PROBLEMS OF EMPIRICAL INFERENCE SCIENCE

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

SUMMARY

1. In the 1960s, two thousand years old the **Empirical Inference** problem became (due to computers) a subject of **Natural Sciences**.

2. Empirical inference theory started in 1930s, when A. Kolmogorov introduced a general model of empirical inference called **Theoretical Statistics**.

3. At the same time, R. Fisher developed the alternative model called **Applied Statistics**. It requires **model (function)** estimation and suffers from the "curse of dimensionality".

4. In the 1970s, fundamentals of the theory of empirical inference, the VC theory, was developed. It requires **(risk) functional** estimation and overcomes the "curse of dimensionality".

5. In the 1990s, methods of empirical inference provided algorithms (e.g. SVMs, Boosting, Neural Nets) that can generalize in high dimensional spaces.

6. New real life problems (e.g. image understanding, information retrieval, microarray analysis) require analysis in high dimensional (10,000 – 1,000,000) spaces. Empirical inference methods can be used for such analysis.

DEDUCTIVE AND INDUCTIVE METHODS IN ANALYSIS OF RANDOM EVENTS

The core problem of Probability Theory:

Given a triplet (X, Ω, P) that defines model of possible random events, estimate a chance of appearance of the event of interest. (**The deductive inference**).

The core problem of Statistics:

Given a pair (X, Ω) and observation of random examples $x_1, ..., x_\ell,\ x_i \in X$, find the statistical law P (the probability model (X, Ω, P)) that generates these examples. (**The inductive inference).**

GLIVENCO-CANTELLI THEOREM

Let

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

be i.i.d. observations from distribution function

$$F(a) = P\{x \leq a\} = EI(x \leq a), \qquad I(x, a) \subset \{0, 1\}.$$

Consider the empirical distribution function

$$F_{emp}(a) = \frac{\text{card}(x_j : \ t_j \leq a)}{\ell} = \frac{1}{\ell} \sum_{i=1}^{\ell} I(x_i \leq a)$$

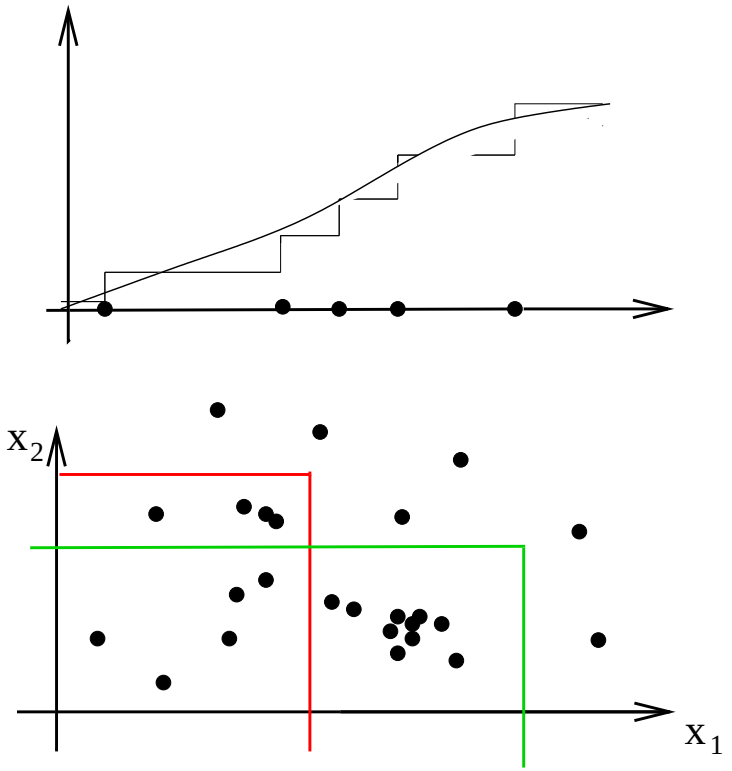
Glivenco-Cantelli Theorem:

$$\lim_{\ell \longrightarrow \infty} P \left\{ \sup_a |F(a) - F_{emp}(a)| > \varepsilon \right\} = 0, \quad \forall \varepsilon > 0.$$

Kolmogorov's Inequality:

$$P \left\{ \sup_a |F(a) - F_{emp}(a)| > \varepsilon \right\} \leq 2 \exp \left\{ -2\varepsilon^2 \ell \right\}.$$

GLIVENCO-CANTELLI THEOREM (illustration)



FISHER'S SIMPLIFICATION — THE APPLIED STATISTICS

- Reduce the problem of estimating the unknown probability measure to the problem of estimating unknown parameters of a density function.
- Use appropriate models to define a parametric family of densities that contain the desired function.
- Use the Maximum Likelihood method to estimate parameters of density.
- Construct efficient algorithms.

EXAMPLE (Regression estimation problem)

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

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

estimate the function $f_0(x)$.

The appropriate model:

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

$$y_i = f(x_i, \alpha_0) + \xi_i, \qquad Ex_i\xi_i = 0.$$

3. The law which defines the noise

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

is known. For example, it is the normal law

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

EXAMPLE (Regression estimation problem)

The idea of inductive inference:

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

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

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:

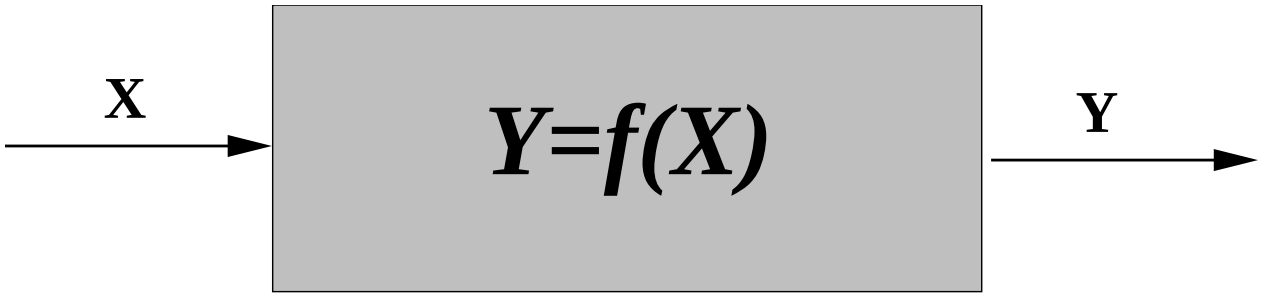
Quality of parameters estimation: consistency, asymptotic normality, efficiency.

