LEARNING WITH NONTRIVIAL TEACHER: LEARNING USING PRIVILEGED INFORMATION

Vladimir Vapnik

Columbia University, NEC-labs

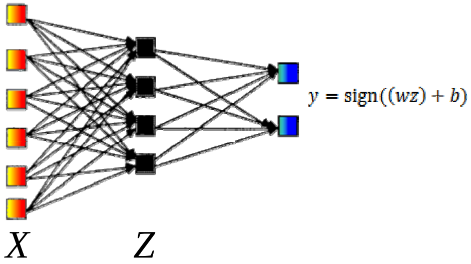
THE ROSENBLATT'S PERCEPTRON AND CLASSICAL MACHINE LEARNING PARADIGM

THE ROSENBLATT'S SCHEME:

1. Transform input vectors of space X into space Z.
2. Using training data

$$(x_1, y_1), ...(x_\ell, y_\ell) \tag{1}$$

construct a separating hyperplane in space Z



GENERAL MATHEMATICAL SCHEME:

1. From a given collection of functions $f(x, \alpha), \alpha \in \Lambda$ choose one that minimizes the number of misclassification on the training data (1)

MAIN RESULTS OF THE VC THEORY

1. There exist two and **only** two factors responsible for generalization:

a) The percent of training errors ν_{train}.

b) The capacity of the set of functions from which
one chooses the desired function (the VC dimension $VCdim$).

2a. The following bounds on probability of test error (P_{test}) are valid

$$P_{test} \leq \nu_{train} + O^* \left(\sqrt{\frac{VCdim}{\ell}} \right)$$

where ℓ is the number of observations.

2b. When $\nu_{train} = 0$ the following bounds are valid

$$P_{test} \leq O^* \left(\frac{VCdim}{\ell} \right)$$

The bounds are achievable.

NEW LEARNING MODEL — LEARNING WITH A NONTRIVIAL TEACHER

Let us include a teacher in the learning process.

During the learning process a teacher supplies training example with additional information which can include comments, comparison, explanation, logical, emotional or metaphorical reasoning, and so on.

This additional (privileged) information is available only for the training examples. It is not available for test examples.

Privileged information exists for almost any learning problem and can play a crucial role in the learning process: it can significantly increase the speed of learning.

THE BASIC MODELS

The classical learning model:. given training pairs

$$(x_1, y_1), ..., (x_\ell, y_\ell), \quad x_i \in X, \quad y_i \in \{-1, 1\}, \quad i = 1, ..., \ell,$$

find among a given set of functions $f(x, \alpha), \alpha \in \Lambda$ the function $y = f(x, \alpha_*)$ that minimizes the probability of incorrect classifications P_{test}

The LUPI learning model: given training triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell), \quad x_i \in X, \quad x_i^* \in X^*, \quad y_i \in \{-1, 1\}, \quad i = 1, ..., \ell,$$

find among a given set of functions $f(x, \alpha), \alpha \in \Lambda$ the function $y = f(x, \alpha_*)$ that minimizes the probability of incorrect classifications P_{test}.

GENERALIZATION OF PERCEPTRON — SVM

Generalization 1: Large margin.

Minimize the functional

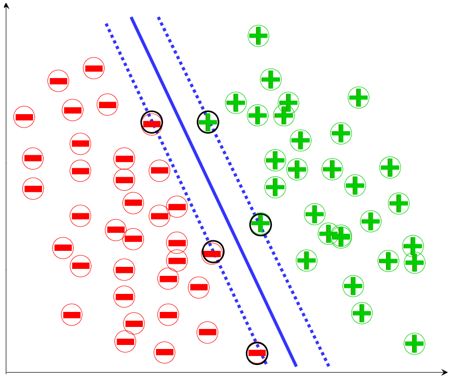
$$R = (w, w)$$

subject to the constraints

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

The solution (w_ℓ, b_ℓ) has the bound

$$P_{test} \leq O^* \left(\frac{VCdim}{\ell} \right).$$



GENERALIZATION OF PERCEPTRON — SVM

Generalization 2: Nonseparable case.

Minimize the functional

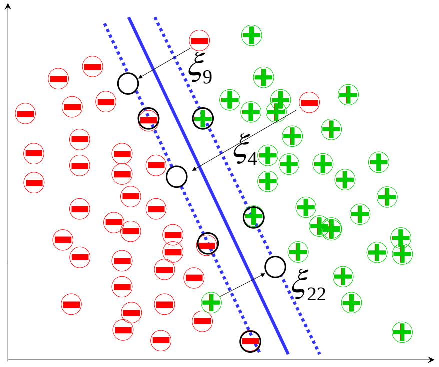
$$R(w, b) = (w, w) + C \sum_{i=1}^{\ell} \xi_i$$

subject to constraints

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

The solution (w_ℓ, b_ℓ) has the bound

$$P_{test} \leq \nu_{train} + O^* \left(\sqrt{\frac{VCdim}{\ell}} \right).$$



WHY WE HAVE SO BIG DIFFERENCE?

