EMPIRICAL INFERENCE SCIENCE

Vladimir Vapnik
Columbia University, New York
NEC Laboratories America, Princeton

CONTENTS

SUMMARY

1. With the appearance of computers the concept of natural science, its methodology, and its philosophy started a painful process of a paradigm change:

The concepts, methodology, and philosophy of a *Simple World* move to very different concepts, philosophy and methodology of a *Complex World*.

2. In such changes an important role belongs to the mathematical facts that were discovered by analyzing the "Drosophila fly" of cognitive science the "Pattern recognition problem" and attempts to obtain their philosophical interpretation.

3. The results of these analyses lead to methods that go beyond the classical concept of science:

– creating generative models of events
– explain-ability of rules
– principles of refutation

SUMMARY

4. The new paradigm introduces direct search for solution (transductive inference, instead of inductive), the meditative principle of decision making, and a unity of two languages for pattern description: technical (rational) and holistic (irrational). These lead to the convergence of the exact science with humanities.

5. The goal of this talk is to demonstrate haw attempts to obtain a better generalization using a *limited* number of observations lead to non-classical paradigm of inference.

6. The main difference between the new paradigm (developed in the computer era) and the classical one (developed before the computer era) is the clame:

To quarantee the success of inference one needs to control the complexity of a computer program for inference rather than complexity of the function that this program produces. Program with low complexity can create a complex function which will generalize well.

PART 1:
PROBLEM OF INDUCTION
(A FRAMEWORK OF THE CLASSICAL PARADIGM)

PATTERN RECOGNITION PROBLEM

Pattern recognition problem can be regarded as the simplest model of the natural science: Given facts (observations)

$$(x_1, y_1), ..., (x_\ell, y_\ell)$$

find the rule

$$y = f(x)$$

where $x \in R^n$ and $y \subset \{-1, 1\}$.

To develop a mathematical model of pattern recognition one has to introduce the corresponding technical framework:

(1) pairs are iid generated by an unknown (but fixed) distribution $P(x, y)$.

(2) the quality of the obtained rule is defined by the expectation of predictive error.

$$Q(f) = 1/2 \int |y - f(x)| dP(x, y)$$

REALISM AND INSTRUMENTALISM APPROACHES IN PHILOSOPHY OF SCIENCE

The philosophy of science distinguishes between two approaches:

(1) *Realism*: the goal of science is to discover real laws of Nature.

(2) *Instrumentalism*: the goal of science is to discover rules which allow one to predict outcomes of events.

REALISM AND INSTRUMENTALISM IN PATTERN RECOGNITION

In pattern recognition the reflection of a *realism approach* is the so-called *generative model*: estimate the unknown distribution function $P(x, y)$ and construct the corresponding rule

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x,y)}{P(x,1) + P(x,-1)}$$

The reflection of *instrumentalist approach is Statistical Learning (or VC) theory*. It requires just to predict well: In a given set of functions

$$\{f(x,\alpha),\ \alpha \in \Lambda\}$$

find the one that minimizes the expected error

$$Q(f) = 1/2 \int |y - f(x,\alpha)| dP(x,y),\ \alpha \in \Lambda.$$

Experiments show that predictive models (SVM, Boosting, NN) are much more accurate than generative models.

Statistical Learning Theory gives up realism in favour of instrumentalism.

COMPLEXITY CONCEPT

In instrumentalist approach, an important role belongs to the concept of *complexity*. The first formulation of this concept, called *Occam razor*, states:

Entities are not to be multiplied beyond necessity

INTERPRETATION:

Entity means:
Thing, existence as opposite to its qualities or relations; thing that has real existence (Oxford Dictionary of Current English).

Beyond necessity means:
Not more than one needs to explain the observed facts.

EQUIVALENT FORMULATION:

Find the function with the smallest number of variables (free parameters, entities) that explains the observed facts.

OCCAM RAZOR IN PHYSICAL SCIENCE

A. Einstein about Simple and Complex World:

When solution is simple, God is answering.

Also

When the number of factors coming into play in phenomenological complexes is too large, scientific methods in most cases fail.

L. Landau about Complex World:

With four free parameters one can draw an elephant, with five one can draw an elephant rotating its tail.

In other words, classical science is an instrument for the simple world. When a world is complex, in most cases classical science fails.

Pattern recognition deals with a complex world where the number of factors coming into play often exceeds a thousand.

PREDICTIVE MODEL OF PATTERN RECOGNITION

In the late 1960s the analysis of the predictive *instrumentalism* model of pattern recognition started:

Predictive model of pattern recognition problem:

Given a set of functions

$$\{f(x,\alpha), \alpha \in \Lambda\}$$

and given iid training data

$$(x_1, y_1), ..., (x_\ell, y_\ell)\ x \in X^n,\ y \subset \{-1,\ 1\}$$

find the function from the given set

$$f(x,\alpha_0) \in f(x,\alpha),\ \alpha \in \Lambda$$

that minimizes the expected error functional (generalises)

$$Q(\alpha) = 1/2 \int |y - f(x,\alpha)| dP(x,y),\ \alpha \in \Lambda$$

THE MAIN TEOREM OF VC THEORY

One of the main results of Vapnik-Chervonenkis theory is:

For any algorithm that selects one function from the set of admissible functions there are two **and only two** *factors responsible for generalization. They are:*

(1) Empirical loss (# of training error made by the chosen function)

(2) The capacity (complexity) measure (VC entropy, VC dimension) of the admissible set of functions $\{f(x,\alpha),\ \alpha \in \Lambda\}$ *from which the desired function was selected*

The measure of capacity which describes diversity of the admissible set of functions plays a crucial role in the VC theory.

CAPACITY CONCEPTS: THE VC ENTROPY AND THE GROWTH FUNCTION

Let $f(x, \alpha) \subset \{-1, 1\}, \alpha \in \Lambda$ be a set of indicator functions and let

$$x_1, ..., x_\ell$$

be an i.i.d. sample from the distribution P. Consider the number

$$N = N^\Lambda(x_1, ..., x_\ell) \leq 2^\ell$$

of *different* separations of the sample by functions from this set.

- We call the quantity

$$H_P^\Lambda(\ell) = \log_2 E_{\{x_1,...,x_\ell\}} N_P^\Lambda(x_1, ..., x_\ell)$$

  the **VC entropy** of the set of indicator functions for samples of size ℓ.

- We call the quantity

$$G^\Lambda(\ell) = \log_2 \max_{x_1,...,x_\ell} N_P^\Lambda(x_1, ..., x_\ell)$$

  the **Growth function**.

