
**PROBLEMS OF EMPIRICAL INFERENCE IN
MACHINE LEARNING AND PHILOSOPHY OF
SCIENCE**

Vladimir Vapnik

FOUR PERIODS OF DEVELOPMENT OF EMPIRICAL INFERENCE SCIENCE

1970 – 1990

Development of Basics of Statistical Learning Theory
(the VC theory)

1992 – 2000

Development of Large Margin Technology (SVMs).

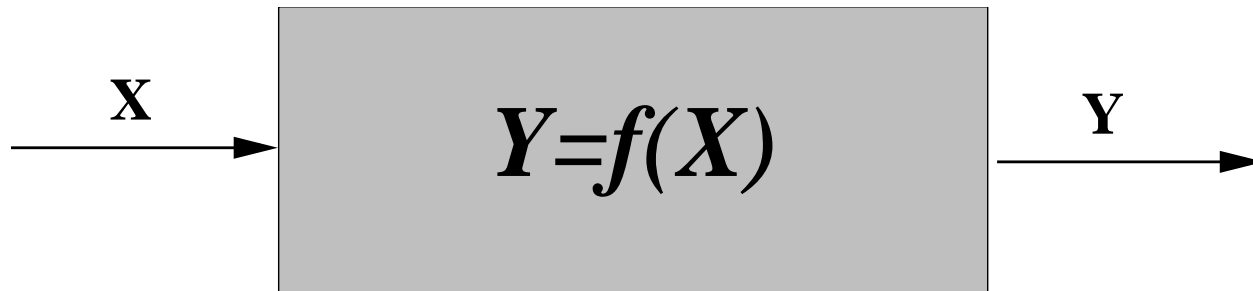
1995 – 2005

Development of Non-Inductive Methods of Inferences.

1998 — ...

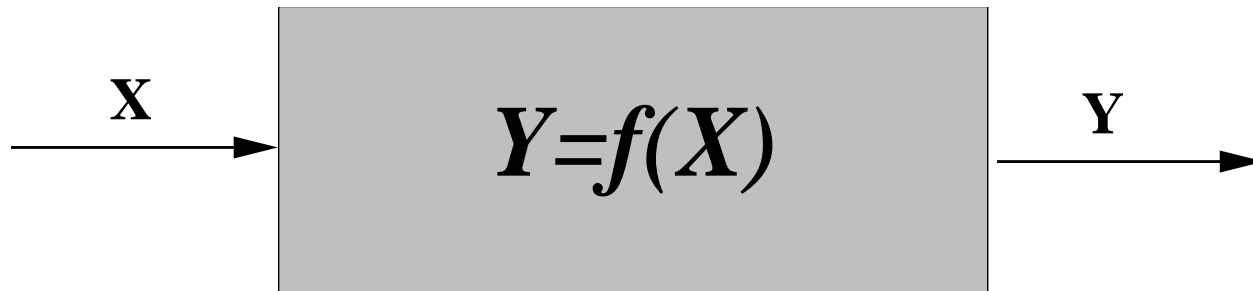
Development of Directed Ad-Hoc Inference (DAHI)

IDENTIFICATION OR IMITATION (REALISM OR INSTRUMENTALISM)?



- The goal of philosophical realism is to **identify** the unknown law.
 - The goal of philosophical instrumentalism is to **imitate** the unknown law.
- The classical statistics reflects positions of philosophical realism while the VC theory reflects position of philosophical instrumentalism.*

CLASSICAL STATISTICS: THE GENERATIVE MODELS



Given data of observation

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

find in a set of admissible models the model that uses Black Box to generate outputs y_i for a given inputs x_i .

Classical statistics reflects ideas of philosophical realism.

Given measurements of function $y = f_0(x)$

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

estimate the function $f_0(x)$.

The generative model of data:

1. An unknown function $f_0(x)$ belongs to the family $f(x, \alpha)$, $\alpha \in R^n$.
2. The measurements have additive noise

$$y_i = f(x_i, \alpha_0) + \xi_i, \quad E x_i \xi_i = 0.$$

3. The law that defines the noise

$$\xi = y - f(x, \alpha_0)$$

is known. For example, it is the normal law

$$P(\xi) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{\xi^2}{2\sigma^2} \right\}.$$

The idea of inductive inference:

Using a model of functions $f(x, \alpha)$, $\alpha \in R^n$ and data

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

estimate parameters of the density from the set

$$P(y - f(x, \alpha)) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y - f(x, \alpha))^2}{2\sigma^2} \right\}.$$

The method of inference:

Use the Maximum Likelihood method:

$$\alpha_\ell = \arg \min_{\alpha} \sum_{i=1}^{\ell} (y_i - f(x_i, \alpha))^2.$$

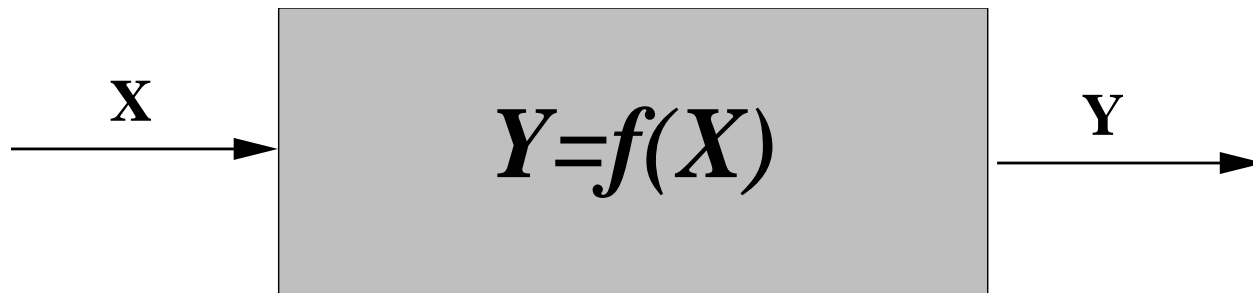
The subject of analysis:

Quality of parameters estimation: consistency, asymptotic normality, efficiency.

With increasing the dimensionality of a problem, the amount of resources that one needs to solve the identification problem increases exponentially.

Example

- Suppose that to approximate a one dimensional function “with fixed smoothness properties” one needs N terms in a Fourier expansion. Then to approximate a d -dimensional function “with the same smoothness properties”, one needs N^d terms in Fourier expansion.
- To estimate N^d parameters of the Fourier expansion well, one needs cN^d , $c > 1$ observations.



In a given set of functions $f(x, \alpha), \alpha \in \Lambda$ find one that has the smallest probability of different classifications from ones given by the Black Box.

Machine learning reflects ideas of philosophical instrumentalism.