- In the separable case using ℓ examples one estimates n parameters of w.
- In the non-separable case one estimates $n+\ell$ parameters (n parameters of vector w and ℓ parameters of slacks).

Suppose that we know set of functions $\xi(x,\delta) \geq 0,\ \delta \in \mathcal{D}$ such that

$$\xi = \xi(x) = \xi(x,\delta_0)$$

and has finite VCdim* (let δ be an m-dimensional vector).

In this situation to find optimal hyperplane in the non-separating case one needs to estimate $n+m$ parameters using ℓ observations.

Can the rate of convergence in this case be faster?

THE KEY OBSERVATION: ORACLE SVM

Suppose we are given triplets

$$(x_1, \xi_1^0, y_1), ..., (x_\ell, \xi_\ell^0, y_\ell),$$

where $\xi_i^0 = \xi^0(x_i), \quad i = 1, ..., \ell$ are the slack values with respect to the best hyperplane. Then to find the approximation (w_{best}, b_{best}) we minimize the functional

$$R(w, b) = (w, w)$$

subject to constraints

$$y_i[(w, x_i) + b] \geq r_i, \quad r_i = 1 - \xi^0(x_i), \quad i = 1, ..., \ell.$$

Proposition 1. For Oracle SVM the following bound holds

$$P_{test} \leq \nu_{train} + O^* \left(\frac{VCdim}{\ell} \right).$$

ILLUSTRATION — I

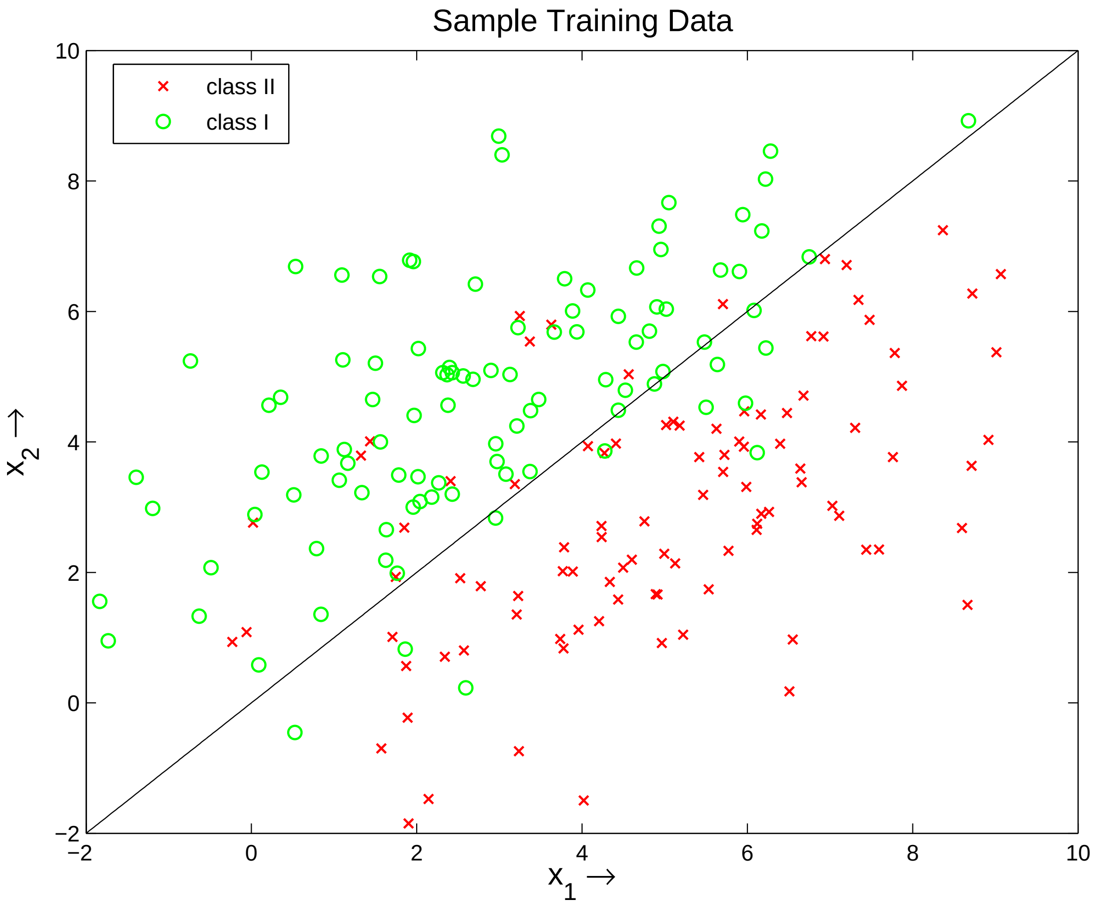
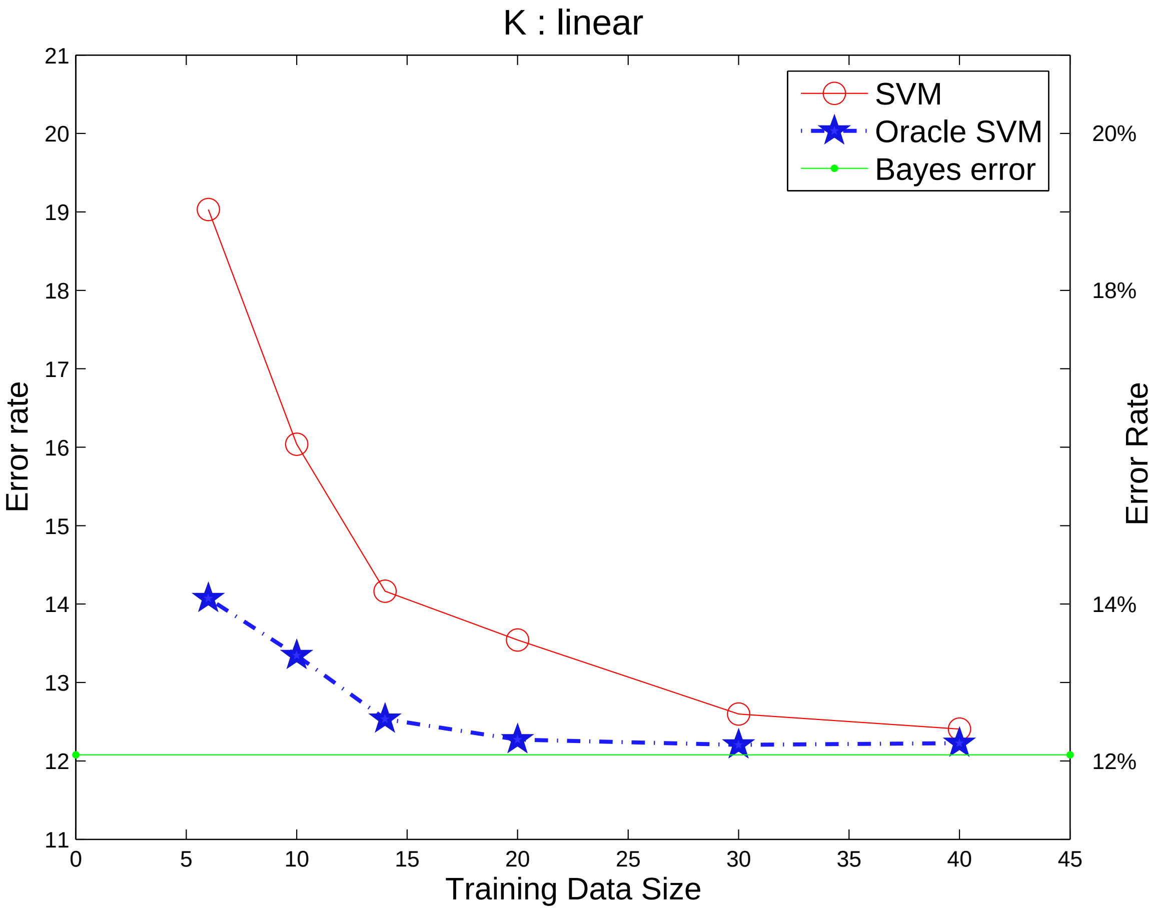


ILLUSTRATION —II



WHAT CAN A REAL TEACHER DO — I

One can not expect that a teacher knows values of slacks. However he can:

- Supply students with a *correcting space* X^* and a set of functions $\xi(x^*, \delta)$, $\delta \in D$, in this space (with VC dimension h^*) which contains a function

  $$\xi_i = \xi(x_i^*, \delta_{best})$$

  that approximates the oracle slack function $\xi^0 = \xi^0(x^*)$ well.
- During training process supply students with triplets

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

  in order to estimate simultaneously both the correcting (slack) function

  $$\xi = \xi(x^*, \delta_\ell)$$

  and the decision hyperplane (pair (w_ℓ, b_ℓ)).