THE STRUCTURE OF THE GROWTH FUNCTION: THE VC DIMENSION

The Growth function is either the linear function

$$G^{\Lambda}(\ell) = \ell \ln 2$$

or bounded by the logarithmic function

$$G^{\Lambda}(\ell) \leq h \ln \left(\frac{e\ell}{h} \right) = h \left(\ln \frac{\ell}{h} + 1 \right),$$

where h is the largest ℓ^* for which

$$G^{\Lambda}(\ell^*) = \ell^* \ln 2.$$

The value h is called the **VC dimension** of the set of indicator functions.

Therefore

$$H_P^{\Lambda}(\ell) \leq G^{\Lambda}(\ell) \leq h \left(\ln \frac{\ell}{h} + 1 \right).$$

THE MAIN RESULTS OF THE VC THEORY

Suppose that our algorithm selects the function from the admissible set that minimizes the number of training errors (minimizes the empirical loss)

1. The algorithm is consistent for **a given probability measure** P, if and only if the VC entropy $H_P^\Lambda(\ell)$ is such that

$$\lim_{\ell\to\infty} \frac{H_P^\Lambda(\ell)}{\ell} = 0.$$

2. The algorithm is consistent for **any** probability measure P if and only if

$$\lim_{\ell\to\infty} \frac{G^\Lambda(\ell)}{\ell} = 0.$$

3. The algorithm is consistent for **any** probability measure P if and only if the VC dimension is finite ($h < \infty$).

CONCEPT OF FALSIFIABILITY

In the 1920s, K. Popper introduced the following concept of falsifiability: The set of vectors

$$x_1, \ldots, x_\ell, \ x_i \in X \tag{1}$$

cannot falsify the set of indicator functions $\{f(x, \alpha), \alpha \in \Lambda\}$ if all 2^ℓ possible separation of vectors (1) into two categories can be accomplished using functions from this set. (The VC dimension of the set is infinite.)

The set of vectors (1) **falsifies** the set $\{f(x, \alpha), \alpha \in \Lambda\}$ if there exists such separation of the set (1) into two categories that cannot be obtained using an indicator function from the set $\{f(x, \alpha), \alpha \in \Lambda\}$.

THE VC DIMENSION

A set of functions $\{f(x,\alpha), \alpha \in \Lambda\}$ has VC dimension h if

(1) **there exist** h vectors that cannot falsify this set and

(2) **any** $h+1$ vectors falsify it.

THE POPPER DIMENSION

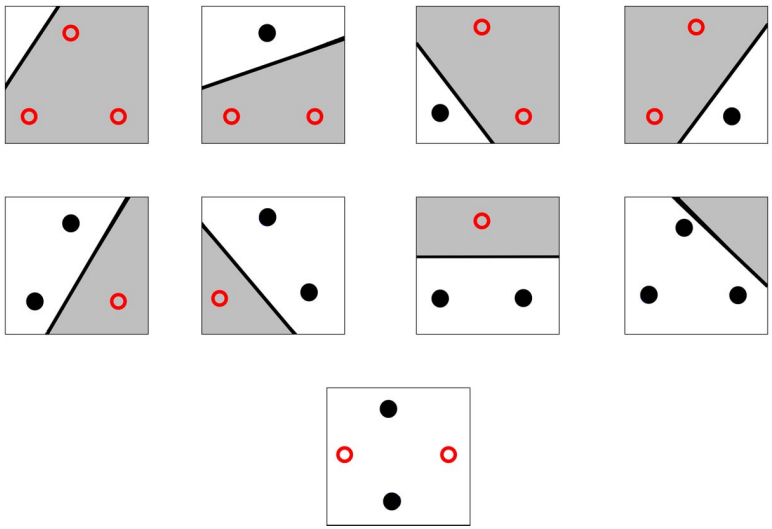
A set of functions $\{f(x,\alpha), \alpha \in \Lambda\}$ has the Popper dimension h if:

(1) **any** h vectors cannot falsify it and

(2) **there exist** $h+1$ vectors that falsify this set.

VC AND POPPER DIMENSION: ILLUSTRATION

The VC dimension of the set of oriented lines in the plane is 3.



There exist three vectors that cannot falsify linear laws.
Any four vectors falsify linear law.

The Popper dimension does not exceed 2 for any dimensionality of the space since it requires non-falsifiability for ANY h points including ones situated on the line.

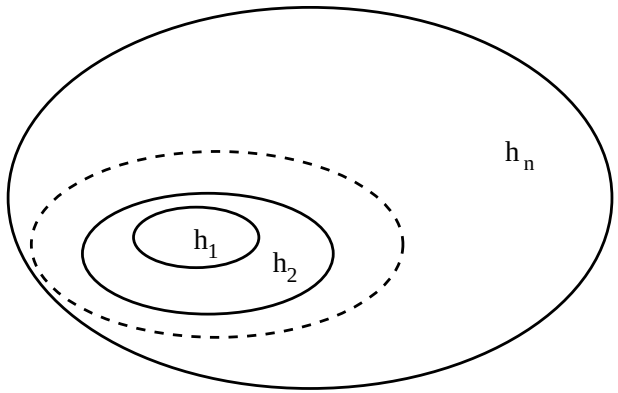
VC BOUNDS AND SRM PRINCIPLE

With probability $1-\eta$ the following inequality holds true

$$R(\alpha) \leq R_{emp}(\alpha) + \Phi\left(\frac{h}{\ell}, \frac{-\ln \eta}{\ell}\right).$$

To minimize the risk $R(\alpha)$ one can minimize the empirical risk $R_{emp}(\alpha)$.

However, one minimizes the risk better if one can make h the controlled variable and minimize the bound over both α and h.



The structural risk minimization principle is strongly universally consistent.

THE OCCAM RAZOR PRINCIPLE AND THE SRM PRINCIPLE

THE OCCAM RAZOR PRINCIPLE:

Entities should not be multiplied beyond necessity.

INTERPRETATION OF OCCAM'S RAZOR PRINCIPLE:

Do not use more concepts than you need to explain the observed facts. CUT EXTRA CONCEPTS.

THE SRM PRINCIPLE:

Explain the observed facts using a function from the subset with the smallest VC dimension (capacity).

INTERPRETATION OF SRM PRINCIPLE:

Explain the observed facts using a theory which is easy to falsify.

THE CRUCIAL POINT

Does the VC dimension describe the number of entities?