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

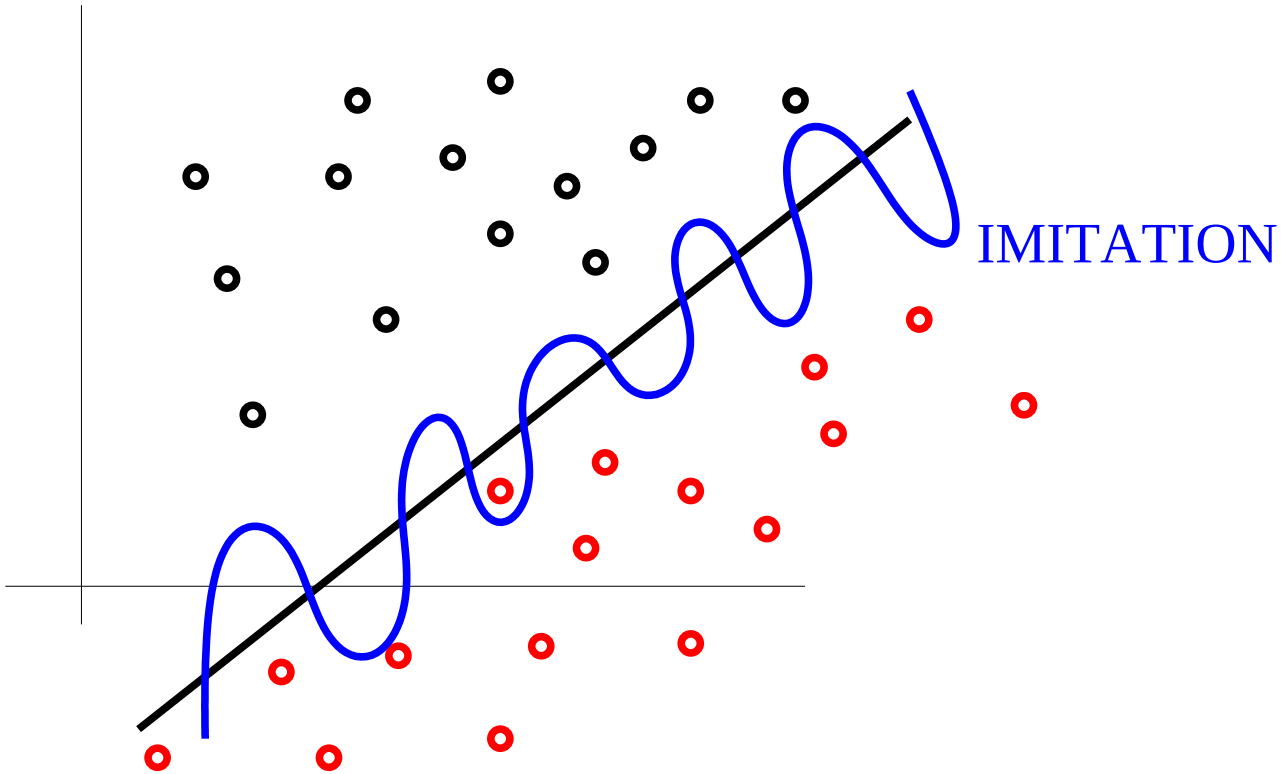
IDENTIFICATION OR IMITATION?



- The main goal of classical statistics is to **identify** the unknown function.
- The main goal of VC theory is to **imitate** the unknown function.

DIFFERENCE BETWEEN IDENTIFICATION AND IMITATION MODES OF INFERENCE

The difference between these paradigms is shown in the figure:



MINIMIZING THE RISK FUNCTIONAL USING DATA

In a given set of functions

$$Q(z, \alpha), \quad \alpha \in \Lambda$$

find one that minimizes the functional

$$R(\alpha) = \int Q(z, \alpha) dP(z)$$

if the probability measure $P(z)$ is unknown but i.i.d. data

$$z_1, ..., z_\ell$$

obtain from $P(z)$ are given.

PARTICULAR PROBLEMS

The following problems are particular cases of the general problem of minimizing a risk functional on the basis of empirical data.

Pattern Recognition:

The observations z are $z = y, x, \quad y \subset \{-1, 1\}, \ x \in X^d$.
The risk function takes the form

$$Q(z, \alpha) = \frac{1}{2}|y - I(x, \alpha)|, \quad I(x, \alpha) \subset \{-1, 1\}.$$

Regression Estimation:

The observations z are $z = y, x, \quad y \in R^1, \ x \in R^d$.
The risk function takes the form

$$Q(z, \alpha) = (y - f(x, \alpha))^2, \quad f(x, \alpha) \in R^1.$$

Density Estimation:

The observations z are $z = x$. The risk function takes the form

$$Q(z, \alpha) = -\ln P(x, \alpha), \quad P(x) \geq 0, \quad \int P(x)dx = 1$$

THE EMPIRICAL RISK MINIMIZATION PRINCIPLE

THE GOAL IS:

Find the function $Q(z, \alpha_0)$ in a set $Q(z, \alpha),\ \alpha \in \Lambda$ that minimizes the **risk functional**

$$R(\alpha) = \int Q(z, \alpha) dP(z), \quad \alpha \in \Lambda$$

if $P(z)$ is unknown but we are given data $z_1, ..., z_\ell$.

THE METHOD IS:

Find the function $Q(z, \alpha_\ell)$ that minimizes the **empirical risk functional**

$$R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha).$$

The problem of the theory is to answer the following questions:

- When the Empirical Risk Minimization (ERM) method is consistent?
- How well does the empirical risk approximate the expected risk?

THE ERM METHOD FOR THE MAIN LEARNING PROBLEMS

THE ERM PRINCIPLE IMPLIES:

- **For Pattern Recognition:** methods that minimize the number of training errors

$$\alpha_\ell = \arg\min_\alpha \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{|y_i - I(x_i,\alpha)|}{2}, \quad y \subset \{-1,1\},\ I(\boldsymbol{x},\alpha) \subset \{-1,1\}.$$

- **For regression estimation:** the least squares method

$$\alpha_\ell = \arg\min_\alpha \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - f(x_i,\alpha))^2 \quad y \in R^1,\ f(x,\alpha) \in R^1.$$

- **For density estimation:** the maximum likelihood method

$$\alpha_\ell = \arg\min_\alpha \frac{1}{\ell}\sum_{i=1}^{\ell} -\ln P(x_i), \quad P(x) \geq 0, \quad \int P(x)dx = 1.$$

DIFFERENCE BETWEEN APPLIED STATISTICAL AND EMPIRICAL INFERENCES

- In classical applied statistics the goal was to find α_ℓ such that

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

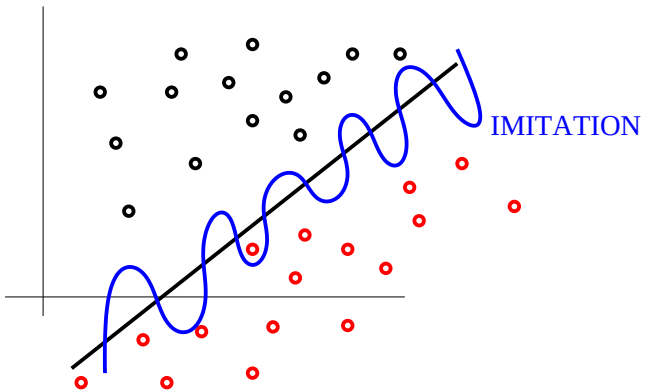
  That is **to find the model of a random process**.