DIFFERENCE BETWEEN IDENTIFICATION AND IMITATION MODES OF INFERENCE

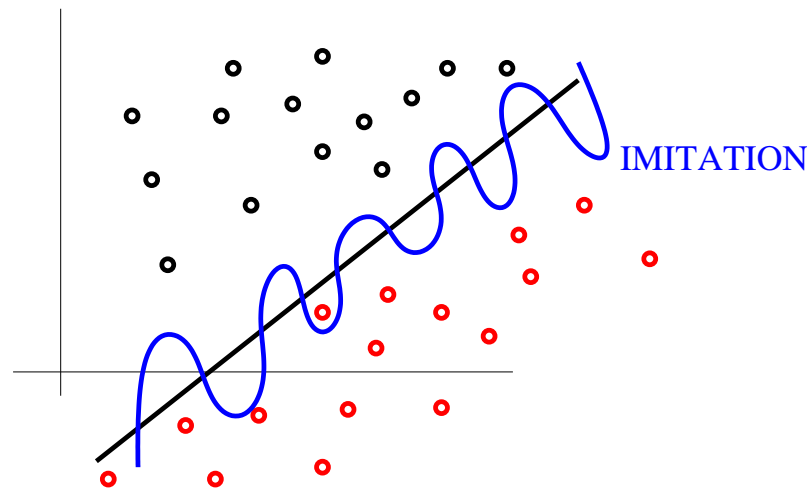
- In **identification** mode of inference the goal is to find α_ℓ such that

$$\|\alpha_\ell - \alpha_0\| \leq \varepsilon, \quad (\|f(x, \alpha_\ell) - f(x, \alpha_0)\| \leq \varepsilon).$$

- In **imitation** mode, the goal is to find α_ℓ such that

$$R(\alpha_\ell) - R(\alpha_0) \leq \varepsilon.$$

- The difference between these modes is shown in the figure:



TWO AND ONLY TWO FACTORS DEFINE PREDICTIVE GENERALIZATION

The main discovery of VC theory is that:

- **Two and only two** factors are responsible for generalization:
 - One (empirical loss) defines how well the function explains data.
 - Another (capacity, e.g. VC entropy or VC dimension) defines the diversity of the set of functions from which one chooses an approximating function.
- If the VC dimension is finite then the uniform law of large number is valid and one can achieve a good generalization.

If it is not finite then for classification problem the generalization is impossible.

Consider a set of vectors

$$z_1, \dots, z_\ell. \quad (*)$$

There exist 2^ℓ different ways to divide this set into two subsets.

We say that vectors $(*)$ **can not falsify** a set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$ if all 2^ℓ separations of $(*)$ are possible by this set of indicators.

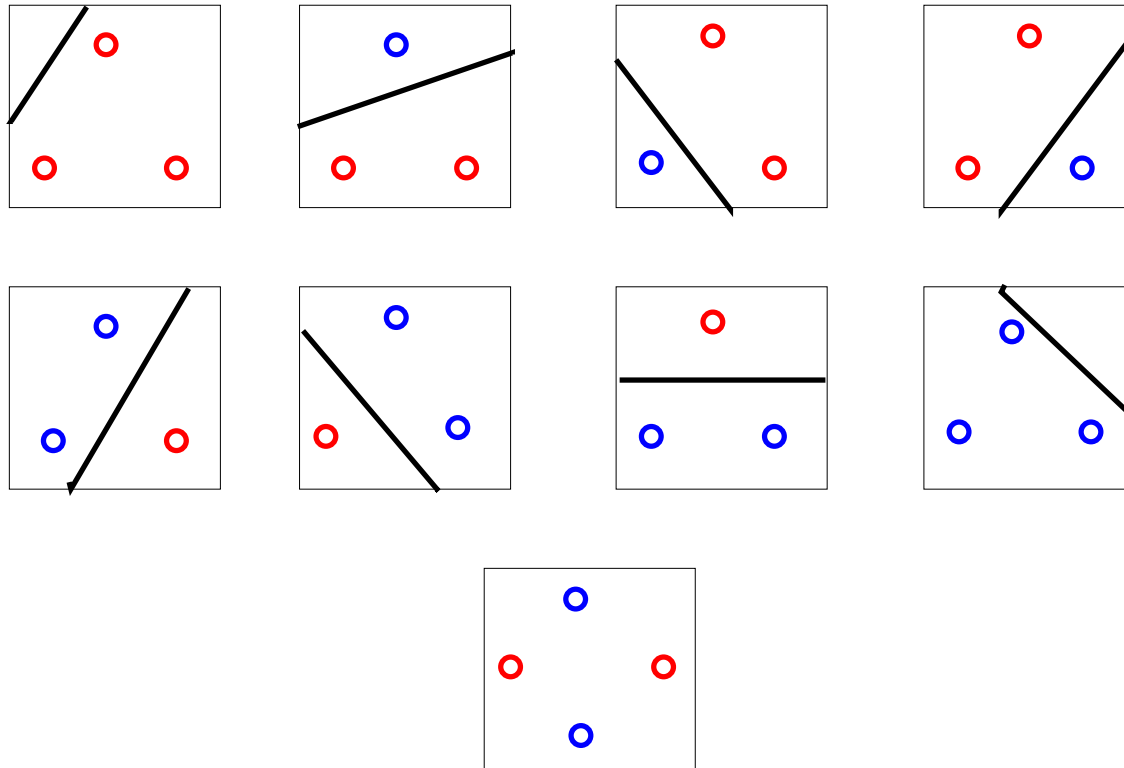
A set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$ has VC dimension h if:

- *There exist h vectors that **can not falsify** this set.*
- *Any $h + 1$ vectors **falsify** this set.*

The VC dimension of the set of indicator functions is equal to the largest number of examples that can not falsify this set.

EXAMPLE

The VC dimension of the set of lines on the plane equals 3.



Four examples can falsify any linear law.

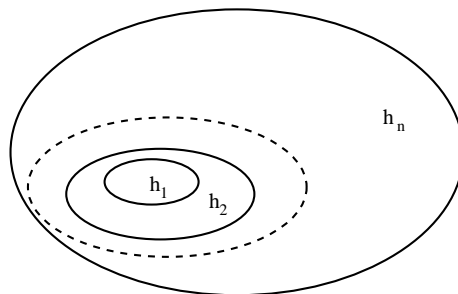
STRUCTURAL RISK MINIMIZATION PRINCIPLE

$$Probability(\text{test err}) \leq Frequency(\text{train err}) + \Phi\left(\frac{h_s}{\ell}\right), \quad (*)$$

The SRM principle requires to minimize r.h.s. of the bounds (*) over elements S_k of structure of a nested set of functions

$$S_1 \subset S_2 \subset \dots \subset S_n = S, \quad h_s = VC(S_n)$$

and functions in the element S_n .



Theorem

SRM 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 (parameters) than you need to explain the facts.

THE SRM PRINCIPLE

Explain facts using a function from the set with the smallest VC dimension.

Interpretation of SRM Principle

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

Does VC dimension describe the number of entities?

