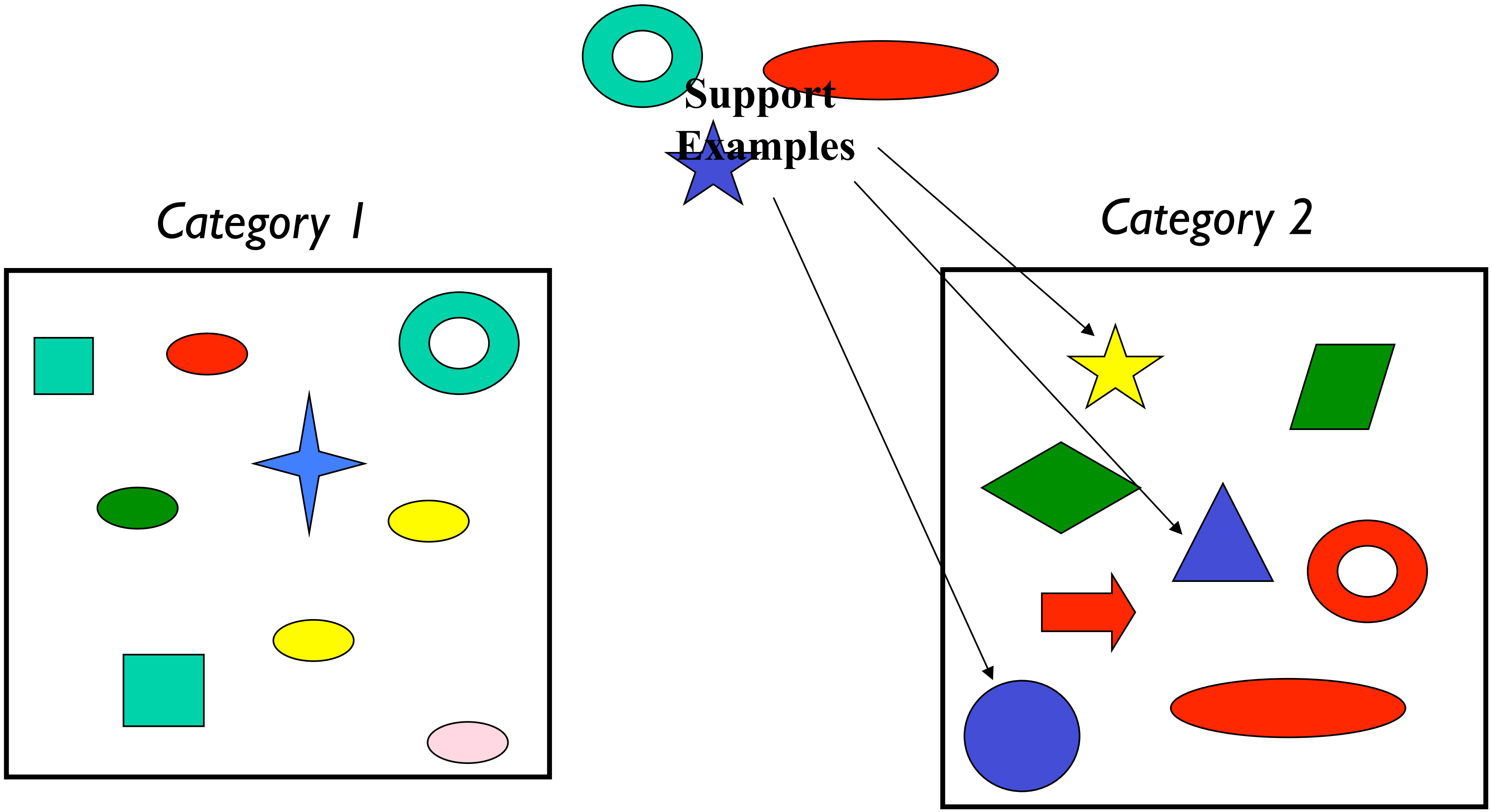
- Overfitting
- How dealing with millions of variables instead of only two?
- How dealing with real world objects instead of real values?