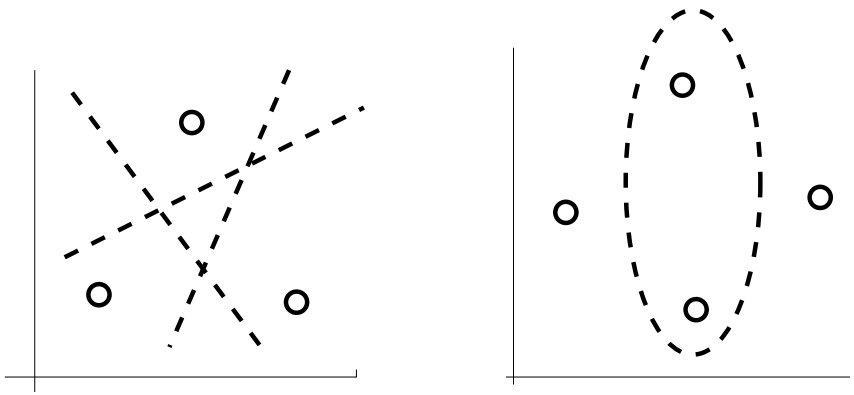
EXAMPLE 1: The VC dimension is equal to the number of entities (parameters)

The VC dimension h of the set of linear indicator functions

$$I(x, w) = \text{sgn}((x, w) + b), \quad x \in R^n, \ w \in R^n$$

is equal to the number of parameters

$$h = n + 1.$$

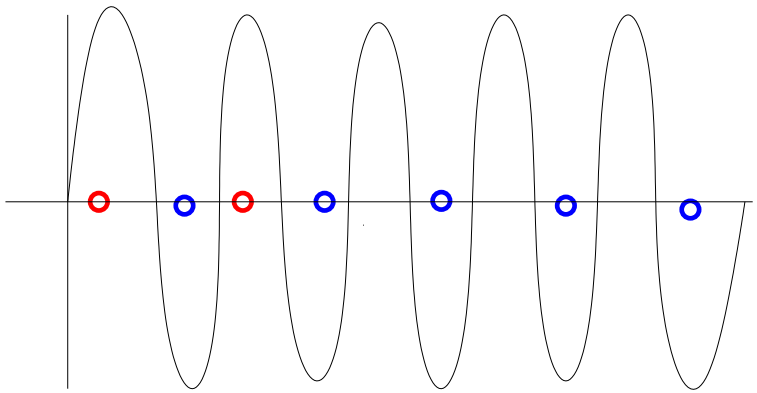


EXAMPLE 2: The VC dimension is larger than the number of entities (parameters)

The VC dimension of the set of functions

$$I(x, a) = \text{sgn}\{\sin ax\}, \quad x \in R^1, \; a \in R^1$$

is infinite.



Quotation from Popper:
"According to common oppinion, the sine-function is a simple one."

EXAMPLE 3: The VC dimensions is less than the number of entities (parameters)

We say that a hyperplane

$$(w^*, x) + b = 0, \quad |w^*| = 1$$

is the Δ-margin separating hyperplane if it classifies vectors x as follows

$$y = \begin{cases} -1, & \text{if } (x, w) + b \geq \Delta \\ 1, & \text{if } (x, w) + b \leq -\Delta. \end{cases}$$

LARGE MARGIN CONCEPT

Let the vectors $x \in R^n$ belong to a sphere of radius R. Then the set of Δ-margin separating hyperplanes has VC dimension h bounded as follows

$$h \leq \min \left\{ \frac{R^2}{\Delta^2}, n \right\} + 1.$$

THE IDEA OF SUPPORT VECTOR MACHINES

- **Increase the number of entities:**

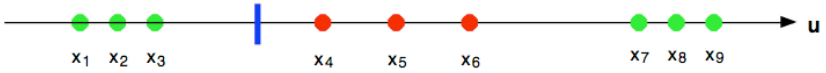
  Map the input vectors $x \in X$ into a high-dimensional space $z \in Z$.

- **Control the VC dimension in high-dimensional space Z:**

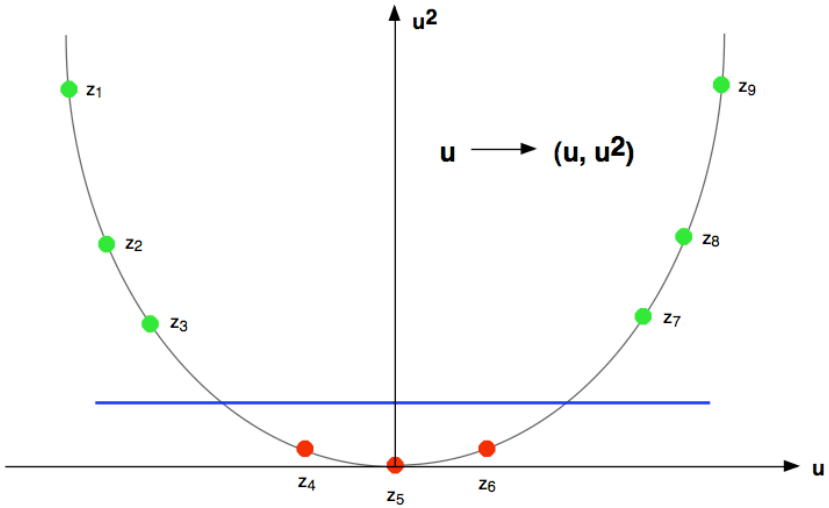
  Construct a hyperplane with a large margin in space Z.

The idea is that with increasing the dimensionality of the space, the ratio of the radius of the sphere to the value of the margin can be small. This will imply a small VC dimension and guarantee good generalization.

ILLUSTRATION



Hyperplane cannot separate red and green points in one-dimensional space



Hyperplanes separate those points with a margin in two-dimensional space.

TECHNICAL DETAILS

The VC bounds

$$R(\alpha) \leq R_{emp}(\alpha) + \Phi\left(\frac{h}{\ell}, \frac{-\ln \eta}{\ell}\right).$$

Realization of VC bounds:

Minimize the functional

$$R^*(w) = \frac{1}{\ell}(w, w) + \frac{C}{\ell}\sum_{i=1}^{\ell}\theta(\xi_i)$$

subject to constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i, \ \ \xi_i \geq 0, \quad \forall i$$

The only compromise

$$R^*(w) = \frac{1}{\ell}(w, w) + \frac{C}{\ell}\sum_{i=1}^{\ell}\xi_i$$

MORE TECHNICAL DETAILS

Mapping into a space Z is equivalent to introducing a Mercer similarity measure $K(x_i, x_j)$ between any two examples x_i and x_j in space X.