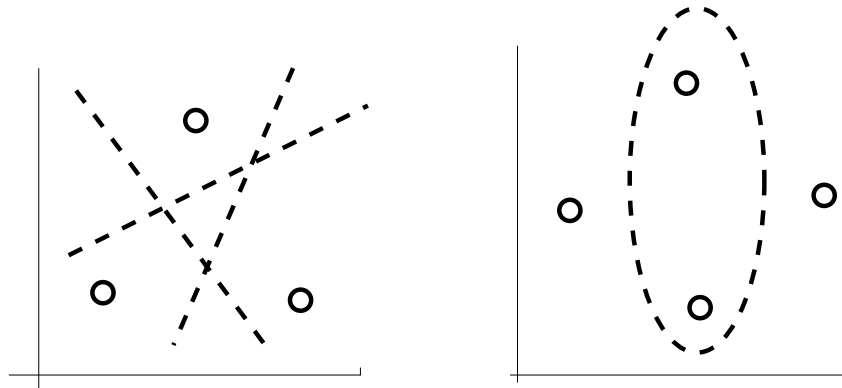
EXAMPLE. VC dimension is equal to number of entities (parameters) ¹⁶

The VC dimension of the set of linear indicator functions

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

is equal to the number of parameters

$$h = n + 1.$$

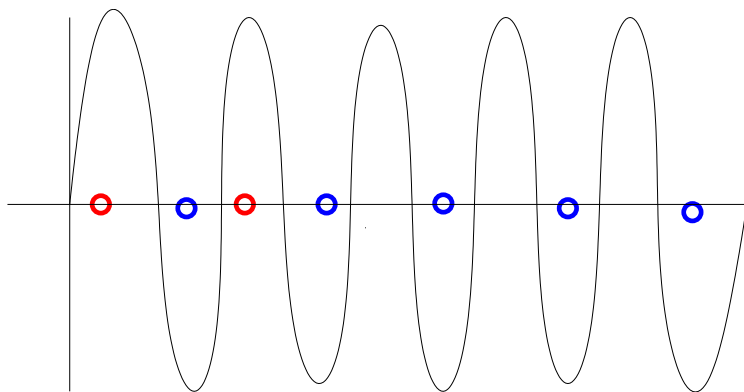


EXAMPLE. VC dimension is larger than the number of entities (parameters)

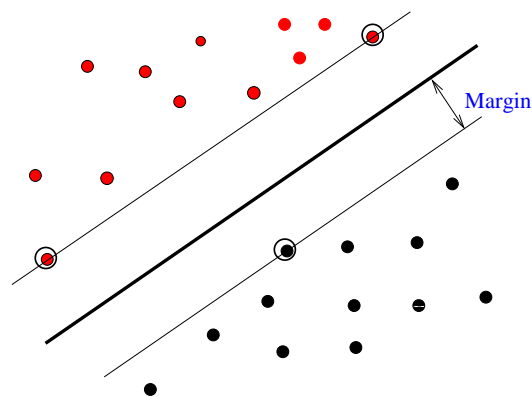
The VC dimension of the set of functions

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

is infinite.



EXAMPLE. VC dimension is less than the number¹⁸ of entities (parameters)



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

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

1992 – 1996

Large Margin Technology (SVMs)

THE IDEA OF SUPPORT VECTOR MACHINES²⁰

- **Increase the number of entities:**

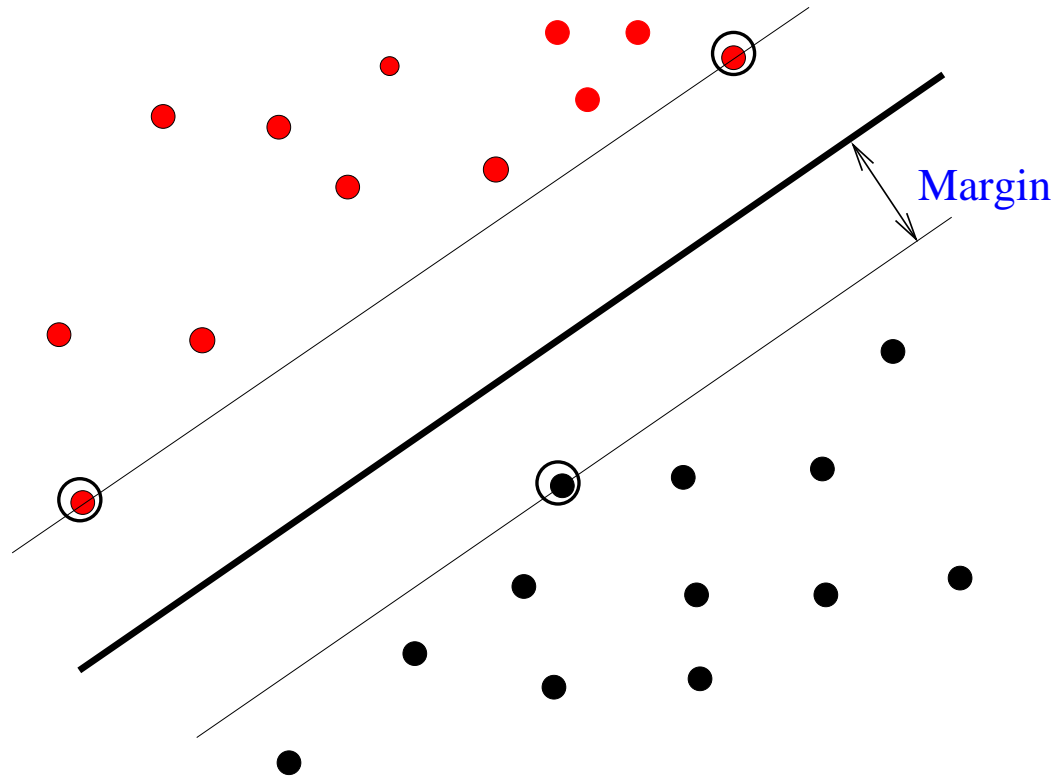
Map the input vectors into a high dimensional (or Hilbert) feature space.

- **Use SRM principle to control generalization:**

Control the VC dimension in high dimensional (feature) spaces constructing a hyperplane with a large margin.

The idea is that with increasing 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.

OPTIMAL SEPARATING HYPERPLANE



$$\text{Probability}(\text{test err}) \leq \text{Frequency}(\text{train err}) + \Phi\left(\frac{h_s}{\ell}\right), \quad (*)$$

Map $x \in X$ into $z \in Z$ such that $|z| \leq 1$. Then Δ -margin hyperplane

$$(w, z) + b = 0$$

belongs to a set with VC dimension bounded by h^* if

$$\Delta^{-2} = (w, w) \leq h^*, \quad (**)$$

Therefore to minimize r.h.s. (*) one has to minimize the functional

$$R = \sum_{i=1}^{\ell} \xi_i$$

subject to constraint (**) and constraints

$$y_i((z_i, w) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, \ell$$

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0.$$

Let us solve the equivalent problem:

Minimize the functional

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

subject to constraints

$$y_i((z_i, w) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \quad i = 1, \dots, \ell$$

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0.$$

Optimal separating hyperplane in feature space has a form

$$\sum_{i=1}^{\ell} y_i \alpha_i^0(z_i, z) + b = 0.$$

To find coefficients α^0 one has to maximize the functional

$$R = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j(z_i, z_j)$$

subject to constraints

$$0 \leq \alpha_i \leq C,$$

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0$$

Finding optimal separating hyperplane in a feature space requires solving simple (box constrained) **QP** problem.

Let vectors $x \in X$ are mapping into elements $z \in Z$ of a Hilbert space H

$$x \longrightarrow z \quad (*)$$

and let (z_i, z_j) be an inner product of elements z_i and z_j of space H .