Learning Models

- Real Values: *regression*
- Finite and integer: *classification*
- Binary Classifiers:
  - 2 classes, e.g.

    $f(x) \rightarrow$ {cats,dogs}

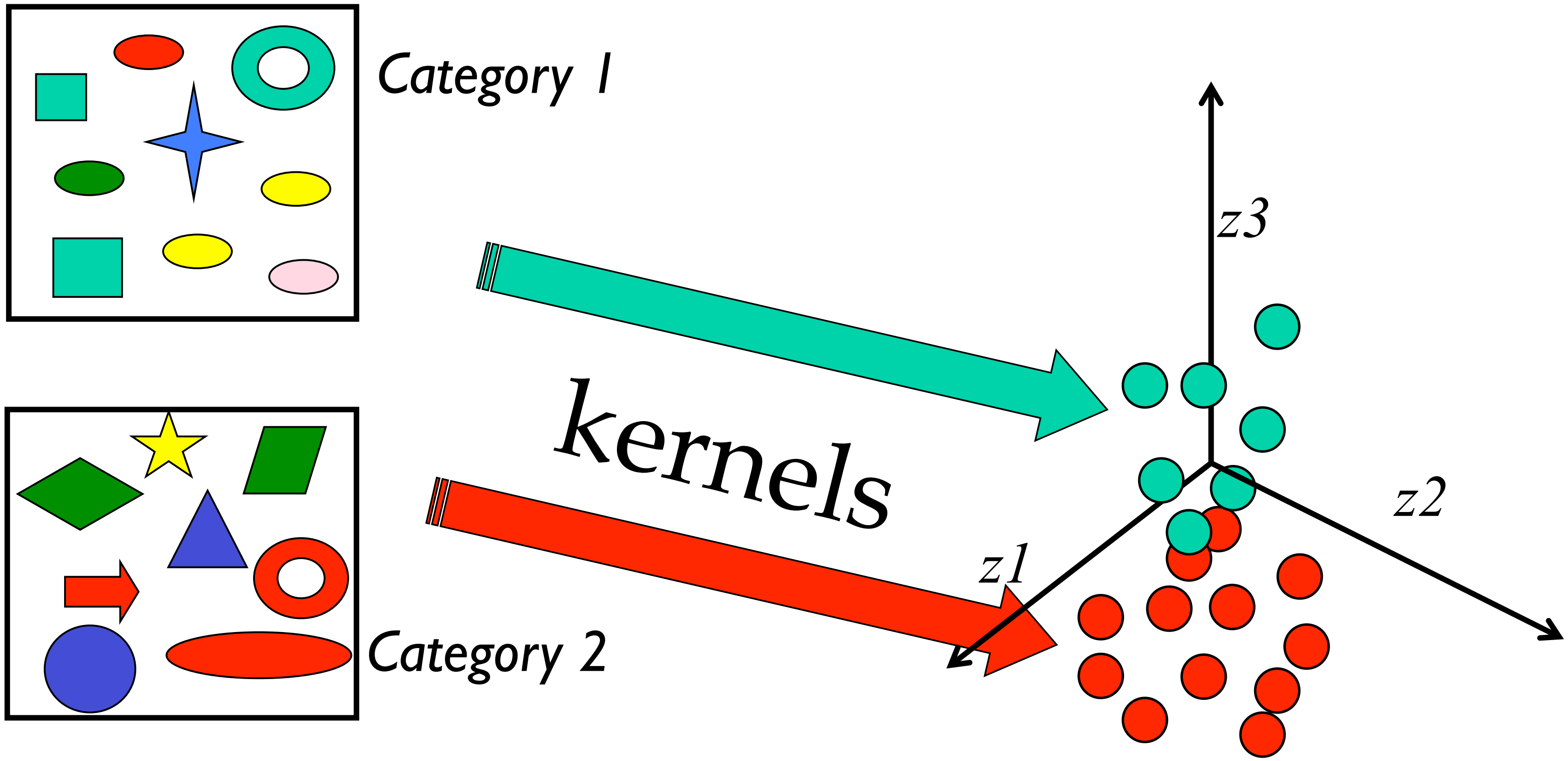
The Idea of Statistical Learning



Similarity in Statistical Learning Theory

- Similarity is intuitively useful to learn classification function
- This does not lead to heuristic models
- In statistical learning theory valid similarities are called ***Kernel Functions***
  - Kernels map examples in vector spaces
  - Examples are classified based on geometric properties
- Formally proved upperbound to the system error
  - Optimize trade-off