An example of Mercer similarity measure (Mercer kernel) is

$$K(x_i, x_j) = \exp\{a|x_i - x_j|^2\}$$

The non-linear in space X (but linear in parameters α) solution is

$$f(x, \alpha_0) = \sum_{i=1}^{\ell} \alpha_i y_i K(x_i, x) + b,$$

where parameters α_i minimize the functional

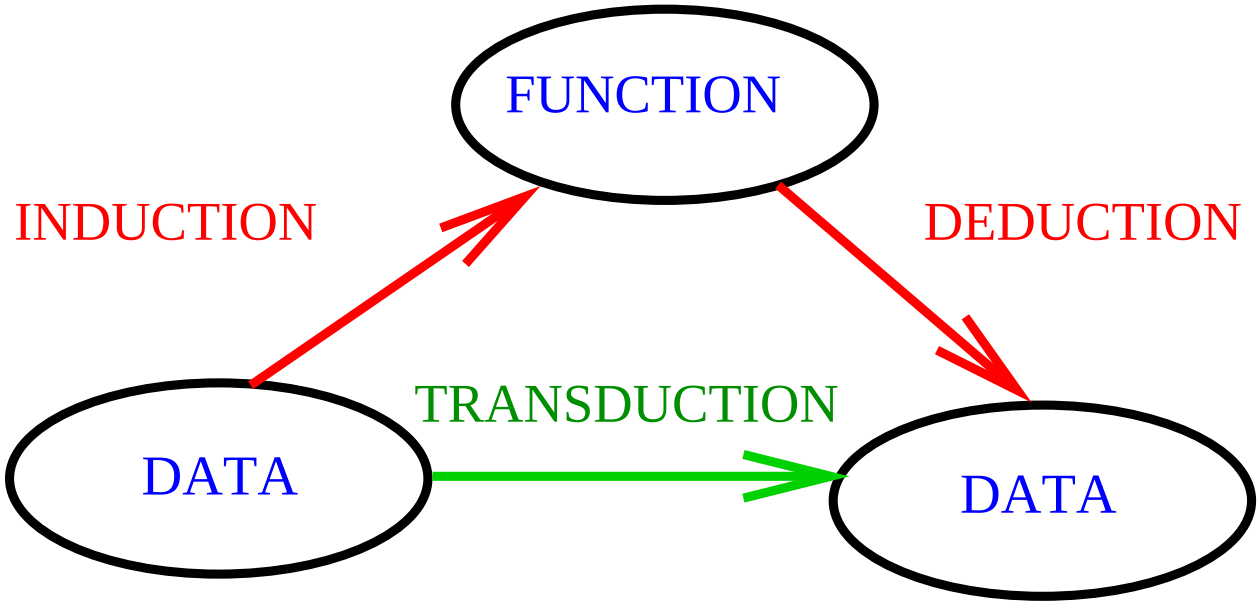
$$\Phi(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to the constraints

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C \quad \forall i$$

PART 2
NON-INDUCTIVE INFERENCES
(TRANSDUCTION VERSUS INDUCTION)

INDUCTIVE AND TRANSDUCTIVE INFERENCES .



The concept of transduction looks more fundamental than concept of induction: The bounds for induction are derived by first obtaining bounds for transduction and only using these bounds one can obtain bounds for induction. Also, bounds for transduction are more accurate than those for induction.

WHAT IS THE TRANSDUCTION PROBLEM

Given a set of training data

$$(x_1, y_1), ..., (x_\ell, y_\ell)$$

and given a set of test data

$$x_1^*, ..., x_k^*$$

find among the admissible set of classification vectors

$$Y^* \in \{Y^* : \ (y_1^*, ..., y_k^*)\}$$

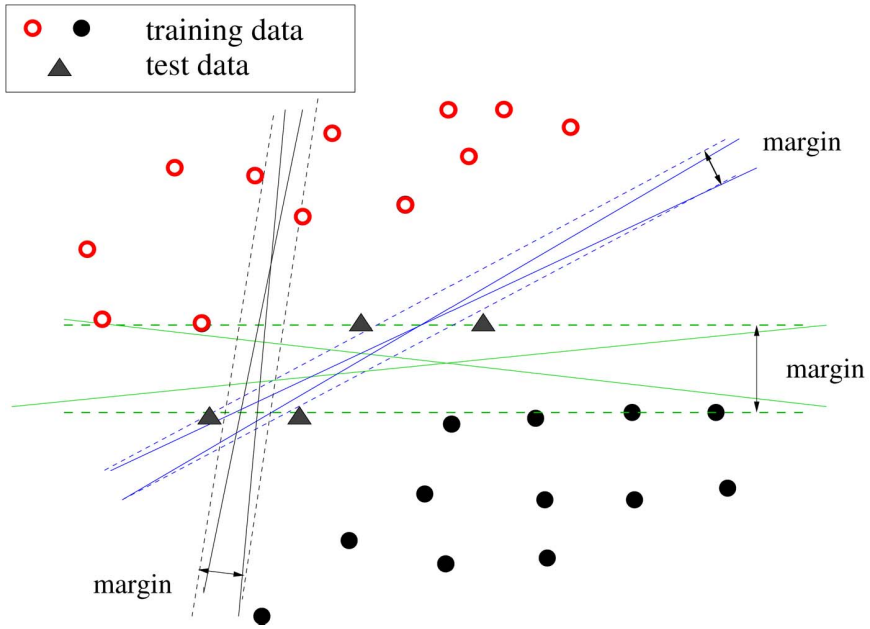
the best classification vector.

An important special case is

$$y_i^* = f(x_i, \alpha_*),$$

where $f(x, \alpha_*) \in f(x, \alpha), \alpha \in \Lambda$.

EQUIVALENCE CLASSES



The infinite set of functions is factorized into a finite set of equivalence classes. Large margins defines a large equivalence class. The structure in SRM is organized in such a way that elements with small numbers include large equivalence classes.

KDD CUP 2001 DATA ANALYSIS

(W,P-C,B,C,E,S, Bioinformatics, V1,#1,2003)

Data was provided by DuPont Pharmaceutical for the KDD competition.

- x_i are 139,351 dimensional binary vectors.
- The training set contained 1,909 examples: 42 (2.2%) of vectors belong to the first class (vectors which bind), 1,867 (97.8%) belong to the second class.
- The test set contained 634 examples: 150 (23.66%) positive and 484 (76.34%) negative examples.
- Result p is evaluated as follows

$$p = \frac{1}{2}(p_1 + p_2),$$

where p_1 and p_2 are the percentages of correct classifications of the positive and negative examples.

PREDICTION OF MOLECULAR BIOACTIVITY

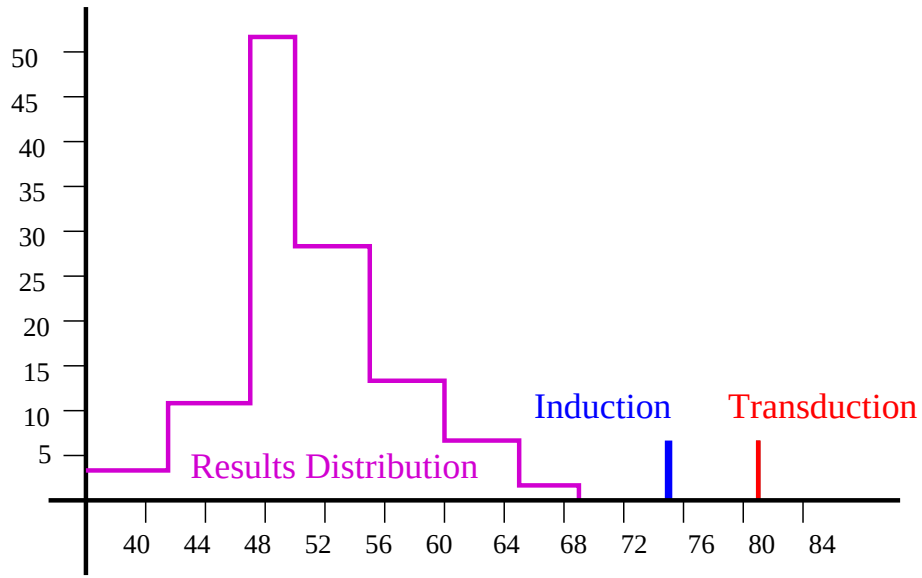
RESULTS OF COMPETITION: The winner's score was 68%.