- In Empirical inference, the goal is to find α_ℓ such that

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

  That is **to predict outcomes of a random process**.

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



THE PROBLEM OF EMPIRICAL INFERENCE

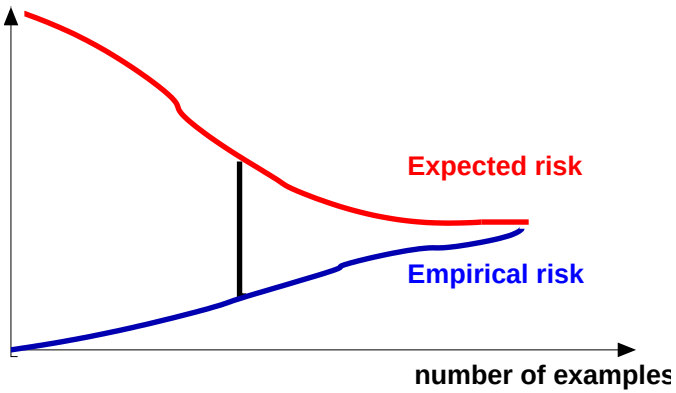
Assume that the function $Q(z, \alpha_\ell)$ minimizes the empirical risk

$$\alpha_\ell = \alpha_\ell(x_1, ..., x_\ell) = \arg\min_\alpha \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha).$$

Consider two functions

$$R(\alpha_\ell) = \int Q(z, \alpha_\ell) dP(z),$$

$$R_{emp}(\alpha_\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_\ell).$$



THE KEY THEOREM OF EMPIRICAL INFERENCE THEORY

Theorem (VC, 1989):

Let $Q(z,\alpha), \alpha \in \Lambda$ be a set of functions that satisfy the condition

$$A \leq \int Q(z,\alpha) dP(z) \leq B, \qquad (A \leq R(\alpha) \leq B).$$

Then the ERM method is consistent if and only if the uniform convergence

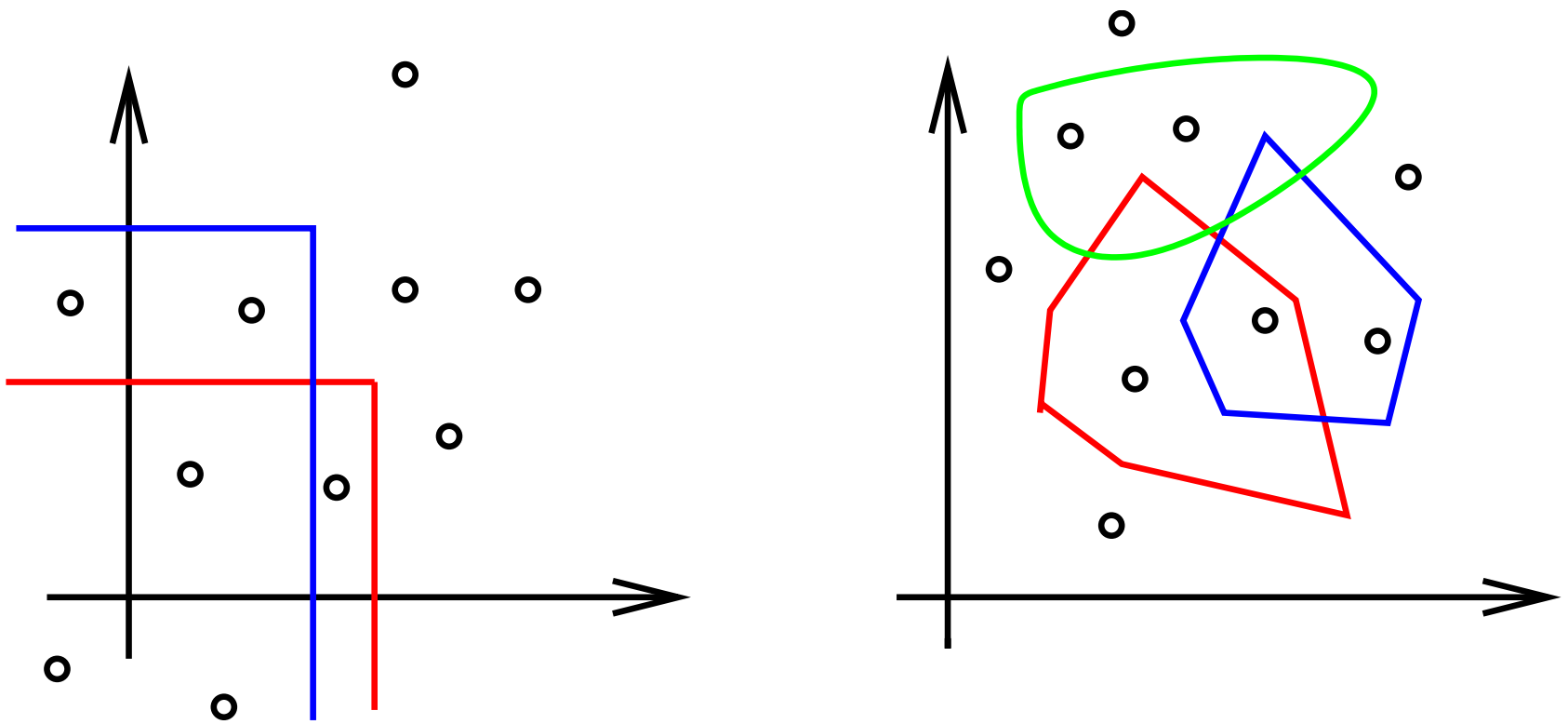
$$\lim_{\ell \longrightarrow \infty} P\left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) \geq \varepsilon \right\} = 0$$

takes place, where

$$R(\alpha) = \int Q(z,\alpha) dP(z),$$

$$R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i,\alpha).$$

EXTENSION OF GLIVENKO-CANTELLI THEORY



PROBLEMS OF THE THEORY OF EMPIRICAL PROCESSES

TWO MAIN PROBLEMS OF EMPIRICAL INFERENCE THEORY

1. Find the conditions when the empirical process is consistent

$$\lim_{\ell \longrightarrow \infty} P\left\{\sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| > \varepsilon\right\} = 0, \quad \forall \varepsilon > 0.$$

2. Find the non-asymptotic rate of convergence of the empirical process (if the empirical process is consistent)

$$P\left\{\sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| > \varepsilon\right\} \leq r(\varepsilon, \ell, \cdot),$$

were $\cdot$ stands for the so-called capacity concepts that determine both the consistency and the rate of convergence of the Empirical Process.