In other words



Vector Spaces

Definition (1)

- A set V is a **vector space** over a field F (for example, the field of real or of complex numbers) if, given
- an operation *vector* ***addition*** defined in V, denoted **v** + **w** (where **v**, **w** $\in V$), and
- an operation, *scalar* ***multiplication*** in V, denoted a * **v** (where **v** $\in$ V and $a \in$ F),
- the following properties hold for all a, $b \in$ F and **u**, **v**, and **w** $\in$ V:
- **v** + **w** belongs to V.
  (Closure of V under vector addition)
- **u** + (**v** + **w**) = (**u** + **v**) + **w**
  (Associativity of vector addition in V)
- There exists a neutral element **0** in V, such that for all elements **v** in V, **v** + **0** = **v**
  (Existence of an additive identity element in V)

Definition (2)

- For all **v** in V, there exists an element **w** in V, such that **v** + **w** = **0**
  (Existence of additive inverses in V)
- **v** + **w** = **w** + **v**
  (Commutativity of vector addition in V)
- *a* * **v** belongs to V
  (Closure of V under scalar multiplication)
- *a* * (*b* * **v**) = (*ab*) * **v**
  (Associativity of scalar multiplication in V)
- If 1 denotes the multiplicative identity of the field F, then 1 * **v** = **v**
  (Neutrality of one)
- *a* * (**v** + **w**) = *a* * **v** + *a* * **w**
  (Distributivity with respect to vector addition.)
- (*a* + *b*) * **v** = *a* * **v** + *b* * **v**
  (Distributivity with respect to field addition.)

An example of Vector Space

- For all *n*, $\mathbf{R}^n$ forms a vector space over **R**, with component-wise operations.
- Let **V** be the set of all n-tuples, $[v_1,v_2,v_3,...,v_n]$ where v_i is a member of **R**={real numbers}
- Let the field be **R**, as well
- Define Vector Addition:

  For all v, w, in **V**, define $v+w=[v_1+w_1,v_2+w_2,v_3+w_3,...,v_n+w_n]$
- Define Scalar Multiplication:

  For all a in **F** and v in **V**, $a*v=[a*v_1,a*v_2,a*v_3,...,a*v_n]$
- Then **V** is a Vector Space over **R**.

Linear dependency

- Linear combination:
- $\alpha_1 \mathbf{v}_1 + \ldots + \alpha_n \mathbf{v}_n = 0$ for some $\alpha_1 \ldots \alpha_n$ not all zero

  $\Rightarrow y = \alpha_1 \mathbf{v}_1 + \ldots + \alpha_n \mathbf{v}_n$ has a unique expression
- In case $\alpha_i > 0$ and the sum is 1 it is called convex combination

Normed Vector Spaces

- Given a vector space V over a field K, a norm on V is a function from V to $\mathbf{R}$,
- it associates each vector $\mathbf{v}$ in V with a real number, $||\mathbf{v}||$
- The norm must satisfy the following conditions:
  - For all a in K and all $\mathbf{u}$ and $\mathbf{v}$ in V,

    1. $||\mathbf{v}|| \geq 0$ with equality if and only if $\mathbf{v} = \mathbf{0}$
    2. $||a\mathbf{v}|| = |a| \, ||\mathbf{v}||$
    3. $||\mathbf{u} + \mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$
- A useful consequence of the norm axioms is the inequality
  - $||\mathbf{u} \pm \mathbf{v}|| \geq | \, ||\mathbf{u}|| - ||\mathbf{v}|| \, |$
- for all vectors $\mathbf{u}$ and $\mathbf{v}$