WHAT CAN A REAL TEACHER DO — II

The problem of learning with a teacher is to minimize the functional

$$R(w,b,\delta) = (w,w) + C\sum_{i=1}^{\ell}\xi(x_i^*,\delta)$$

subject to constraints $\xi(x^*,\delta) \geq 0$ and constraints

$$y_i((w,x)+b) \geq 1 - \xi(x_i^*,\delta), \quad i = 1, ..., \ell.$$

Proposition 2. *With probability* $1-\eta$ *the following bound holds true*

$$P(y[(w_\ell,x)+b_\ell] < 0) \leq P(1-\xi(x^*,\delta_\ell) < 0) + A\frac{(n+h^*)(\ln\frac{2\ell}{(n+h^*)}+1)-\ln\eta}{\ell}.$$

The problem is how good is the teacher: *how fast the probability* $P(1-\xi(x^*,\delta_\ell) < 0)$ *converges to the probability* $P(1-\xi(x^*,\delta_0)) < 0)$.

THE BOTTOM LINE

The goal of a teacher is by introducing both the space X^* and the set of slack-functions in this space $\xi(x^*, \delta), \delta \in \Delta$ to try speed up the rate of convergence of the learning process from $O(\frac{1}{\sqrt{\ell}})$ to $O(\frac{1}{\ell})$.

The difference between standard and fast methods is in the number of examples needed for training:
ℓ for the standard methods and $\sqrt{\ell}$ for the fast methods (i.e. 100,000 and 320; or 1000 and 32).

IDEA OF SVM ALGORITHM

- **Transform the training pairs**

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

into the pairs

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

by mapping vectors $x \in X$ into $z \in Z$.

- **Find in Z the hyperplane that minimizes the functional**

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

subject to constraints

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

- **Use inner product in Z space in the form**

$$(z_i, z_j) = K(x_i, x_j).$$

DUAL SPACE SOLUTION FOR SVM

The decision function has a form

$$f(x, \alpha) = \mathbf{sgn}\left[\sum_{i=1}^{\ell} \alpha_i y_i K(x_i, x) + b\right] \tag{2}$$

where $\alpha_i \geq 0,\ i = 1, ..., \ell$ are values which maximize the functional

$$R(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{3}$$

subject to constraints

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

Here kernel $K(\cdot, \cdot)$ is used for two different purposes:
1. In (2) to define a set of expansion-functions $K(x_i, x)$.
2. In (3) to define similarity between vectors x_i and x_j.

IDEA OF SVM+ ALGORITHM

- Transform the training triplets $(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell)$

into the triplets $(z_1, z_1^*, y_1), ..., (z_\ell, z_\ell^*, y_\ell)$

by mapping vectors $x \in X$ into vectors $z \in Z$ and $x^* \in X^*$ into $z^* \in Z^*$.

- Define the slack-function in the form

$$\xi_i = (w^*, z_i^*) + b^*$$

and find in space Z the hyperplane that minimizes the functional

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

subject to constraints

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

- Use inner products in Z and Z^* spaces in the kernel form

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

DUAL SPACE SOLUTION FOR SVM+

The decision function has a form

$$f(x,\alpha) = \operatorname{sgn}\left[\sum_{i=1}^{\ell} \alpha_i y_i K(x_i, x) + b\right]$$

where $\alpha_i, \;\; i = 1, ..., \ell$ are values that maximize the functional

$$R(\alpha, \beta) =$$

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

subject to the constraints

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0, \qquad \sum_{i=1}^{\ell} (\alpha_i - \beta_i) = 0.$$