- *For any mapping*

$$x \longrightarrow z \quad (*)$$

there exist positive definite (PD) function $K(x, x_)$ such that*

$$(z_i, z_j) = K(x_i, x_j). \quad (**)$$

- *For any PD function $K(x_i, x_j)$ there exists a mapping $(*)$ that $(**)$ holds.*

Optimal hyperplane in feature space has a form

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x) + b = 0.$$

To find coefficients α_i^0 one has to maximize the functional

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

subject to constraints

$$0 \leq \alpha_i \leq C, \quad i = 1, \dots, \ell$$

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0$$

- They always converge to best possible solutions.
- They minimize guarantee bounds for finite number of observations.
- They have a standard way to incorporate singularities of real-life problems using appropriate similarity measure $K(x_i, x_j)$ between two vectors x_i, x_j .
- The similarity measure can be defined for non-vectorial data (e.g. chemical formulas, or poetry texts, or political situations).
- They have universal generalization engine (simple **QP** solver).
- They construct non-linear decision rules using technologies of linear analysis.

Combination of these properties is unique in applied analysis.

The problem: Given i.i.d. data

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

find the optimal decision rule.

The non-parametric statistics solution for Parzen kernels is:

$$f(x) = \frac{1}{\ell_1} \sum_{\{i: y_i=1\}} K_\delta(x, x_i) - \frac{1}{\ell_2} \sum_{\{j: y_j=-1\}} K_\delta(x, x_j)$$

The Support Vector Machine solution for Mercer kernels is:

$$f(x) = \sum_{\{i: y_i=1\}} \lambda_i K_\delta(x, x_i) - \sum_{\{j: y_j=-1\}} \lambda_j K_\delta(x, x_j), \quad \lambda \geq 0$$

Geometrical (Mercer) interpretation in a feature space:

- Non-parametric solution is the hyperplane defined by the vector connecting the two centers of mass.
- The Support Vector solution is the optimal hyperplane.

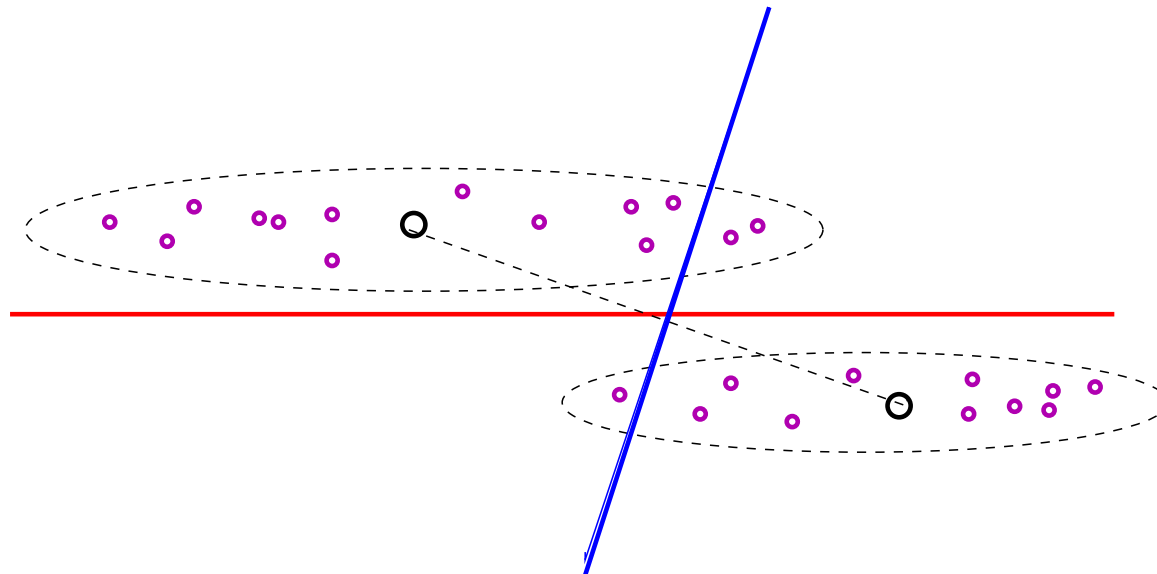
NON-PARAMETRIC METHODS AND THE SVM²⁹

(Illustration)

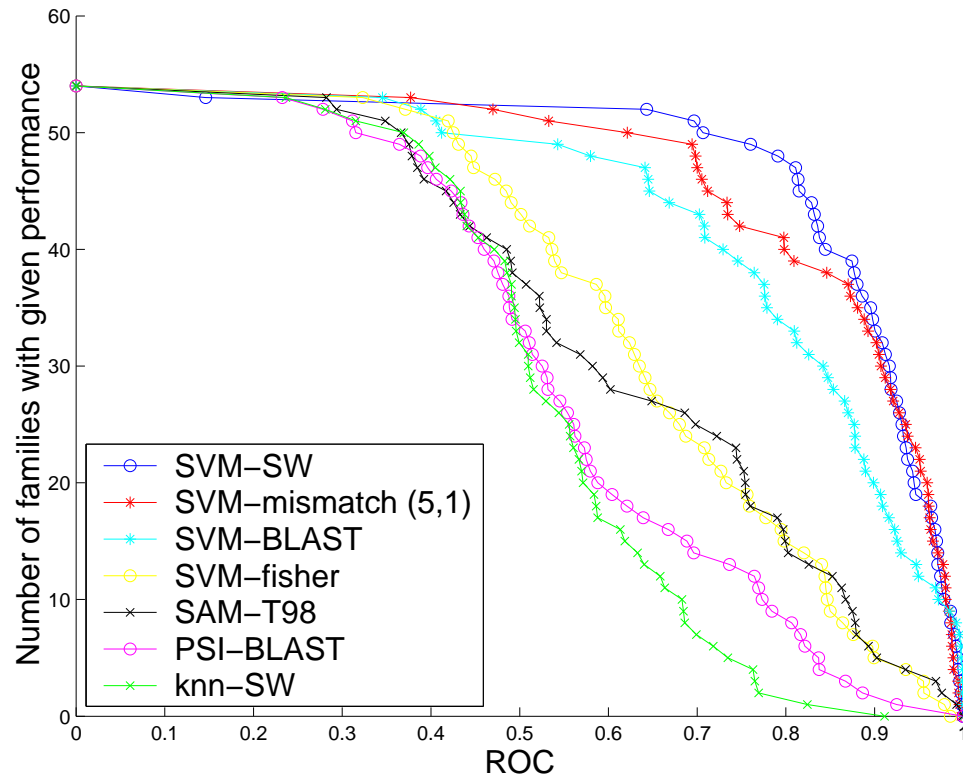
If $K_\delta(x, x_i)$ is a Mercer kernel then there exist mapping X -space into U -space such that

$$K_\delta(x, x_i) = (u, u_i),$$

In U -space both methods construct a hyperplane.