SVM scores:

For **inductive** inference (using training data only): **74.5**%.

For **transductive** inference (using also unlabeled test data): **82.3**%.

Comparison to other 119 contestants of the competition.



SELECTIVE INFERENCE

Given ℓ training examples

$$(x_i, y_1), ..., (x_\ell, y_\ell)$$

and n candidates vectors

$$x_1^*, ...x_n^*$$

select from n candidates the k vectors with the highest probability of belonging to the first class.

Drug bioactivity: From n given candidates select k representatives with the highest probability of belonging to the group with a high bioactivity.

National security: From given candidates select k representatives with the highest probability of belonging to a terrorist group.

Selective Inference is less demanding than Transductive. It can have a more accurate solution than one obtained from Transductive Inference.

THE IMPERATIVE FOR THE COMPLEX WORLD

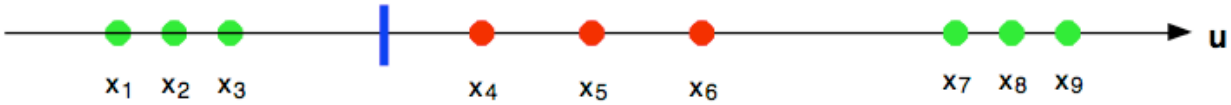
When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need, but not a more general one.

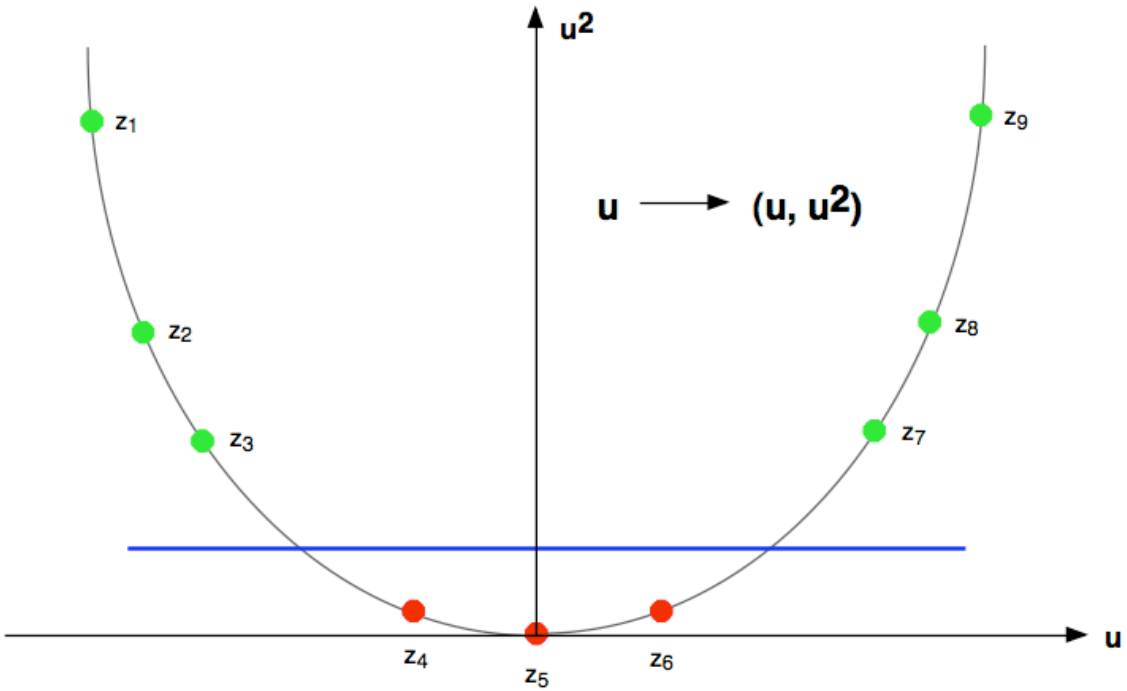
Example

- Do not estimate a density if you need to estimate a function.
  (Do not use classical statistics paradigm for prediction.)
- Do not estimate a function if you need to estimate its values at given points.
  (Try to perform transduction instead of induction.)
- Do not estimate predictive values if your goal is to act well.
  (A good action strategy does not necessary rely on good prediction.)