THE VC ENTROPY AND THE GROWTH FUNCTION

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

$$z_1, ..., z_\ell$$

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

$$N = N^\Lambda(z_1, ..., z_\ell)$$

of different separations of on the sample by the set of indicator functions.

- We call the quantity

$$H_P^\Lambda(\ell) = E_{\{z_1,...,z_\ell\}} \log_2 N_P^\Lambda(z_1, ..., z_\ell)$$

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

- We call the quantity

$$G^\Lambda(\ell) = \max_{z_1,...,z_\ell} \log_2 N_P^\Lambda(z_1, ..., z_\ell)$$

the **Growth function**.

THE VC DIMENSION

Theorem. (VC, 1968,1971):
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.

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

COMBINATORIC DEFINITION OF THE VC DIMENSION

Consider the set of vectors

$$z_1, ..., z_\ell.$$

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

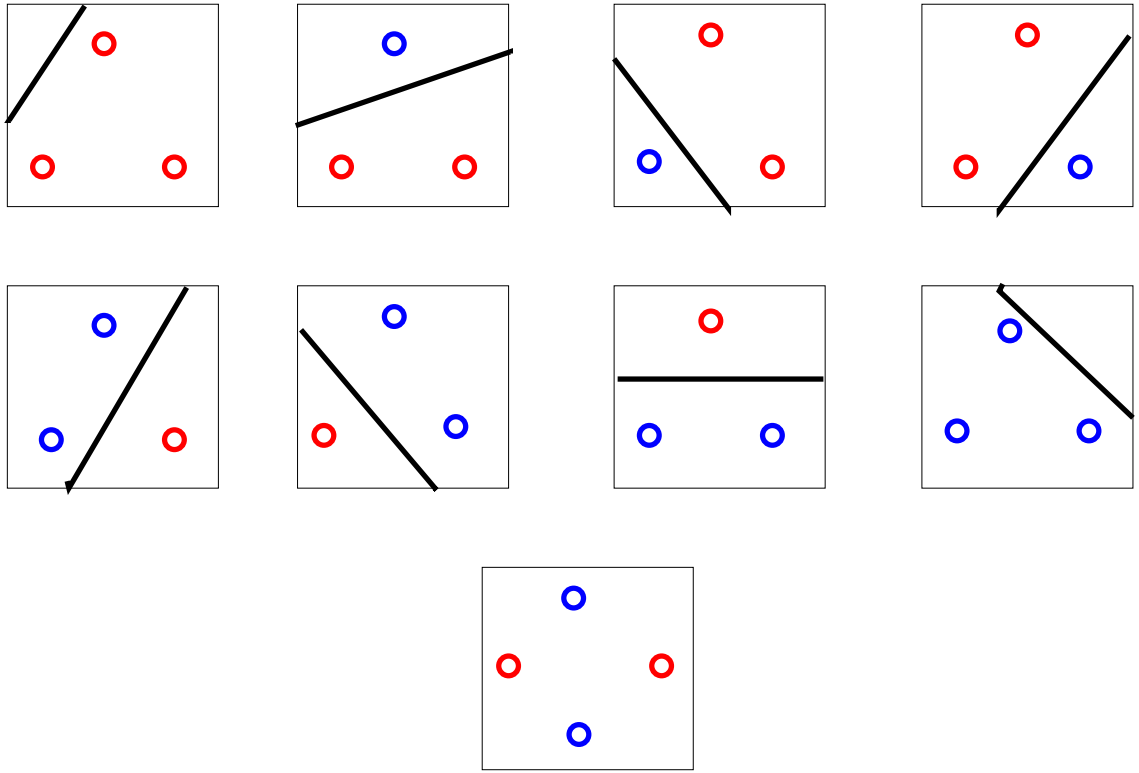
We say that the set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$ **shatters** these vectors if all 2^ℓ separations are possible using this set of indicators.

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

- **There exist h vectors that can be shattered using this set.**
- **There are no $h+1$ vectors that can be shattered using this set.**

EXAMPLE

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



Four examples can falsify any linear law.

TWO AND ONLY TWO FACTORS DEFINE GENERALIZATION

The key discovery of VC theory is that:

- **Two and only two** factors are responsible for generalization:
  – One (empirical loss) defines how well the function approximates data.
  – Another (capacity, e.g. VC dimension) defines the diversity of the set of functions from which one chooses an approximating function.
- If the VC dimension is finite, then one can achieve a good generalization. *If it is not finite then the generalization is impossible.*

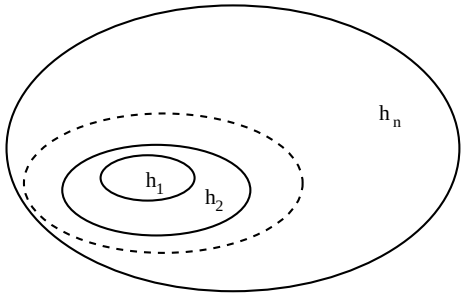
STRUCTURAL RISK MINIMIZATION PRINCIPLE

Using these two factors the VC theory gives the generalization bounds:

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

where ℓ is the number of training examples.

To minimize r.h.s of (*) one creates *a structure*



and minimizing r.h.s of (*) over both factors.

The 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.

THE CRUCIAL POINT

Does VC dimension describe the number of entities?

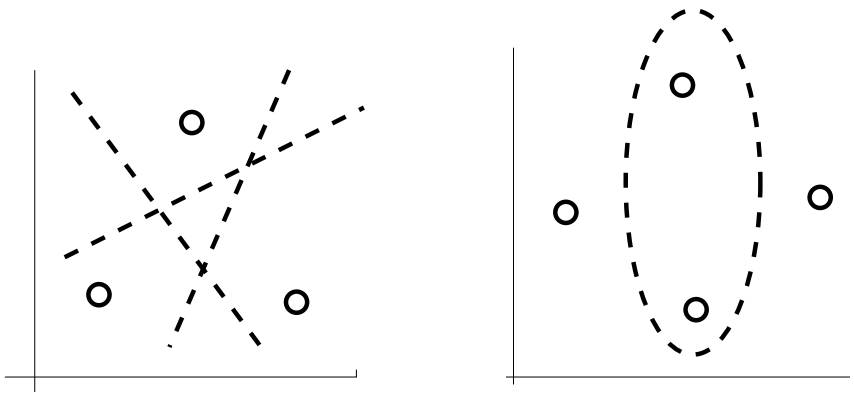
EXAMPLE. VC dimension is equal to number of entities (parameters)