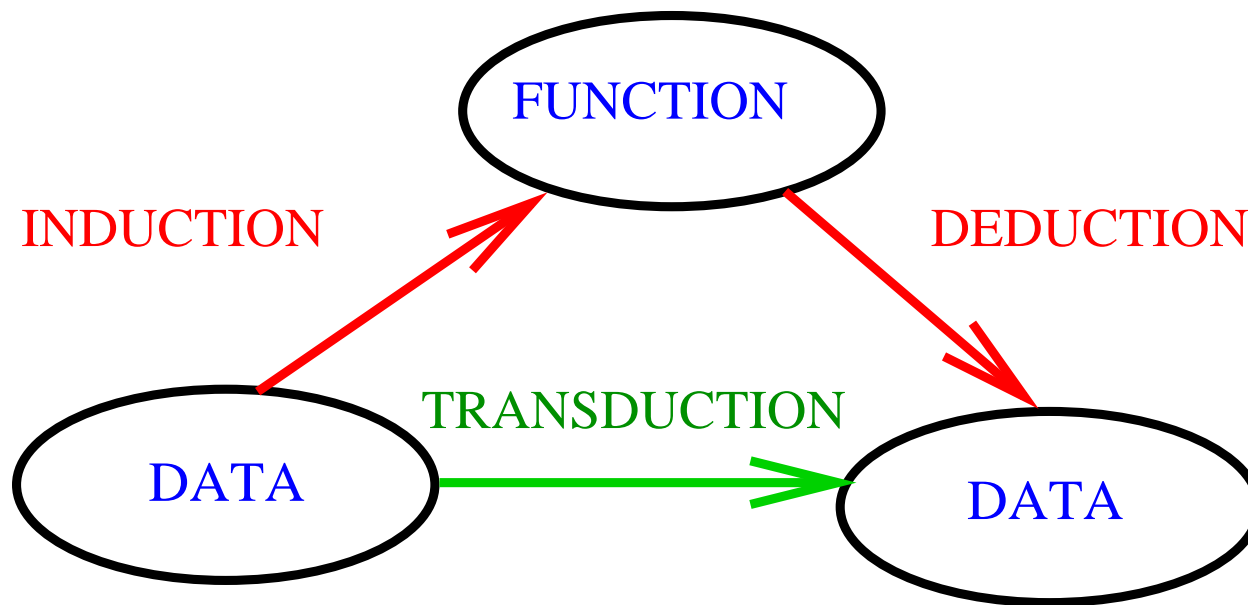
EXAMPLE: PROTEINS CLASSIFICATION



1996 – 2000

Creation of Non-Inductive Methods of Inferences

INDUCTIVE AND TRANSDUCTIVE INFERENCE



Given a set of training data

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

and given a set of test data

$$x_1^*, \dots, x_k^*$$

find among admissible set of classification vectors

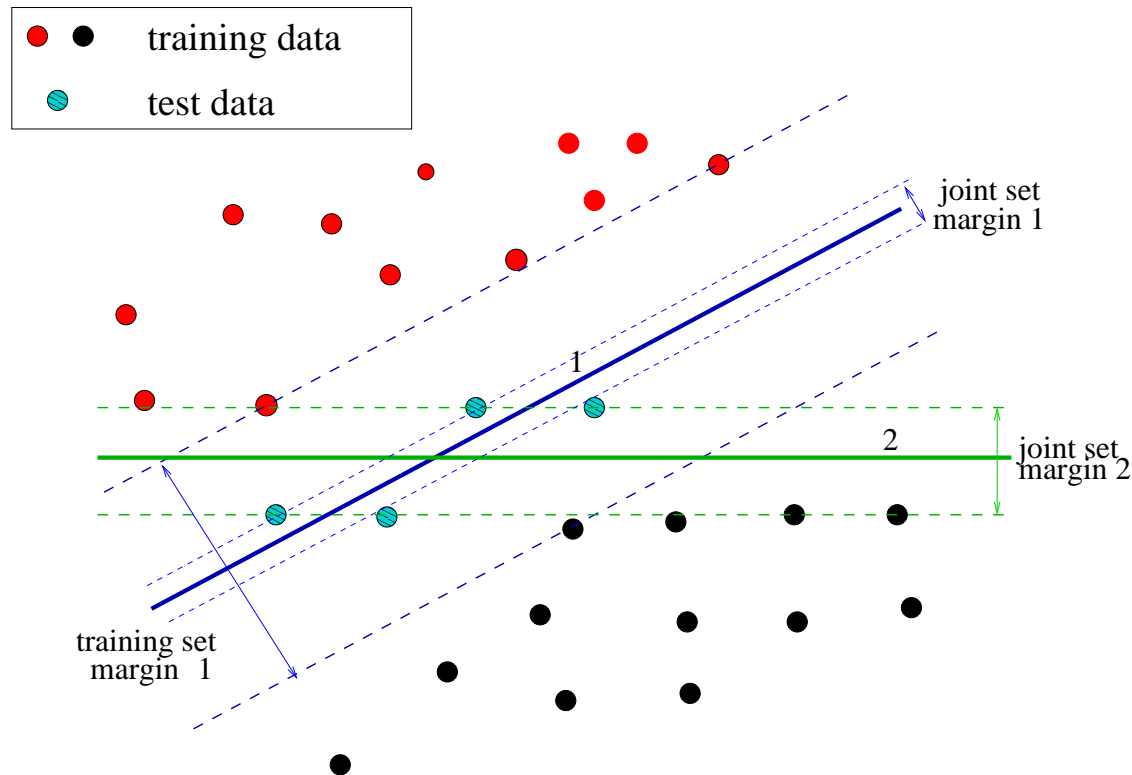
$$Y^* \in \{Y^* : (y_1^*, \dots, y_k^*)\}$$

the best classification vector.

TREE APPROACHES TO TRANSDUCTIVE INFERENCE

1. Inference based on size of margin.
2. Inference based on number of contradictions on Universum.
3. Inference based on value of VC entropy on Universum.

1. INFERENCE BASED ON SIZE OF MARGIN 35



Classify test data by hyperplane that separates training data and has the largest margin on the joint set of training and test data.

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 1909 examples: 42 (2.2%) of vectors belong to the first class (which bind), 1867 (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 percentage of correct classifications of the positive and negative examples.

PREDICTION OF MOLECULAR BIOACTIVITY³⁷

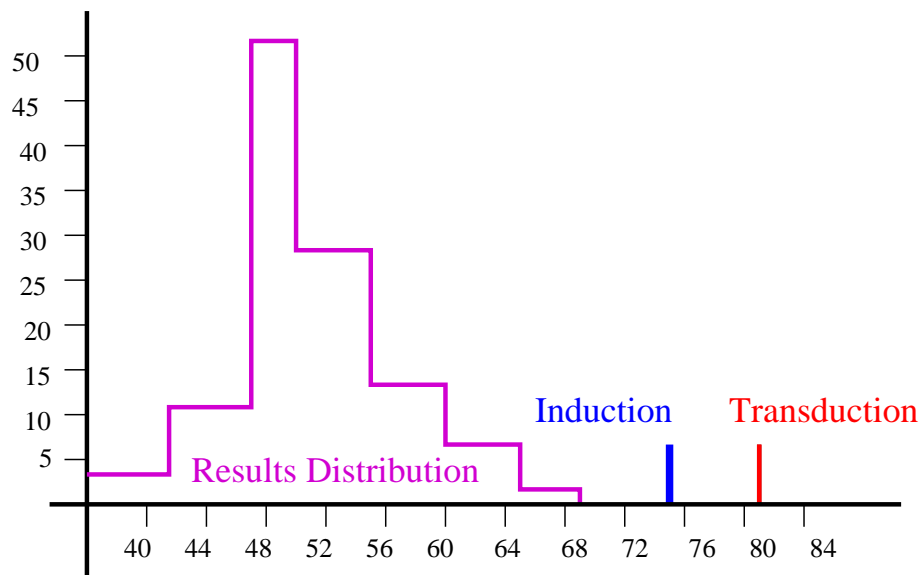
RESULTS OF COMPETITION: Winner's score was 68%.