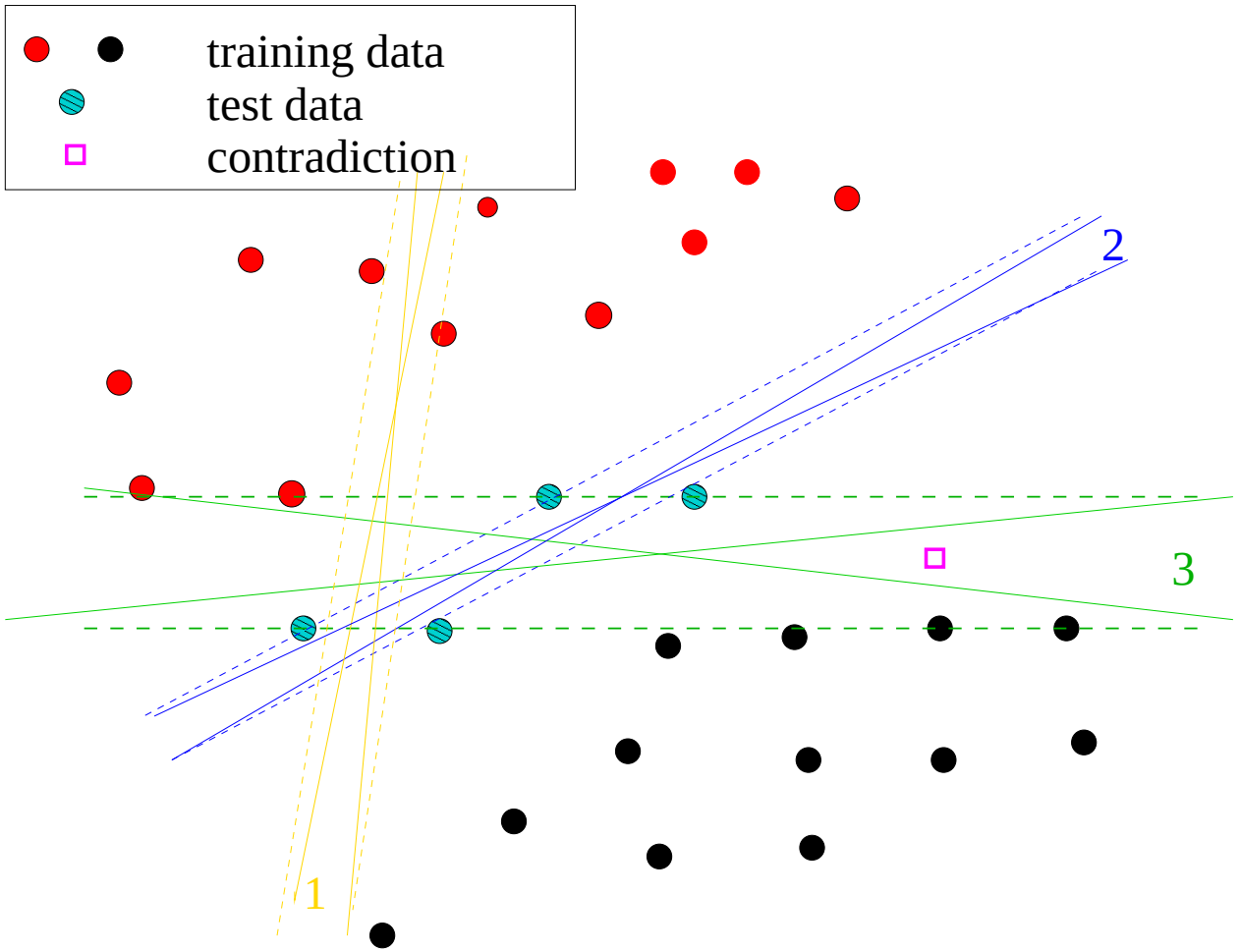
PART 3
USING REFINED PART OF THE VC THEORY (CONVERGENCE OF SCIENCES AND HUMANITIES)

WHAT IS WRONG WITH LARGE MARGIN?





BACK TO VC ENTROPY: THE CONCEPT OF CONTRADICTION



IDEA OF UNIVERSUM

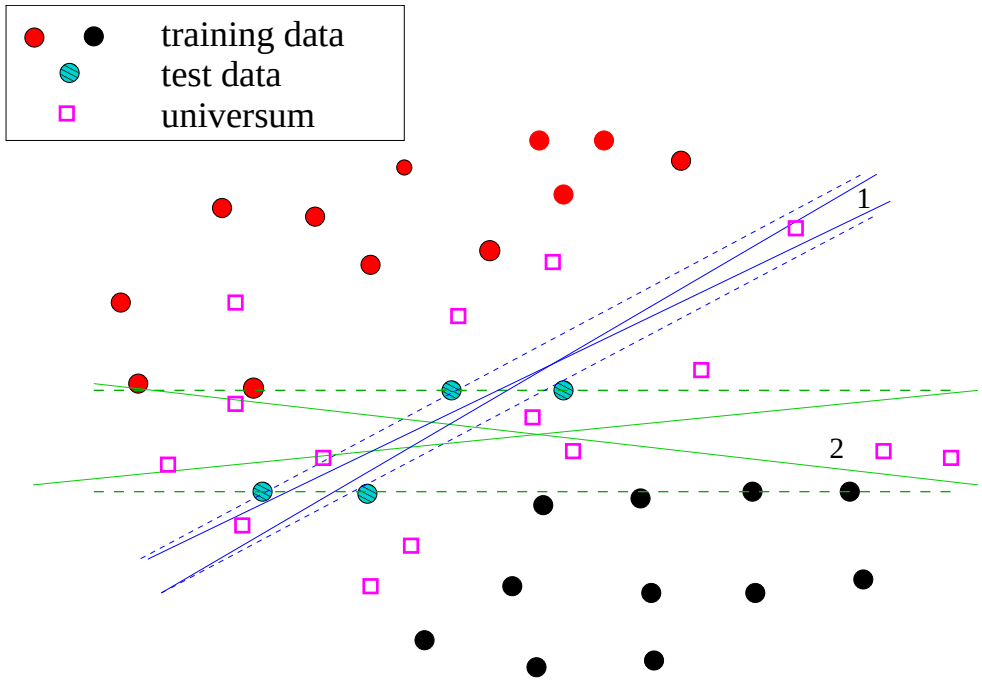
The universum is prior information about data which form our problem
This prior information is different from one used in Bayesian inference:

- Bayesian inference uses prior information about decision rules appropriate for the problem one solves.
- Universum is prior information about world (data) which appear during training and test stages.

It is very difficult to have prior information on a set of decision rules. It seems that obtaining (virtual) information about data is much easier.

In some sense literature (and art) is a sort of such (virtual) information about the real world.

INFERENCE BASED ON THE NUMBER OF CONTRADICTIONS ON UNIVERSUM



Classify test data by the equivalence class that separates training data well and has the maximal number of contradictions on the Universum.

EXPERIMENTS WITH DIGIT RECOGNITION

Choose the equivalence class that separates digits 5 and 8 and makes the maximal number of contradiction on the Universum (not large margin).

No. of train. examples	250	500	1000	2000	3000
Test Err. SVM (%)	2.83	1.92	1.37	0.99	0.83
Test Err. SVM+U_1 (%)	2.43	1.58	1.11	0.75	0.63
Test Err. SVM+U_2 (%)	1.51	1.12	0.89	0.68	0.60
Test Err. SVM+U_3 (%)	1.33	0.89	0.72	0.60	0.58

The Universums (5,000 elements) were constructed as follows:

U_1: Selects random digits from the other classes (0,1,2,3,4,6,7,9).

U_2: Creates an artificial image by first selecting a random 5 and a random 8, and then for each pixel of the artificial image choosing with probability 1/2 the corresponding pixel from the image 5 or from the image 8.

U_3: Creates an artificial image by first selecting a random 5 and a random 8, and then constructing the mean of these two digits.

(J. Weston, R. Collobert, F. Sinz, (2006))

EXPERIMENTS WITH FACES RECOGNITION

Choose the equivalence class that separates male faces from female faces and makes a maximal number of contradiction on Universum U_3 (Xue Bai, (2007)).

Size of Universum	0 (SVM)	10	50	100	300	500
Test Err%. Exp 1	20	17	07	04	02	00
Test Err%. Exp 2	20	17	06	06	02	00
Test Err%. Exp 3	13	13	13	14	13	12
Test Err%. Exp 4	16	08	10	08	02	01

Namber of training examples 40 (20+20).
Number of test examples 200.