The VC dimension 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

$$h = n + 1.$$

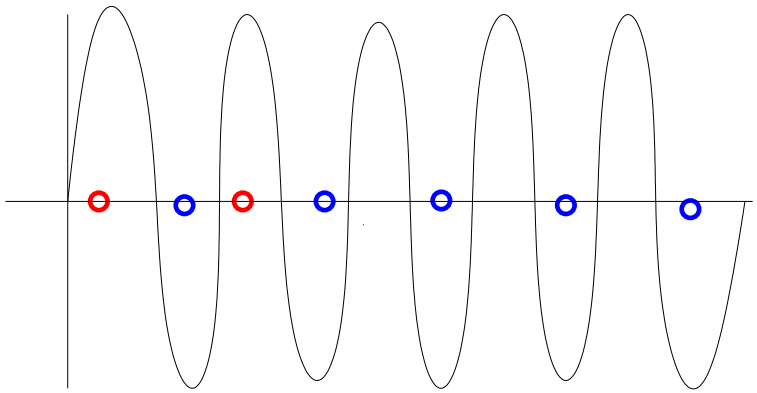


EXAMPLE. VC dimension is larger than the number of entities (parameters)

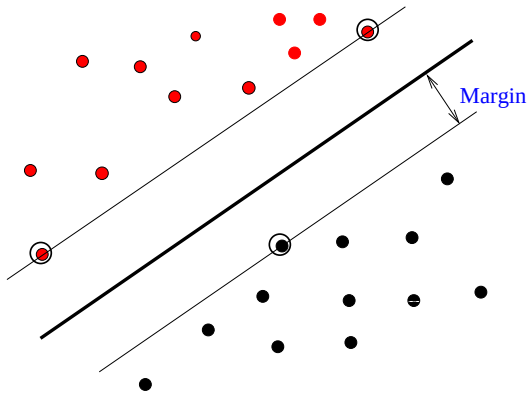
The VC dimension of the set of functions

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

is infinite.



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



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

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

PRESENT: 1992 – 2004

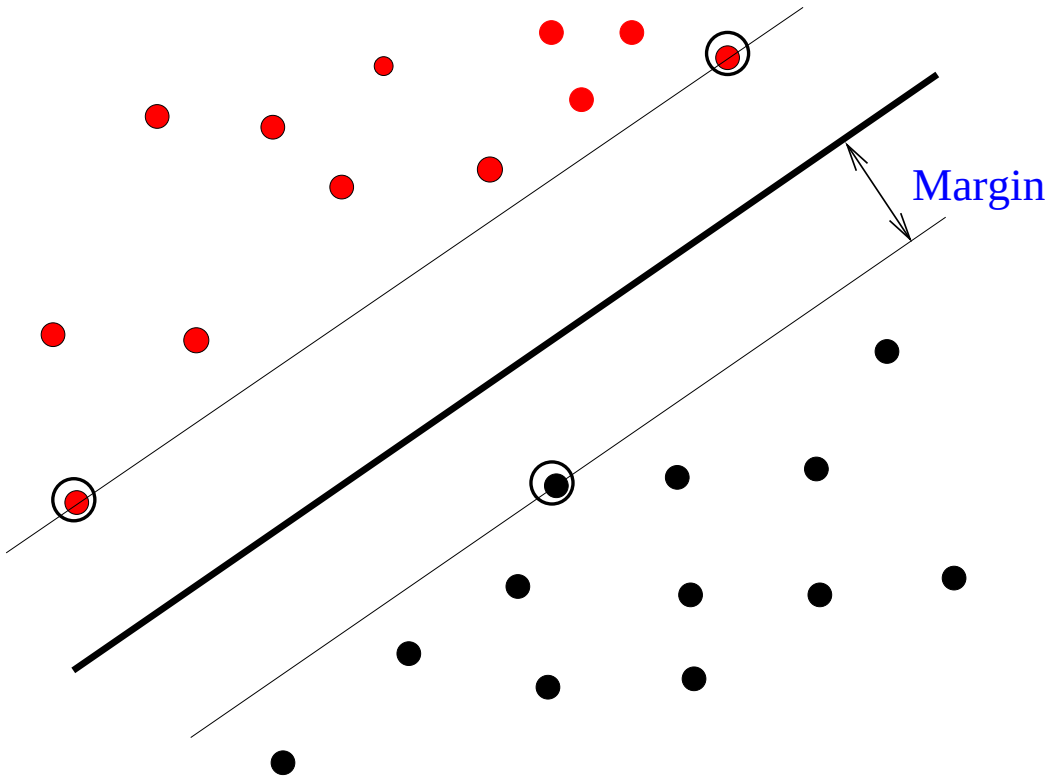
SVM technology

THE IDEA OF SUPPORT VECTOR MACHINES

- **Increase the number of entities:**
  Map the input vectors into a high dimensional (or Hilbert) feature space.
- **Control the VC dimension in high dimensional spaces:**
  Construct in the feature space 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



ALGORITHMS FOR CONSTRUCTING SVMs

- Map the input vectors x_i into a feature space z_i.

$$x_i \longrightarrow z_i$$

- Construct in feature space a hyperplane with a large margin

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

To find such a hyperplane one has to maximize the quadratic form

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

subject to constraints $\alpha_i \geq 0, \ i = 1, ..., \ell$

THE KERNEL TRICK

For any mapping

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

there exist *positive definite* (PD) function $K(x_i, x_j)$ such that

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

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

Therefore the optimal hyperplane in the *image space* has the form

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

where the coefficients α_i^0 are those that maximize

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

subject to constraints $\alpha_i \geq 0$.

SVM IS UNIVERSAL LEARNING METHOD

SVM introduced a universal method for solving learning problems:

• Mapping into Hilbert space and controlling capacity factor are key elements for many problems such as:

– Regression estimation.

– Operator estimation.

– Estimation of density support.

– Non-linear factor analysis, and so on

• One can map into Hilbert space not only vectors but also some abstract elements. Therefore SVM can be used for learning problems that have non-vectorian inputs. These inputs can be:

– Sequences of different size (bioinformatics and linguistics).

– Can belong to a space of chemical formulas.

The problem: Given i.i.d. data

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

find the optimal decision rule.