SVM scores:

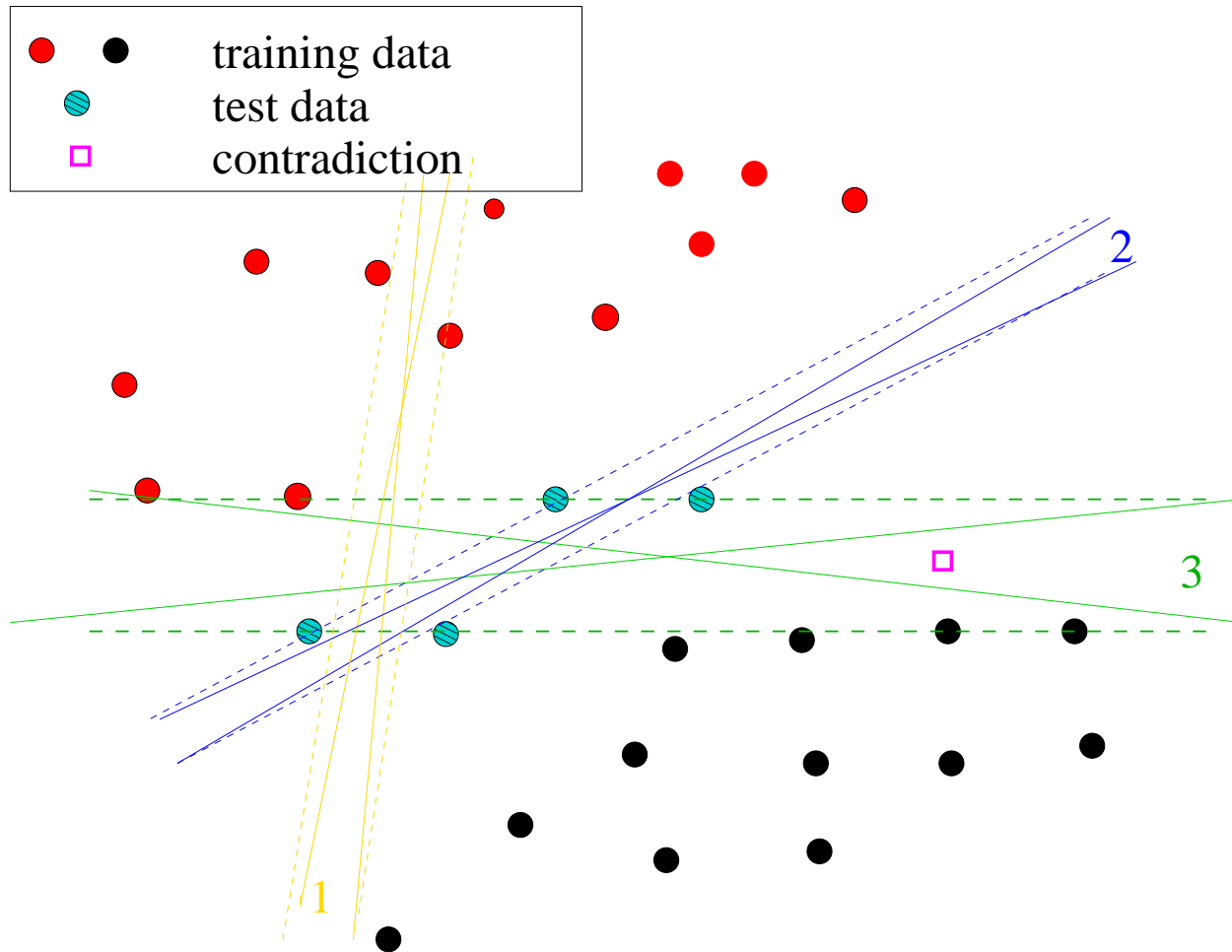
For **inductive** inference (using training data only): **74.5%**.

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

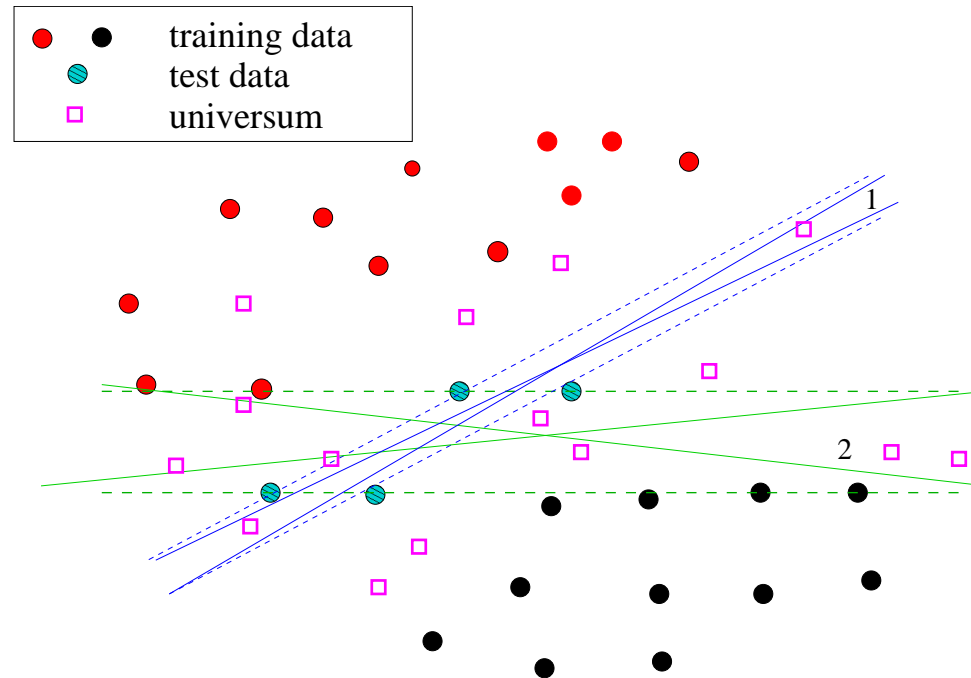
Comparison to other 119 entrants of the competition.