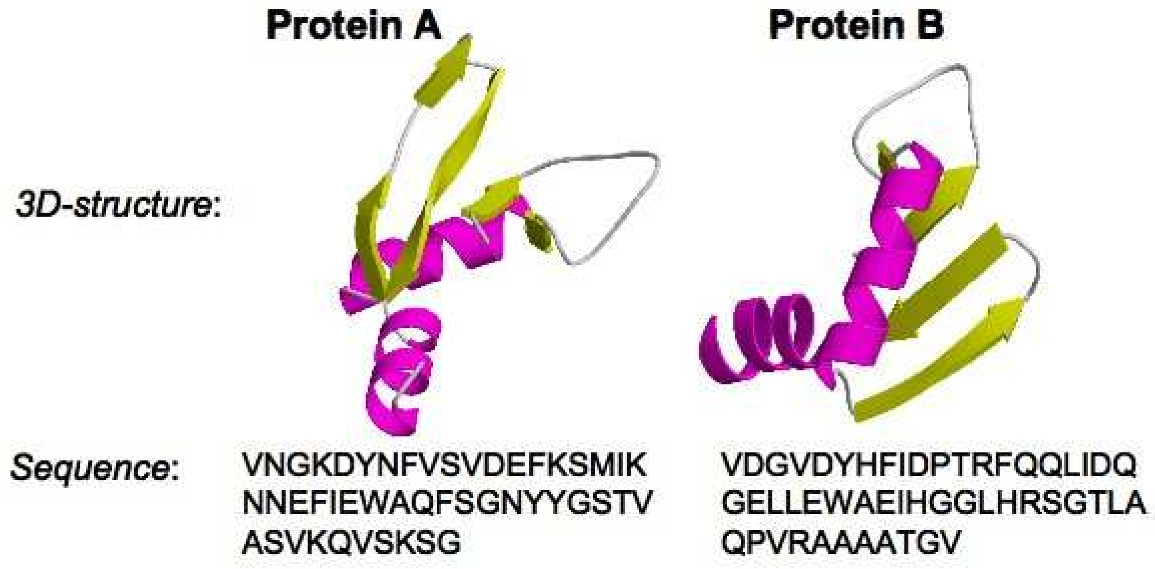
and the constraints

$$\alpha_i \geq 0, \qquad 0 \leq \beta_i \leq C$$

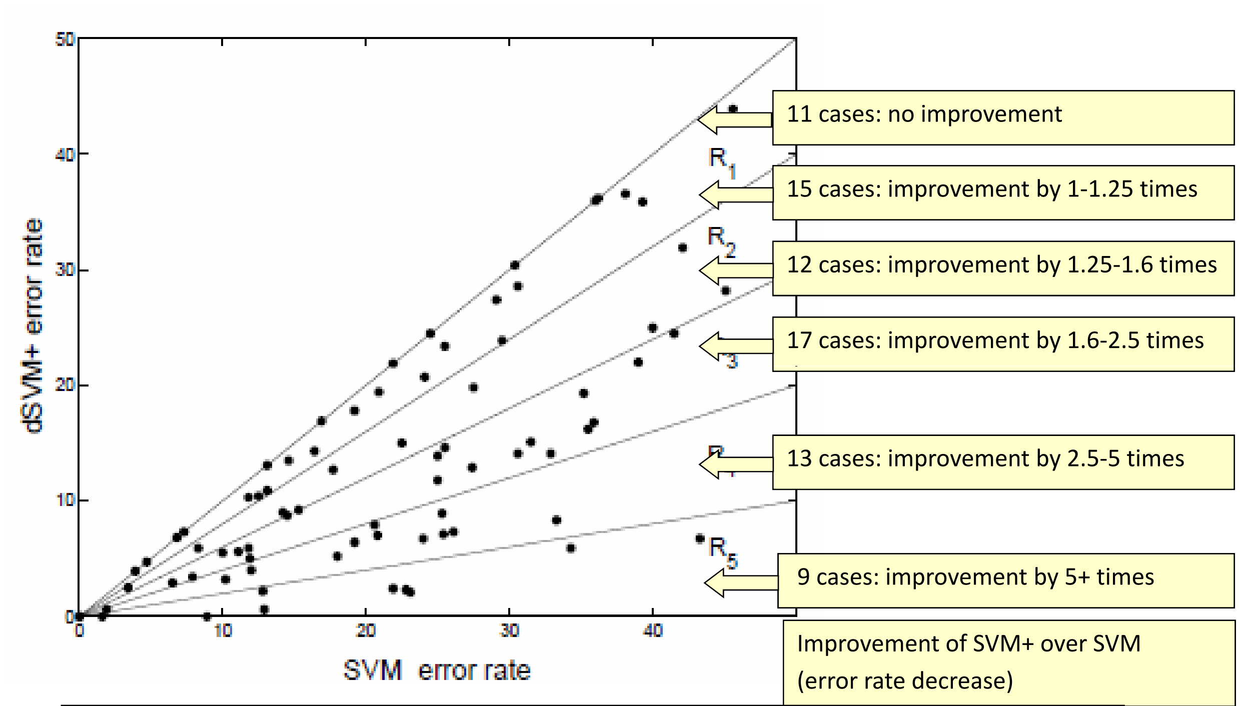
ADVANCED TECHNICAL MODEL AS PRIVILEGED INFORMATION

Classification of proteins into families

The problem is : Given amino-acid sequences of proteins construct a rule to classify families of proteins. The decision space X is the space of amino-acid sequences. The privileged information space X^* is the space of 3D structure of the proteins.