The non-parametric statistics solution for Parzen kernels is:

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

The Support Vector Machine solution for Mercer kernels is:

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

Geometrical interpretation in a feature space:

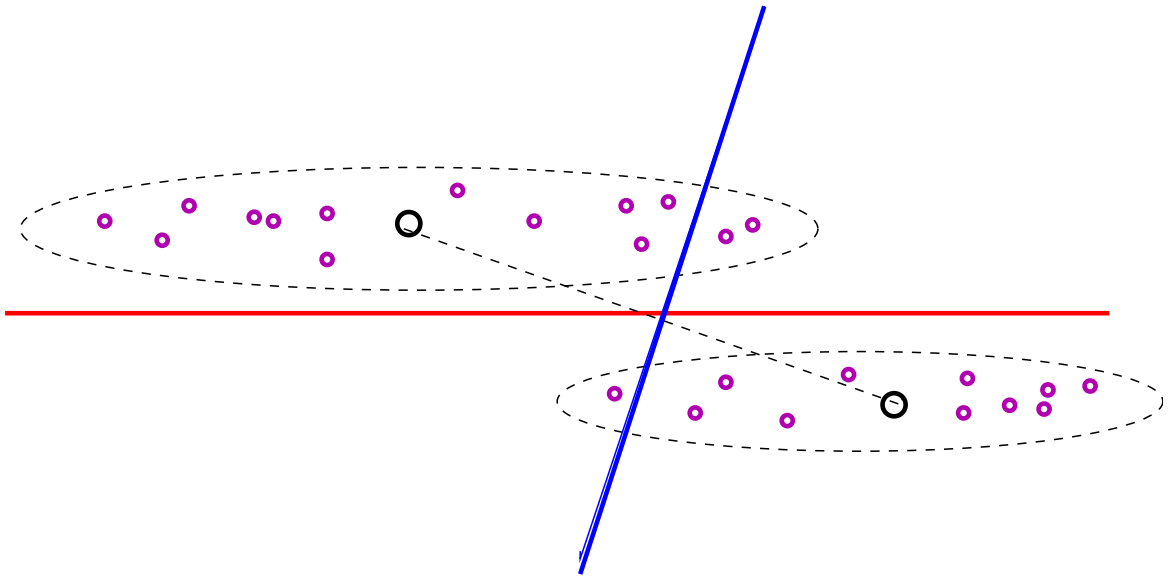
- Non-parametric solution is the hyperplane defined by the vector connecting the two centers of mass.
- The Support Vector solution is the optimal hyperplane.

NON-PARAMETRIC METHODS AND THE SVM (Illustration)

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

$$K(x, x_i) = (u, u_i),$$

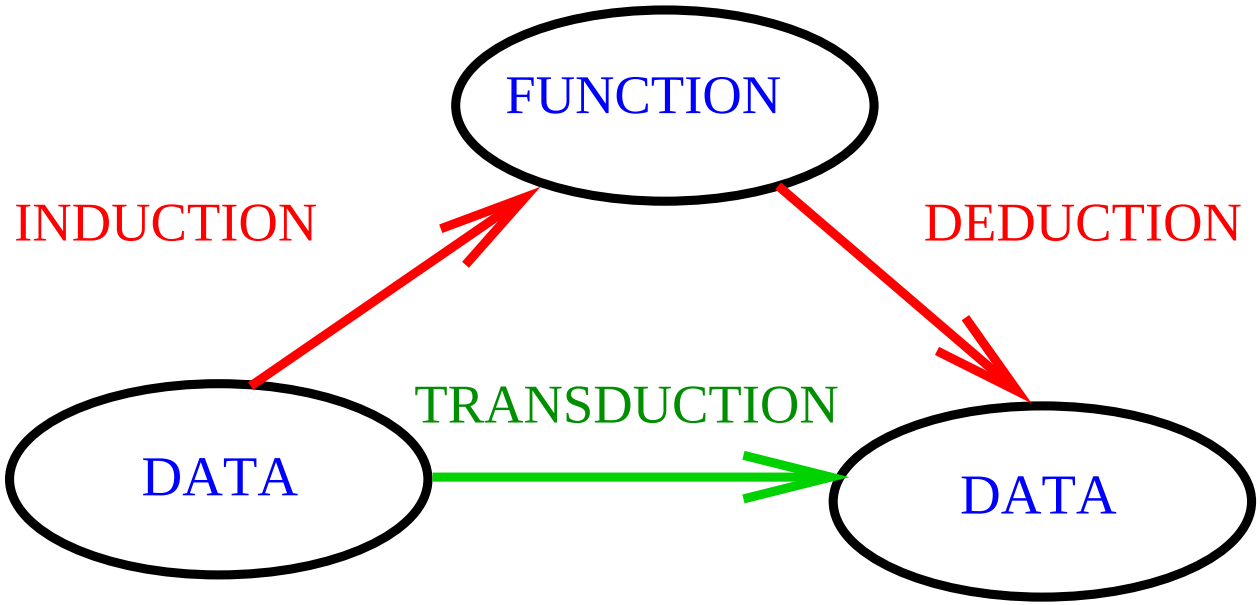
In U-space both methods construct a hyperplane.



FUTURE: 2005 – ...

Creation of non-inductive methods of inferences

INDUCTIVE AND TRANSDUCTIVE INFERENCES



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 admissible set of classification vectors

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

the best classification vector.

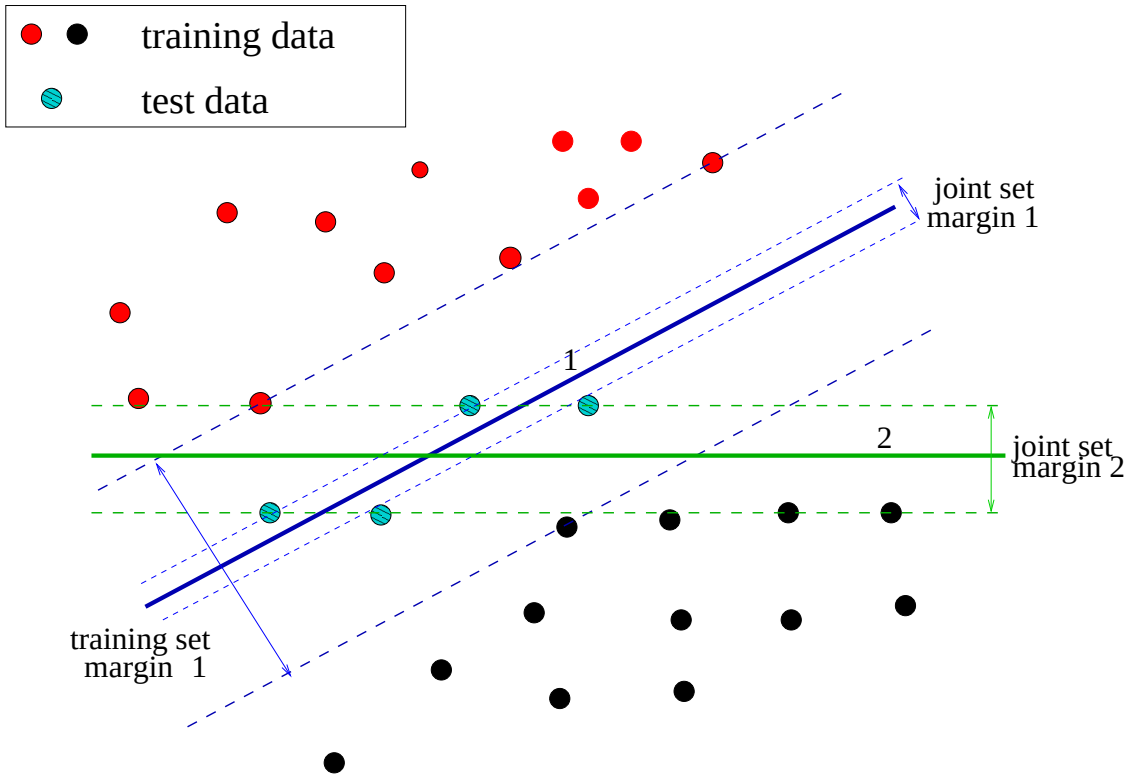
THREE APPROACHES TO TRANSDUCTIVE INFERENCE