DEFINITION OF EQUIVALENCE CLASSES AND CONTRADICTIONARY VECTOR

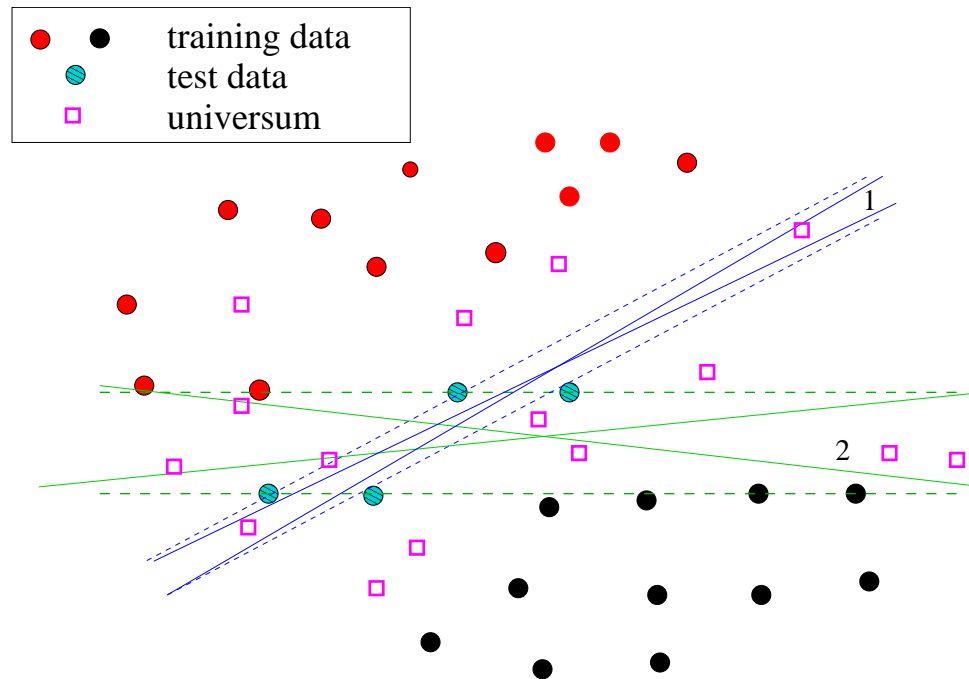


2. INFERENCE BASED ON THE NUMBER OF CONTRADICTIONS ³⁹



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

3. INFERENCE BASED ON THE VALUE OF THE VC ENTROPY



Classify test data by the equivalence class that separates training data and has the maximal VC entropy on Universum.

BEYOND TRANSDUCTION: SELECTIVE INFERENCE

Given ℓ training examples

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

and n candidate vectors

$$x_1^*, \dots, x_n^*$$

select among n candidate these k vectors with the highest probability of belonging to the first class.

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

National security: Among 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.

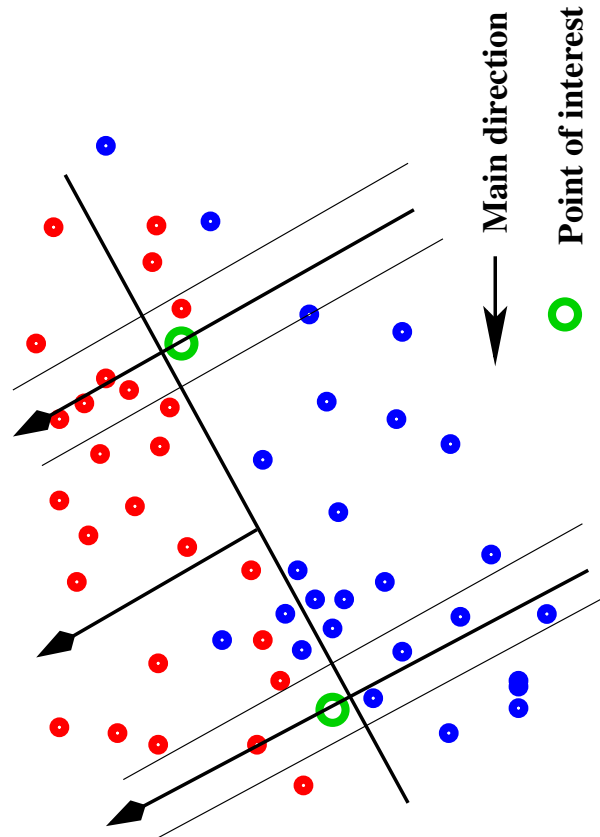
2005 – ...

Development of Directed Ad-Hoc Inference (DAHI)

1. During the training stage , DAHI looks for one principal direction (concept) used to construct different rules for future inferences.
2. During the test stage DAHI uses this principal direction to construct a specific rule for each given test vector (the ad-hoc rule). It constructs one specific rule for each specific example.

Principal direction is a unit vector ν that provides the property: The larger value of the projection $\xi = (\nu, z)$, the higher probability that vector z belongs to the positive class. Principal direction is not unique.

ALGORITHMIC IDEA OF DAHI



ESTIMATING CONDITIONAL PROBABILITY ALONG THE LINE

To estimate conditional probability $P(y = 1|\xi, z_0)$ along the line in R^n passing in the direction ν through point z_0 one must solve the equation

$$\int_a^\xi P(y = 1|\xi', z_0)dF(\xi'|z_0) = F(\xi, y|z_0), \quad \xi = (\nu, z_i), \quad y \in \{1, -1\}$$

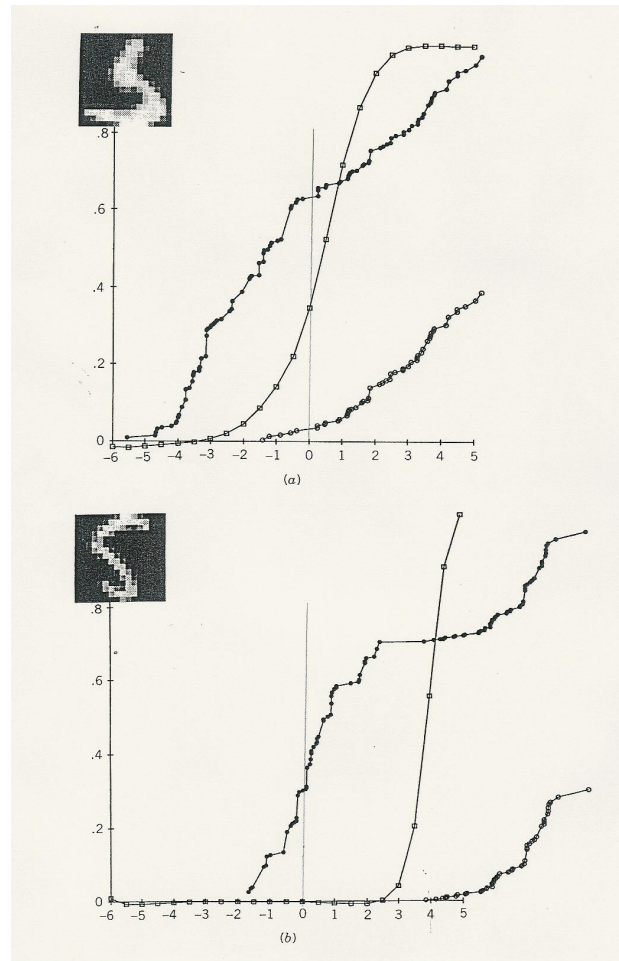
when functions $F(\xi'|z_0)$, $F(\xi, y|z_0)$ unknown but data