CLASSIFICATION OF PROTEIN FAMILIES



11 cases: no improvement

15 cases: improvement by 1-1.25 times

12 cases: improvement by 1.25-1.6 times

17 cases: improvement by 1.6-2.5 times

13 cases: improvement by 2.5-5 times

9 cases: improvement by 5+ times

Improvement of SVM+ over SVM
(error rate decrease)

- Significant improvement in classification accuracy achieved without additional data
- Two factors contribute to classification accuracy improvement:
  - Privileged information during training (3D structures)
  - New mathematical algorithm (SVM+)

CLASSIFICATION OF PROTEINS: DETAILS

Protein superfamily pair	SVM	SVM+	SVM (3D)	
a.26.1-vs-c.68.1	7.3	7.3	0	
a.26.1-vs-g.17.1	16.4	14.3	0	
a.118.1-vs-b.82.1	19.2	6.4	0	★
a.118.1-vs-d.2.1	41.5	24.5	3.8	
a.118.1-vs-d.14.1	13.1	13.1	2.2	
a.118.1-vs-e.8.1	22.8	2.3	2.3	★
b.1.18-vs-b.55.1	14.6	13.5	0	
b.18.1-vs-b.55.1	31.5	15.1	0	★
b.18.1-vs-c.55.1	36.2	36.2	0	☆
b.18.1-vs-c.55.3	38.1	36.6	0	☆
b.18.1-vs-d.92.1	25	11.8	0	★
b.29.1-vs-b.30.5	16.9	16.9	3.6	
b.29.1-vs-b.55.1	10	5.5	0	
b.29.1-vs-b.80.1	8.3	5.9	0	
b.29.1-vs-b.121.4	35.9	16.8	5.3	★

Protein superfamily pair	SVM	SVM+	SVM (3D)	
b.29.1-vs-c.47.1	10.2	3.2	0	★
b.30.5-vs-b.80.1	43.3	6.7	0	★
b.30.5-vs-b.55.1	25.5	14.6	0	
b.55.1-vs-b.82.1	11.8	10.3	0	
b.55.1-vs-d.14.1	20.9	19.4	0	
b.55.1-vs-d.15.1	17.7	12.7	0	
b.80.1-vs-b.82.1	4.7	4.7	0	
b.82.1-vs-b.121.4	7.9	3.4	0	★
b.121.4-vs-d.14.1	29.5	23.9	0	
b.121.4-vs-d.92.1	15.3	9.2	0	
c.36.1-vs-c.68.1	8.9	0	0	★
c.36.1-vs-e.8.1	12.8	2.2	0	★
c.47.1-vs-c.69.1	1.9	0.6	0	★
c.52.1-vs-b.80.1	11.8	5.9	0	
c.55.1-vs-c.55.3	45.1	28.2	22.5	

☆ 3D structure is essential for classification; SVM+ does not improve classification of SVM

★ SVM+ provides significant improvement over SVM (several times)

FUTURE EVENTS AS PRIVILEGED INFORMATION

Time series prediction

Given pairs

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

find the rule

$$y_t = f(x_{t+\Delta}),$$

where

$$x_t = (x(t), ..., x(t-m)).$$

For regression model of time series:

$$y_t = x(t+\Delta).$$

For classification model of time series:

$$y_t = \begin{cases} 1, & \text{if} \quad x(t+\Delta) > x(t), \\ -1, & \text{if} \quad x(t+\Delta) \leq x(t). \end{cases}$$

MACKEY-GLASS TIME SERIES PREDICTION

Let data be generated by the Mackey-Glass equation:

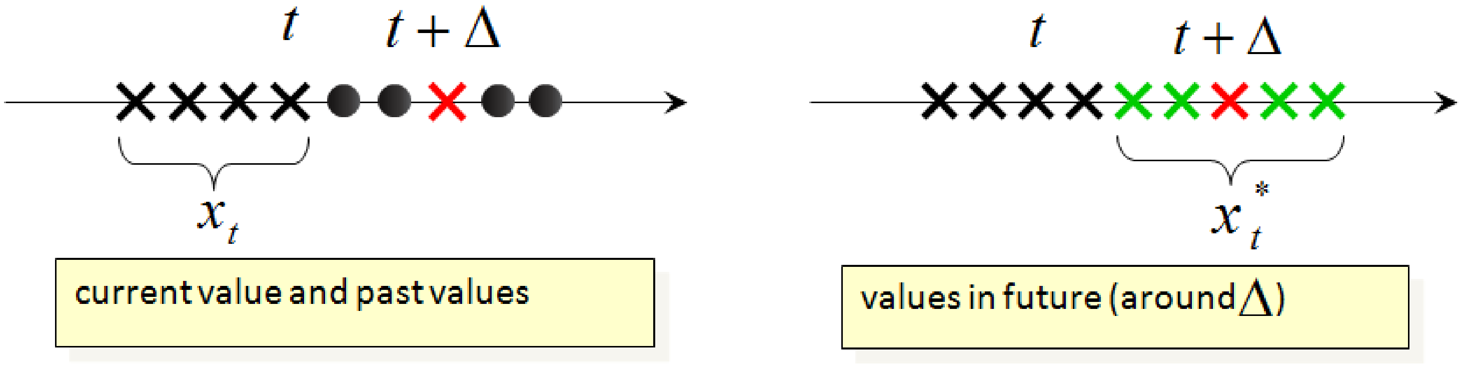
$$\frac{dx(t)}{dt} = -ax(t) + \frac{bx(t-\tau)}{1+x^{10}(t-\tau)},$$

where a, b, and τ (delay) are parameters.