FURTHER CAPACITY CONTROL: SVM+

When constructing the desired hyperplane in a non-separable case the SVM minimizes the functional

$$R = C\sum_{i=1}^{\ell}\xi_i + (w, w)$$

over vector w and ℓ slack variables subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i,\ \ i = 1, ..., \ell$$

Now we would like to control the capacity of the slack variables.

Let slack variables be a realization of some correcting function

$$\xi_i = \phi(x_i, \beta_*),$$

that belongs to a set of admissible functions $\phi(x_i, \beta),\ \ \beta \in \mathcal{B}$ with restricted capacity. Let us chose both the decision function $f(x, \alpha_*)$ and the correcting function $\phi(x_i, \beta)$.

SVM+: FORMULATION

Let us map vector x into two different conjugate spaces: the space of decision functions Z and the space of correcting functions Z^*.

$$x \longrightarrow z,$$
$$x \longrightarrow z^*$$

In these spaces, let the decision and correcting functions be linear functions

$$y = f(x, \alpha) = (w, z) + b,$$
$$\xi = \phi(x, \beta) = (w^*, z^*) + d$$

The problem is to minimize the functional

$$R = [(w, w) + \gamma(w^*, w^*)] + C \sum_{i=1}^{\ell} [(w^*, z_i^*) + d]$$

subject to the constraints

$$y_i[(w, z_i) + b] \geq 1 - [(w^*, z_i^*) + d],$$
$$[(w^*, z_i^*) + d] \geq 0$$

SVM+: DUAL SPACE SOLUTION

The decision and correcting functions have the form

$$f(x,\alpha) = \sum_{i=1}^{\ell} \alpha_i y_i K(x_i, x) + b$$

$$\phi(x,\beta) = \frac{1}{\gamma}\sum_{i=1}^{\ell}(\alpha_i + \beta_i - C)K^*(x_i, x) + d$$

where $\alpha \geq 0,\ \beta \geq 0$ maximize the functional

$$R = \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \frac{1}{2\gamma}\sum_{i,j=1}^{\ell}(\alpha_i+\beta_i-C)(\alpha_j+\beta_j-C)K^*(x_i, x_j)$$

subject to the constraints

$$\sum_{i=1}^{\ell}\alpha_i y_i = 0 \quad \text{and} \quad \frac{1}{\ell}\sum_{i=1}^{\ell}(\alpha_i + \beta_i) = C$$

ONE STEP MORE: LEARNING HIDDEN INFORMATION

Given iid triplets

$$(x_1, x_1^*, y_1), ...(x_\ell, x_\ell^*, y_\ell)$$

find the function $f(x, \alpha_0)$ in the set $\{f(x, \alpha), \alpha \in \Lambda\}$ that minimizes the loss

$$Q(f) = 1/2 \int |y - f(x)| dP(x, y)$$

where the x_i^* are *hidden information* that available only for the training stage and will **not** be available for the test stage.

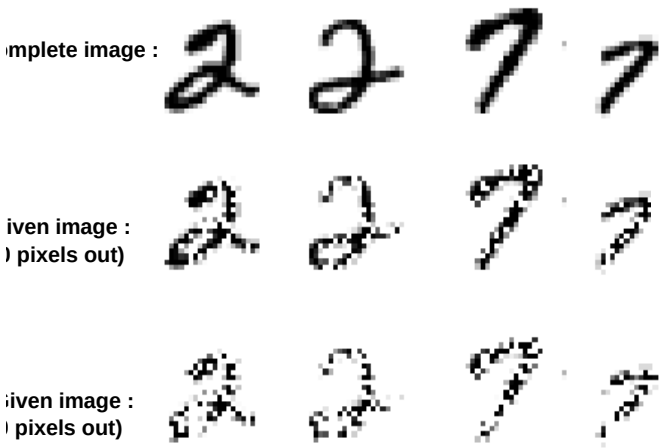
EXAMPLES

(1) Given medical conditions x, one has to predict outcome of treatment in a year. For the training stage, one possesses additional information x^* about medical conditions half a year before prediction. Can this help?

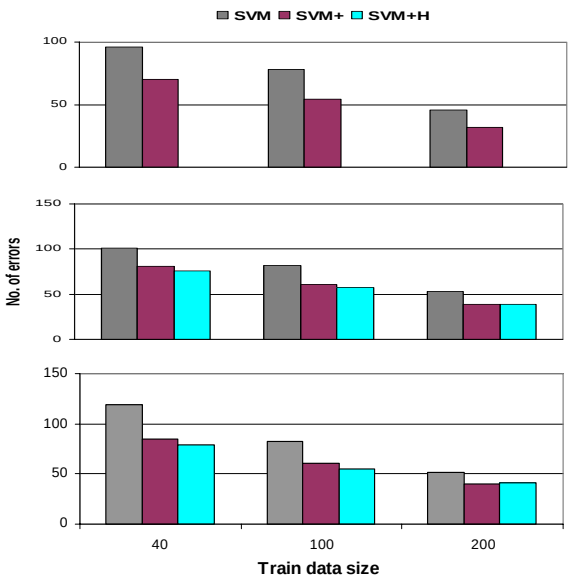
(2) One has to predict an outcome using "cheap" features. During training one possesses both "cheap and expensive" features. Can this help?

SVM+ solves this problem by mapping x in Z and x^* in Z^*.

ILLUSTRATION



Error Comparison



Errors on 2,060 test digits

Algo. / Train Size	0 pixels out		200 pixels out			400 pixels out		
	SVM	SVM+	SVM	SVM+	SVM+H	SVM	SVM+	SVM+H
40	96	70	101	81	76	119	85	79
100	78	54	82	61	57	83	61	55
200	46	32	53	39	39	52	40	41
Gain over SVM		27-30%			25-30%			21-34%

MASTER-CLASS LEARNING

SVM+ allows one to use the scheme that admits two learning languages:

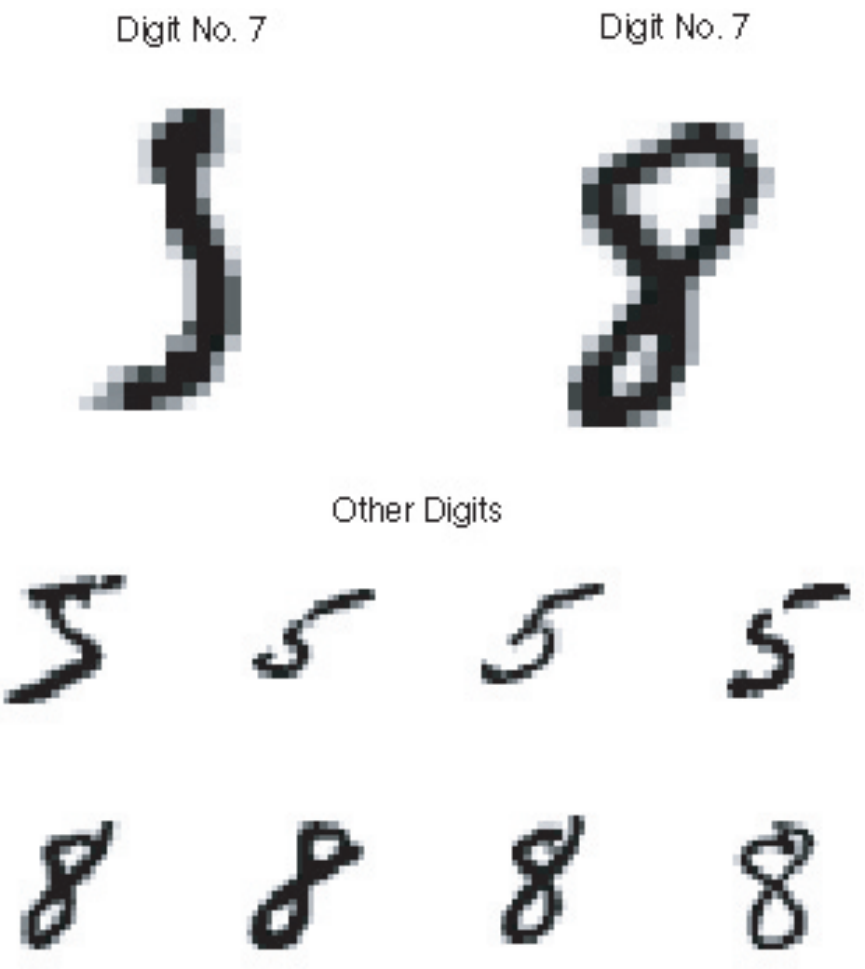
1. A technical language (pixel space for digit recognition)

2. A holistic description language (metaphoric, gestalt).

The role of holistic information is to correct decision rules that are constructing in the technical space.

This technique is thus using master-class teachers.

MASTER-CLASS DIGIT RECOGNITION LEARNING: TECHNICAL SPACE



MASTER-CLASS DIGIT RECOGNITION LEARNING: HOLISTIC SPACE

Holistic description of digit FIVE (# 7)

Has almost one part. Has some snakeness. Has flexibility. A snake with almost no head. Slightly wriggling. Standing on the end of its tail (one point). Unsteady. Waving and hesitating. Only one pocket – very very insignificant – very shallow. A hardly noticeable upper part is a small handle with a bulb and the lower part is a long young crescent. No hill. Everything seems unclear. The creature is looming. Something unfinished. No infinity. A piece of a rope. The rope is rather thick and not old. Everything is too oblong. Asymmetrical. Almost non-slanting to the right. No movement. No criss crosses. Almost strait line. No curling of the ends.

MASTER-CLASS DIGIT RECOGNITION LEARNING: HOLISTIC SPACE

Holistic description of digit EIGHT (# 9)

Two-part creature. Slightly slanted to the right. It is absolutely closed without any appendix. Not beautiful. Not regular. The head is much bigger than the bottom. You cannot move along with such a heavy head. If it could be turned upside down it would be regular. The head is curved. The left part of the head has a dent which is no good (there should be no dents). It is a bit lopsided (due to the left part of head). The creature does not look any way. The criss-cross angles are not equal. The criss-cross lines are curved, flexible. The upper angle is more obtuse than lower one. The infinity way is absolute. The snake is coiled up in two rings. It is uncomfortable. No movement and no sleep.

CODES FOR HOLISTIC DESCRIPTION

1 Two-part-ness
2. Slant, tiltness (to which side, big or small)
3. Head-first (bottom-first)
4. Sharp-or-piercing-tool-ness
5. Roundishness
6. Flexibility (hardness)
7. Movement (speed)
8. Running
9. Walking
10. Flying
11. Crawling
12. Standing
13. Openness (outsideness)
14. Cavity-pocket-ness (the depth of the cavity)
15. Aggressiveness (peacefulness)

.....

BIG PICTURE

1. About fifty years ago machine learning researchers started analysing the problem of generalization (problem of induction).

2. They introduced complexity concepts (VC entropy and VC dimension) and showed that there are two and only two very simple factors responsible for generalization (complexity and accuracy of training data classification).

3. They showed (theoretically and experimentally) that the classical principle of generalization (Occam razor) is wrong (SVMs and Boosting are built in direct contradiction with this principle).

4. They discovered a strongly universally consistent principle for generalization (SRM) and corresponding algorithms (SVM).

5. They discovered theoretically and later confirmed experimentally that inductive inference is less accurate (and less universal) than transductive.

6. They showed that advanced methods of inference use techniques that could have human interpretation and human-like searching for truth (meditation, virtual world, holistic (metaphoric) language, transmitting hidden information).

This leads to a convergence of exact science and humanities.

MACHINE LEARNING AND EMPIRICAL INFERENCE SCIENCE

1. The classical paradigm of science was inspired by success of physics, which stresses the simplicity of the world.

2. Computers allowed us to enter the complex world, which brings into play many factors. This world is very different from the one where physicists act.

3. The only advanced theory and methodology related in a complex (high dimensional) world is pattern recognition.

4. Pattern recognition was very successful in solving many practical problems.

5. However, it is much more significant if the facts about the complex world discovered by pattern recognition bring radical changes in our understanding of science.

6. To advance a new paradigm scientists of different specialities (philosophers, linguists, biologists, behaviour scientists) must know these facts. Jointly they can create interpretation and develop a new discipline which can be called

Empirical Inference Science.

CONCLUSION: TWO METAPHORS FOR A SIMPLE WORLD

I want to know God's thoughts ...

When the solution is simple, God is answering.

A. Einstein

INTERPRETATION:

Nature is a realization of the simplest conceivable mathematical ideas. I am convinced that we can discover by means of purely mathematical constructions concepts and laws, connecting them each to other, which furnish the key to understanding of natural phenomena.

ALSO

When the number of factors coming into play in a phenomenological complexes is too large, scientific methods in most cases fail. *A. Einstein.*

CONCLUSION: TWO METAPHORS FOR A COMPLEX WORLD

The Devil imitates God.

Definition of the Devil.

INTERPRETATION

Actions (decisions) based on one's understanding of God's thoughts can bring one to catastrophe. Understanding God's thoughts is an ill-posed problem.

Subtle is Lord, but malicious He is not.

A. Einstein

INTERPRETATION

Subtle is Lord — one cannot understand His thoughts, but

malicious He is not — one can act well without understanding them.

WHAT IS EMPIRICAL INFERENCE SCIENCE ABOUT?

The Empirical Inference Science has to answer the question:

How to act well without understanding God's thoughts?