$$(z_1, y_1), \dots, (z_\ell, y_\ell)$$

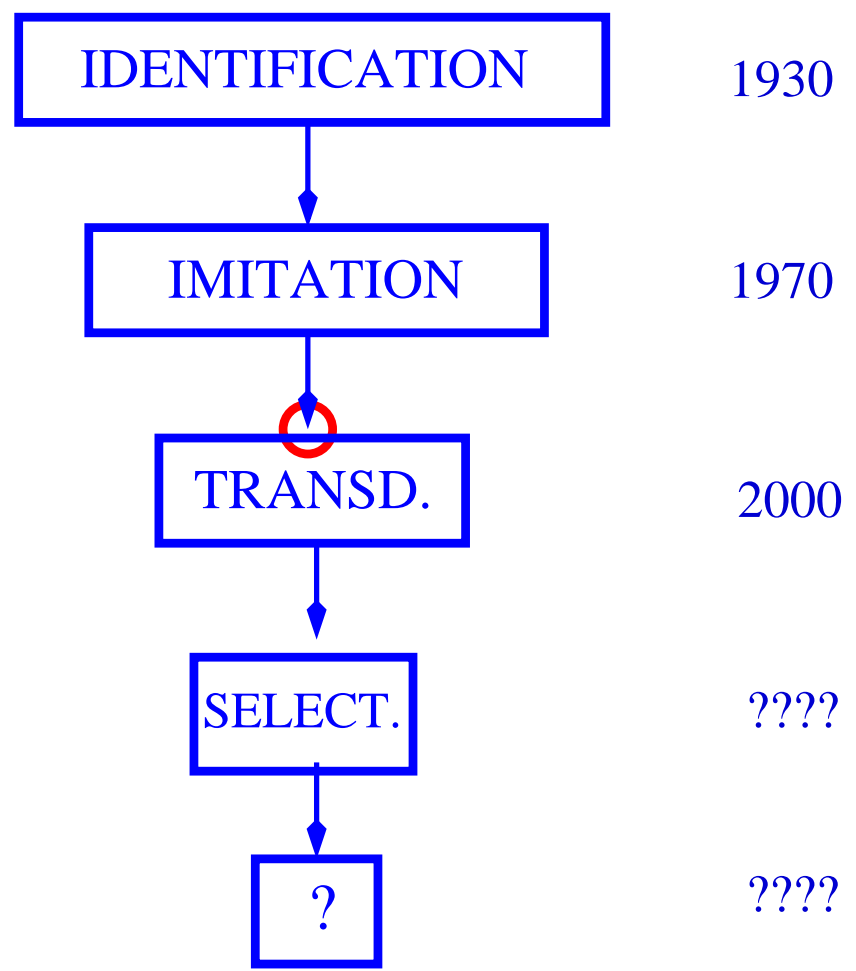
are given.

To do this we construct corresponding approximation to $F(\xi|z_0)$ and $F(\xi, y|z_0)$.

CONDITIONAL PROBABILITIES OF DIGIT 3 GIVEN IMAGE



BIG PICTURE: TRADING AMBITIONS FOR DIMENSIONALITY



THE GREAT 1930s

1. A. Kolmogorov introduced axiomatization of probability theory. It immediately connected the general problem of statistics to the analysis of **empirical (Glivenco-Cantelli) processes**.
2. K. Popper defined a demarcation between Metaphysical and Empirical Sciences based on the concept of **falsifiability** of theory.
3. R. Fisher introduced the paradigm of applied statistics as the idea of estimating a model of observed events. For model estimation, he suggested (**maximum likelihood**) method. He defined the key elements of a future theory of model (parameter) estimation:
 - sufficient statistics,
 - information matrix,
 - consistency and asymptotic normality,
 - efficiency.

THE GREAT 1960s

1. Tikhonov, Ivanov, and Phillips developed the main elements of the theory of ill-posed problems.
2. Kolmogorov and Tikhomirov developed capacity concepts (ε -entropy, covering numbers, width) for sets of functions.
3. Solomonov, Kolmogorov, and Chaitin developed the concept of algorithmic complexity.
4. Vapnik and Chervonenkis developed basics of Empirical Inference Science.
5. The empirical inference problem became a problem of Natural Science.

THE GREAT 1990s

1. Necessary and sufficient conditions for consistency of the empirical risk minimization principle were discovered.
2. Estimation of high dimensional functions became an actual problem.
3. Large margin methods based on the VC theory of generalization (SVM, Boosting, Neural Networks) prove advantageous over classical statistics methods.

THE GREAT 2000s

1. The problem of Transductive and Ad-Hoc inference have become hot topics in Empirical Inference.
2. A new of generation of reseachers in computer learning: instead of practitioners that rely on the applied statistics paradigm, the new generation of reseachers with good theoretical background in VC theory.
3. In data mining competitions empirical inference, techniques based on VC theory dominate over classical statistics techniques.

THE GREAT 200?

1. Creating theory of non-inductive methods of inference and proof of advantage of non-inductive methods over inductive methods for a complex World..
2. Creating a new Philosophy of Science that considers non-inductive methods of inference as main elements of inference for a complex World.
3. Reconsideration of psychological and behavioral sciences based on transductive type of inference.
4. Reconsideration of goals and methods of pedagogical science: Teaching not just inductive inferences but also direct inference and its connections to human cultural Universum.

- At the end of the 1960s it became clear that **classical statistics is too restrictive.**
(It can not be applied to high dimensional problems.)
 - At the end of the 1990s it became clear that **the Occam Razor principle of induction is too restrictive.**
(Experiments with SVM, Boosting, and Neural Nets contradict it.)
 - At the beginning of the 2000s it becoming clear that **the classical model of Science is too restrictive.**
(It does not include Transductive and Ad-Hoc inferences which in high dimensional situations can be more accurate than inductive inference.)
 - At the beginning of the 2000s it became clear that **in creating a new philosophy of science the problem of empirical inference will play the same role that physics played in creating the old philosophy of science.**
-

I want to know God's thoughts ... the rest are details.

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.

A. Einstein.

- When solution is simple, God is answering.

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

A. Einstein

- If something is said not to be science, it does not mean that there is something wrong with it ... it just means that it is not a science.

R. Feynman

INTERPRETATION

- Classical Science is an instrument for a Simple World. When World is complex in most cases Classical Science fails.

- For a Complex World there are methods that do not belong to the Classical Science.

FIRST METAPHOR

Subtle is Lord, but malicious He is not.

A. Einstein

INTERPRETATION

Subtle is Lord — one can not understand His thoughts,
but malicious He is not — one can act well without understanding them.

SECOND METAPHOR

The Devil imitates God.

Definition of the Devil.

VC INTERPRETATION

Actions based on your understanding of God's thoughts can bring you to catastrophe.

THIRD METAPHOR

If God does exist then many things must be forbidden.

F. Dostoevsky.

INTERPRETATION

If a subtle and non-malicious God exists, then many ways of generalization must be forbidden. Subject of the new philosophy of science is to define a corresponding imperative (to define what should be forbidden). This philosophy determines the success of generalization in real life high dimensional problems.

THE IMPERATIVE FOR HIGH DIMENSIONAL EMPIRICAL INFERENCE ⁵⁹

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. (1995).

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 not induction.)
- Do not estimate predictive values if your goal is to act well.
(Good strategy of action not necessarily rely on good prediction.)