We analyze three approaches to transductive inference:

- Inference based on size of margin
- Inference based on number of contradictions on Universum
- Inference based on similarity of sets of data

The problem is to construct a general theory of transduction

INFERENCE BASED ON SIZE OF MARGIN



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 BIOACTIVITY

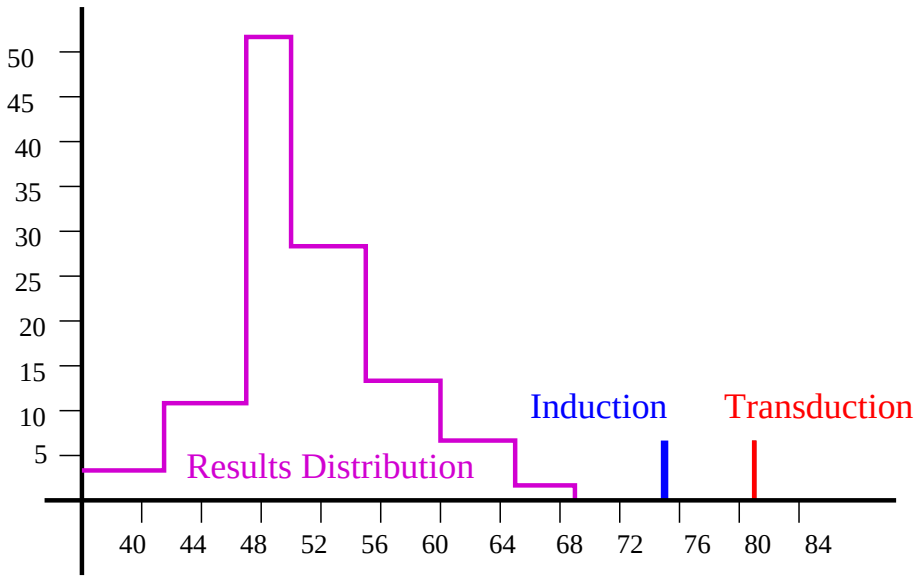
RESULTS OF COMPETITION: Winner's score was 68%.

SVM scores:

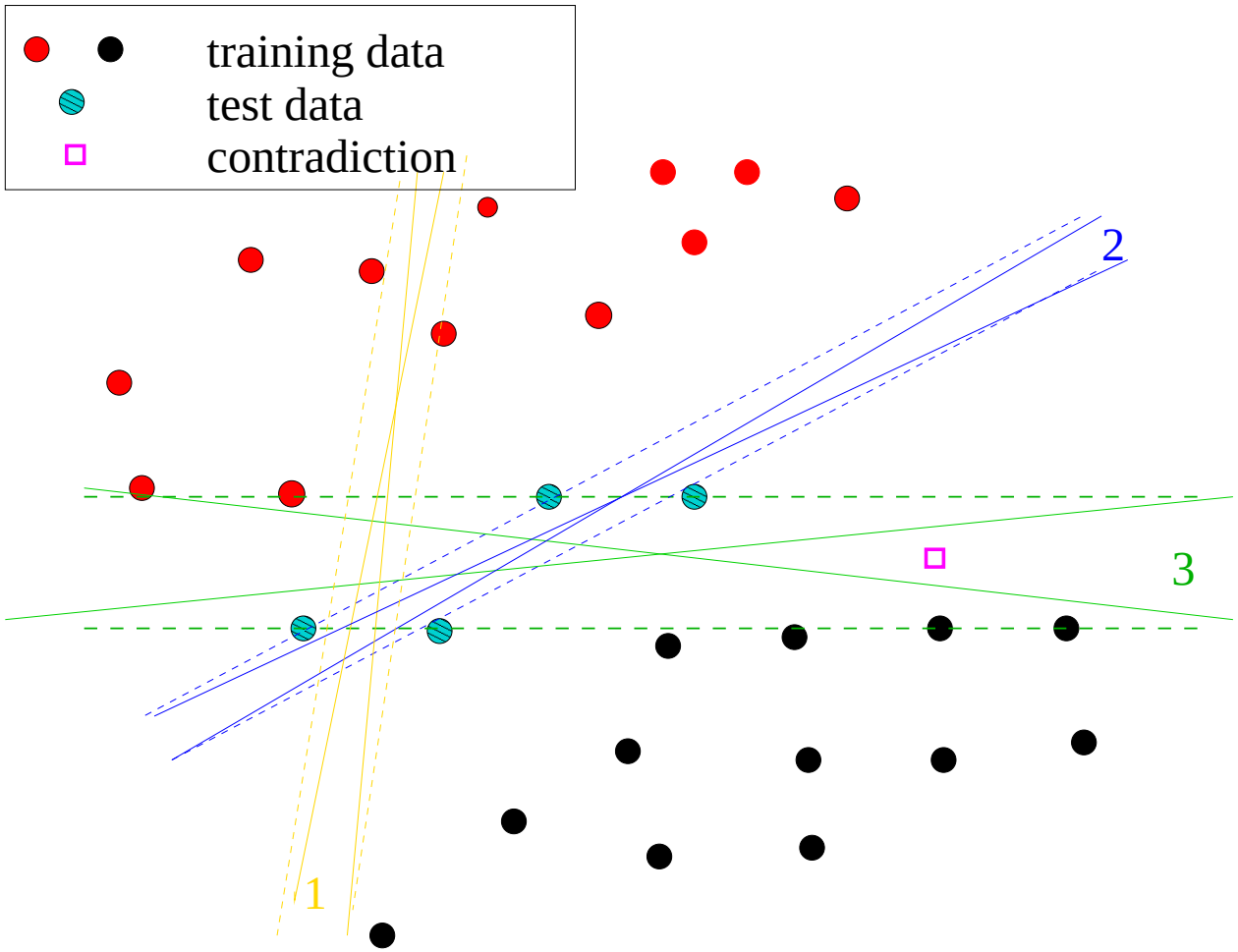
For **inductive** inference (using training data only): **74.5**%.