Inner Product Spaces

- Let V be a vector space and **u**, **v**, and **w** be vectors in V and c be a constant.
- Then, an *inner product* (,) on V is
  - a function with domain consisting of pairs of vectors and
  - range real numbers satisfying
  - the following properties:
    1. $(\mathbf{u}, \mathbf{u}) \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{0}$.
    2. $(\mathbf{u}, \mathbf{v}) = (\mathbf{v}, \mathbf{u})$
    3. $(\mathbf{u} + \mathbf{v}, \mathbf{w}) = (\mathbf{u}, \mathbf{w}) + (\mathbf{v}, \mathbf{w})$
    4. $(c\mathbf{u}, \mathbf{v}) = (\mathbf{u}, c\mathbf{v}) = c(\mathbf{u}, \mathbf{v})$

Example

- Let V be the vector space consisting of all continuous functions with the standard + and *. Then define an inner product by

$$(f,g)=\int_0^1 f(t)g(t)dt$$

- For example: $(x,x^2) = \int_0^1 (x)(x^2)dx = \dfrac{1}{4}$

- The four properties follow immediately from the analogous property of the definite integral:

$$(f+g,h)=\int_0^1 (f+g)(t)h(t)\;dt$$

$$=\int_0^1 \left[f(t)h(t)+g(t)h(t)\right]\,dt=\int_0^1 f(t)h(t)\;dt+\int_0^1 g(t)h(t)\;dt$$

$$=(f,h)+(g,h)$$

Inner Product Properties

- $(\mathbf{v}, \mathbf{0}) = 0$
- $\| v \| = \sqrt{(v, v)}$
- If $(\mathbf{v}, \mathbf{u}) = 0$, $\mathbf{v}, \mathbf{u}$ are called orthogonal
- Schwarz Inequality:
  - $[(\mathbf{v}, \mathbf{u})]^2 \leq (\mathbf{v}, \mathbf{v})\,(\mathbf{u}, \mathbf{u})$
- The classical scalar product is the component-wise product
- $(x_1, x_2, \ldots, x_n)\,(y_1, y_2, \ldots, y_n) = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$
- $\cos(u, v) = \dfrac{(u, v)}{\| u \| \cdot \| v \|}$

Projection

- From $\cos(\vec{x}, \vec{w}) = \dfrac{\vec{x} \cdot \vec{w}}{\|\vec{x}\| \cdot \|\vec{w}\|}$