The training triplets $(x_1, , x_1^*, y_1), \ldots., (x_\ell, x_\ell^*, y_\ell)$ are defined as follows:

$$x_t = (x(t), x(t-1), x(t-2), x(t-3))$$

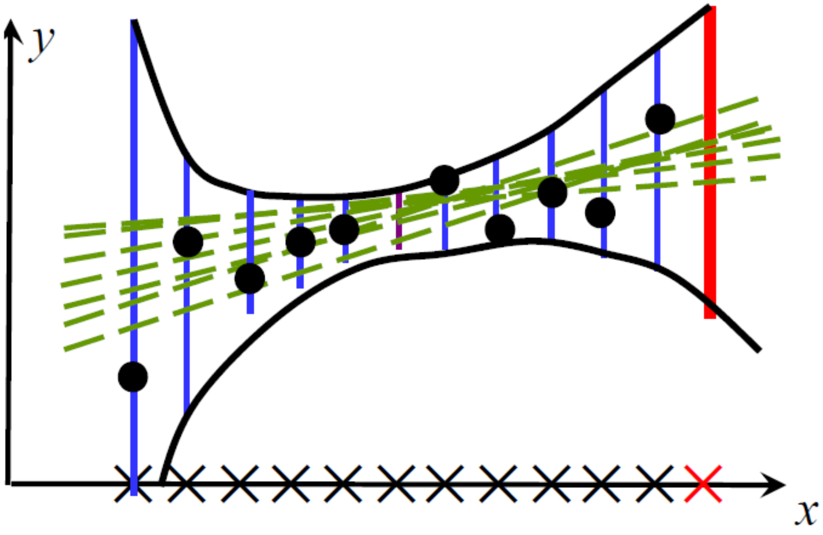
$$x_t^* = (x(t+\Delta-1), x(t+\Delta-2), x(t+\Delta+1), x(t+\Delta+2))$$

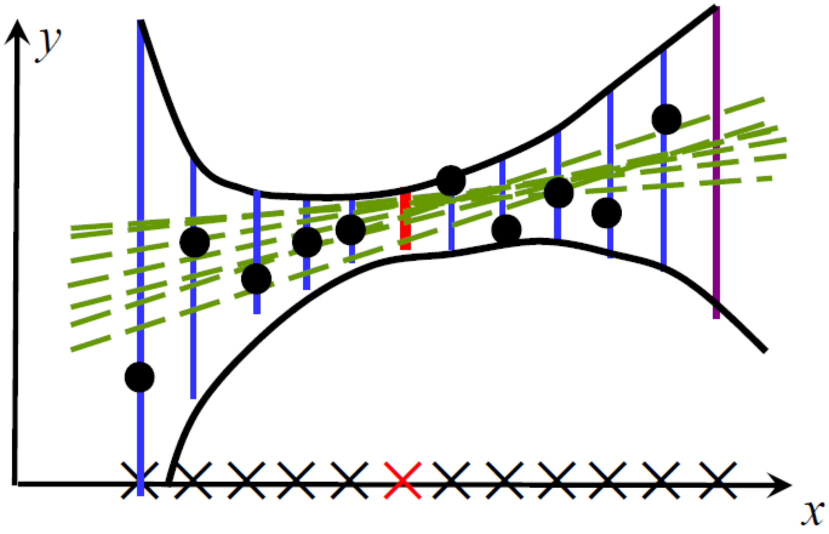


current value and past values

values in future (around Δ)