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

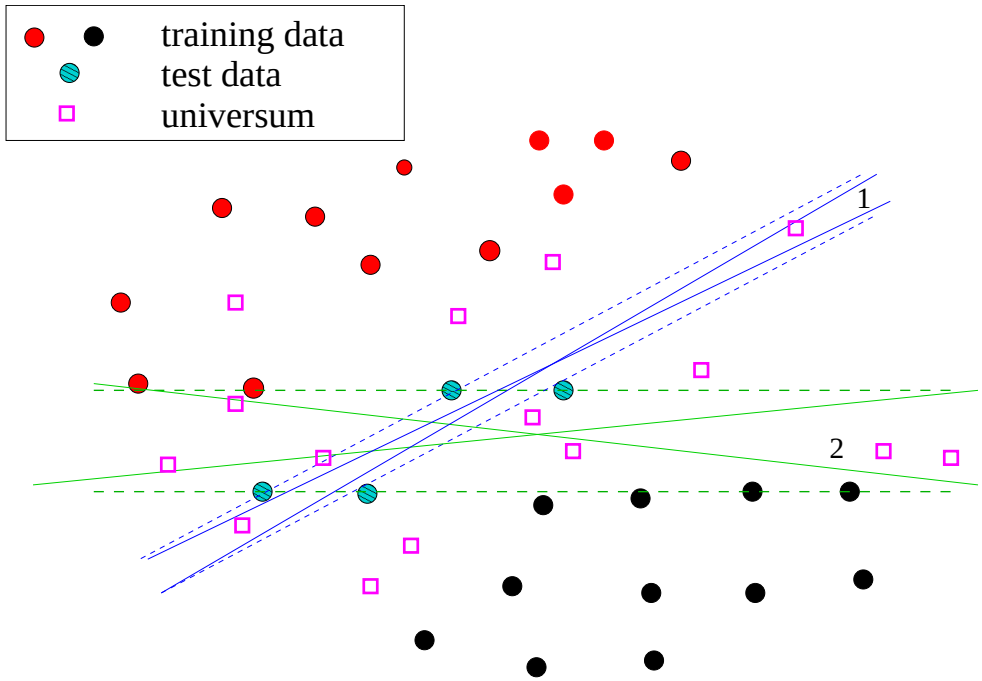
Comparison to other 119 entrants of the competition.



DEFINITION OF EQUIVALENCE CLASSES AND CONTRADICTORY VECTOR



INFERENCE BASED ON THE NUMBER OF CONTRADICTIONS



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

INFERENCE BASED ON MEASURE OF SIMILARITY OF SETS OF DATA

Problem:

Given a set of training data

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

and a set of test data

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

choose a classification vector $(y_1^*, ...y_k^*)$ from an admissible set of vectors.

Method:

1. For any admissible vector $(y_1^*, ..., y_k^*)$ find similarity measure between set of training data $(*)$ and the set of pairs

$$(x_1^*, y_1^*), ..., (x_k^*, y_k^*)$$

2. Choose the vector that provides the best similarity measure.

We have the example of an analytic solution of linear regression problem by this method that provides better accuracy than least squares method.

BEYOND TRANSDUCTION: SELECTIVE INFERENCE

Given ℓ training examples

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

and n candidates vectors

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

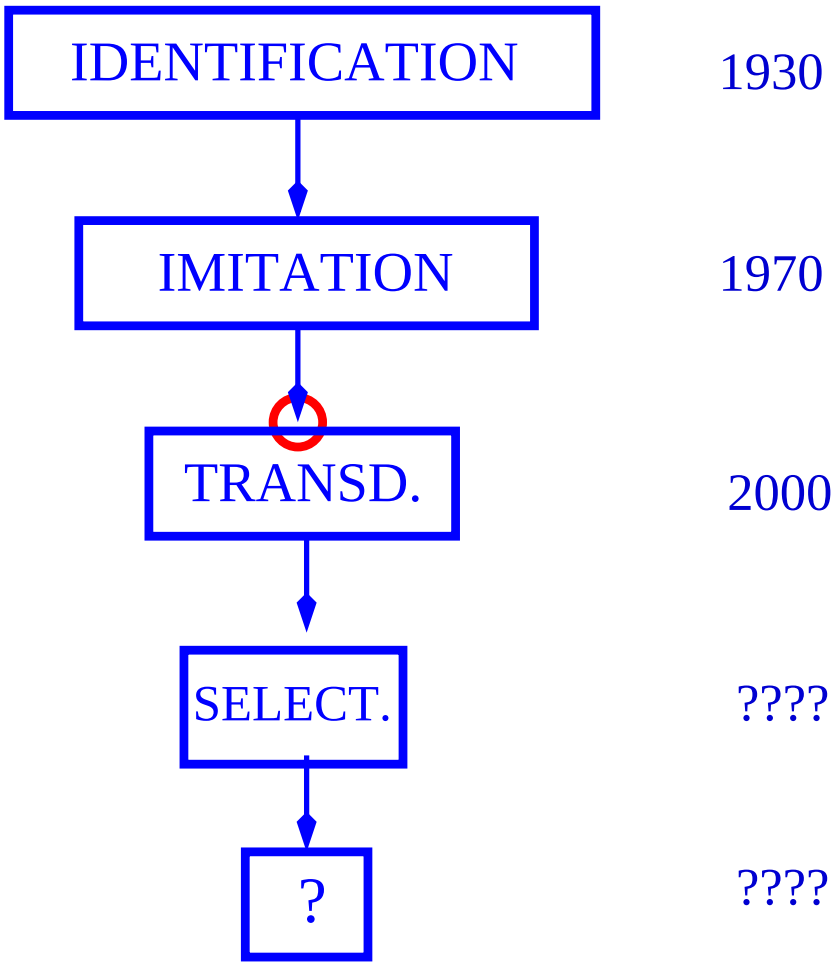
select among n candidate these k vectors with the highest probability of belonging to the first class.

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.

BIG PICTURE: TRADING PHILOSOPHICAL AMBITIONS FOR DIMENSIONALITY



CONCLUSION

It is difficult (maybe even impossible) to discover good models in a high-dimensional World. However using our past experience we can act well in this World. We can predict, make reasonable decisions, and control complex objects. The Empirical Inference Science addresses the question:

How can one effectively use past experience to act well in a high-dimensional complex World?

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 became 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.

PHILOSOPHICAL RETROSPECTIVE

- 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.

TWO METHAPHORS FOR A SIMPLE WORLD

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.

FIRST METAPHOR

Subtle is Lord, but malicious He is not.

A. Einstein

VC 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 catastrophy.

THIRD METAPHOR

If God does exist then many thins are forbidden.

F. Dostoevsky.

VC 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 VC IMPERATIVE FOR HIGH DIMENSIONAL EMPIRICAL INFERENCE

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.)
- Do not estimate a function if you need to estimate it values at given points.
  (Try to perform transduction not induction.)
- Do not estimate predictive values if your goal to act well.
  (Good strategy of action not necessary rely on good prediction.)