- It follows that

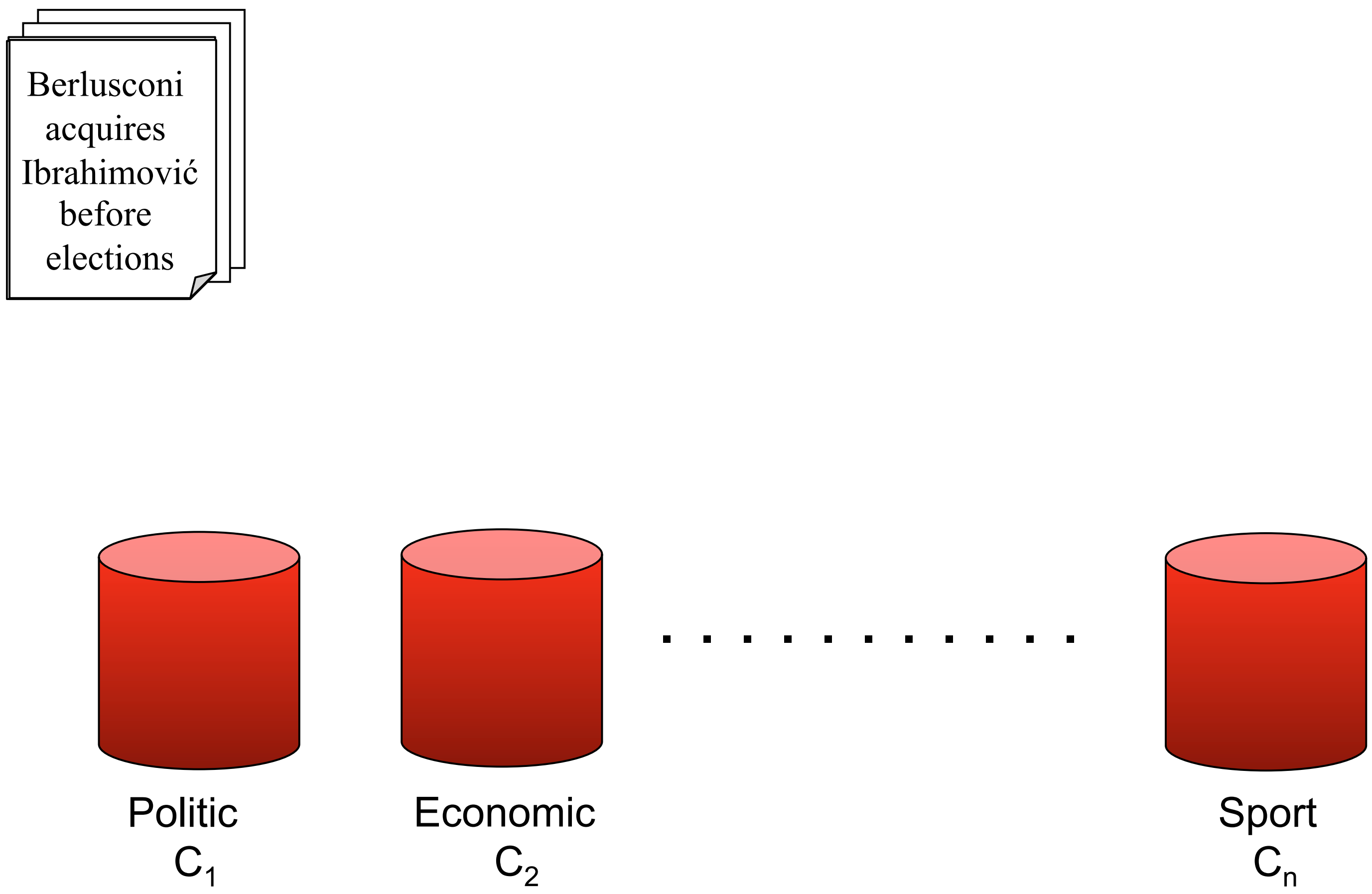
$$\|\vec{x}\| \cos(\vec{x}, \vec{w}) = \frac{\vec{x} \cdot \vec{w}}{\|\vec{w}\|} = \vec{x} \cdot \frac{\vec{w}}{\|\vec{w}\|}$$

- Norm of $\vec{x}$ times the cosine between $\vec{x}$ and $\vec{w}$, i.e. the projection of $\vec{x}$ on $\vec{w}$

Similarity Metrics

- The simplest distance for continuous *m*-dimensional instance space is *Euclidian distance*.
- The simplest distance for *m*-dimensional binary instance space is *Hamming distance* (number of feature values that differ).
- Cosine similarity is typically the most effective

A Simple Example: Text Categorization

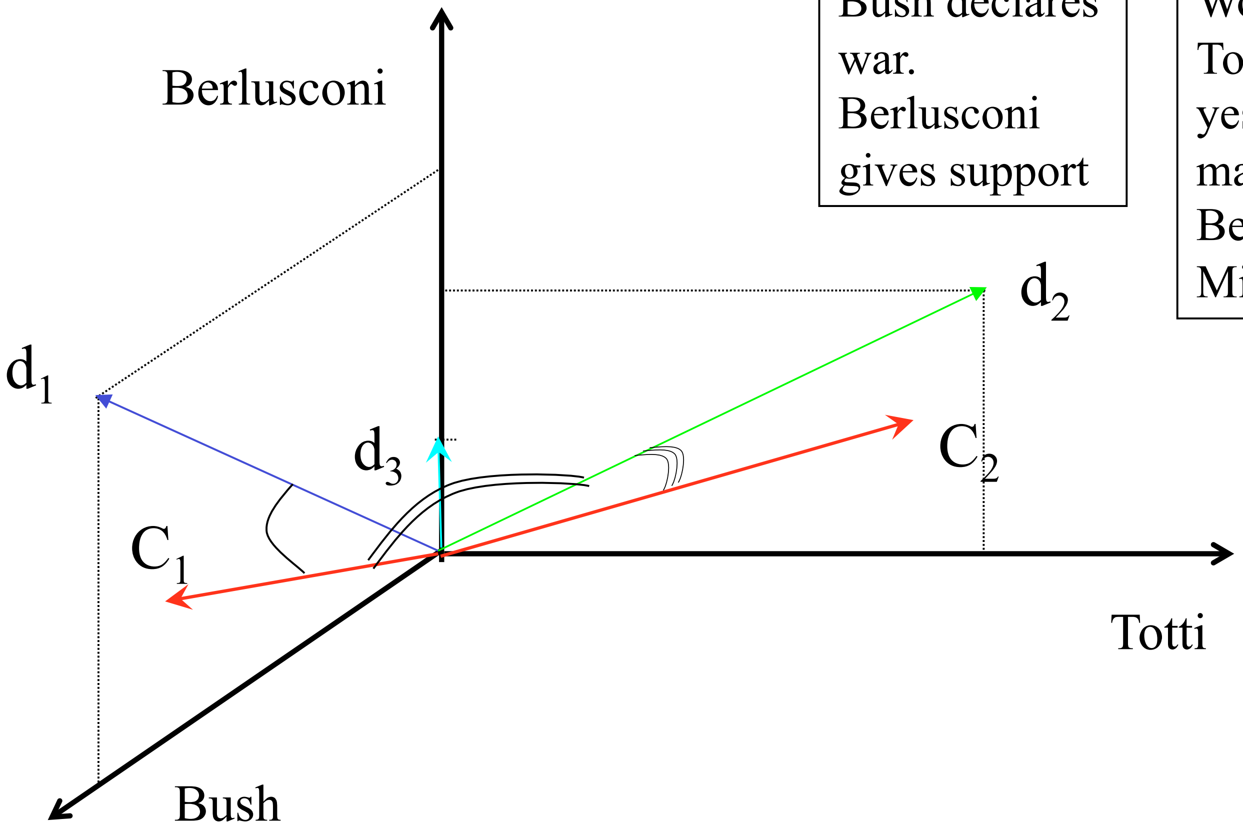


Text Classification Problem

- Given: $C = \{C^1,.., C^n\}$
  - a set of target categories:
  - the set *T* of documents,

define $f : T \rightarrow 2^C$

The Vector Space Model (VSM)



d_1: Politic

Bush declares war. Berlusconi gives support

d_2: Sport

Wonderful Totti in the yesterday match against Berlusconi's Milan

d_3: Economic

Berlusconi acquires Ibrahimović before elections

C_1 : Politics Category

C_2 : Sport Category

Summary of VSM

- VSM (Salton89')
  - Features are dimensions of a Vector Space **Linear Kernel**
  - Documents and Categories are vectors of feature weights.
  - *d* is assigned to C^i if $\vec{d} \cdot \vec{C}^i > th$
- Changing symbols

$$\vec{w} \cdot \vec{x} - th > 0 \Rightarrow \vec{w} \cdot \vec{x} + b > 0$$

Summary of Today Machine Learning Concepts

- Positive and Negative examples
- Feature representation
  - Kernels
- Learning Algorithm
- Training and test set
- Accuracy measurement
- Generalization/Empirical error Trade-off

What Next?

- Can we learn any function?
- Statistical Learning Theory
  - PAC learning

END

Several Kinds of Learning Algorithms

- Logic boolean expressions, (e.g. Decision Trees).
- Probabilistic Functions, (Bayesian Classifier).
- Separating Functions working in vector spaces
  - Non linear: KNN, neural network multiple-layers,…
  - **Linear**: **SVMs**, neural network with one neuron,…
- These approaches are largely applied In language technology
- Very Simple Example: Text Categorization