INTERPOLATION AND EXTRAPOLATION

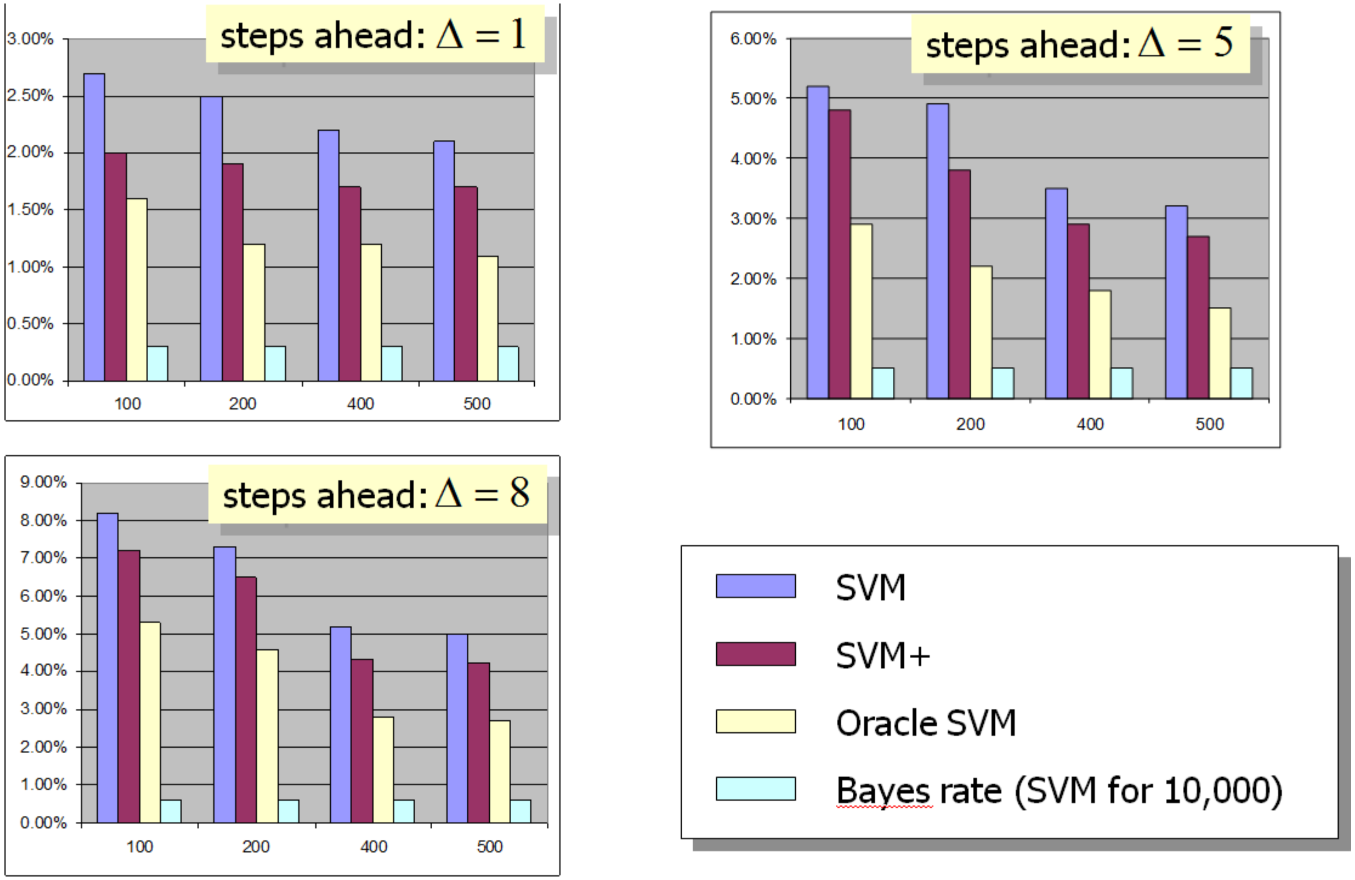




- Extrapolation of trends has to face large conditional variance

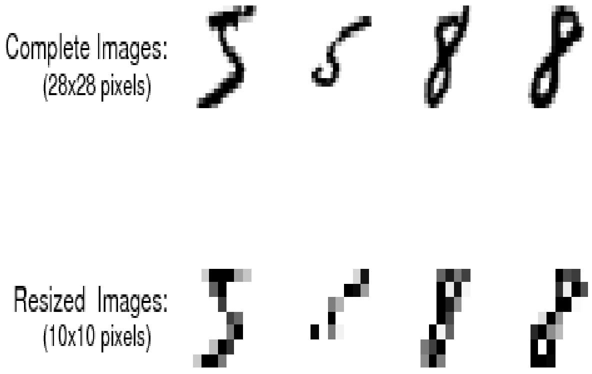
- Interpolation of trends faces small conditional variance

ILLUSTRATION



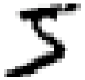
HOLISTIC DESCRIPTION AS PRIVILEGED INFORMATION

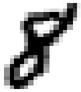
Classification of digit 5 and digit 8 from the NIST database.



Given triplets (x_i, x_i^*, y_i), $i = 1, ..., \ell$ find the classification rule $y = f(x)$, where x_i^* is the holistic description of the digit x_i.

YING YANG STYLE DESCRIPTIONS

 Straightforward, very active, hard, very masculine with rather clear intention. A sportsman or a warrior. Aggressive and ruthless, eager to dominate everybody, clever and accurate, more emotional than rational, very resolute. No compromise accepted. Strong individuality, egoistic. Honest. Hot, able to give much pain. Hard. Belongs to surface. Individual, no desire to be sociable. First moving second thinking. Will never give a second thought to whatever. Upward-seeking. 40 years old.

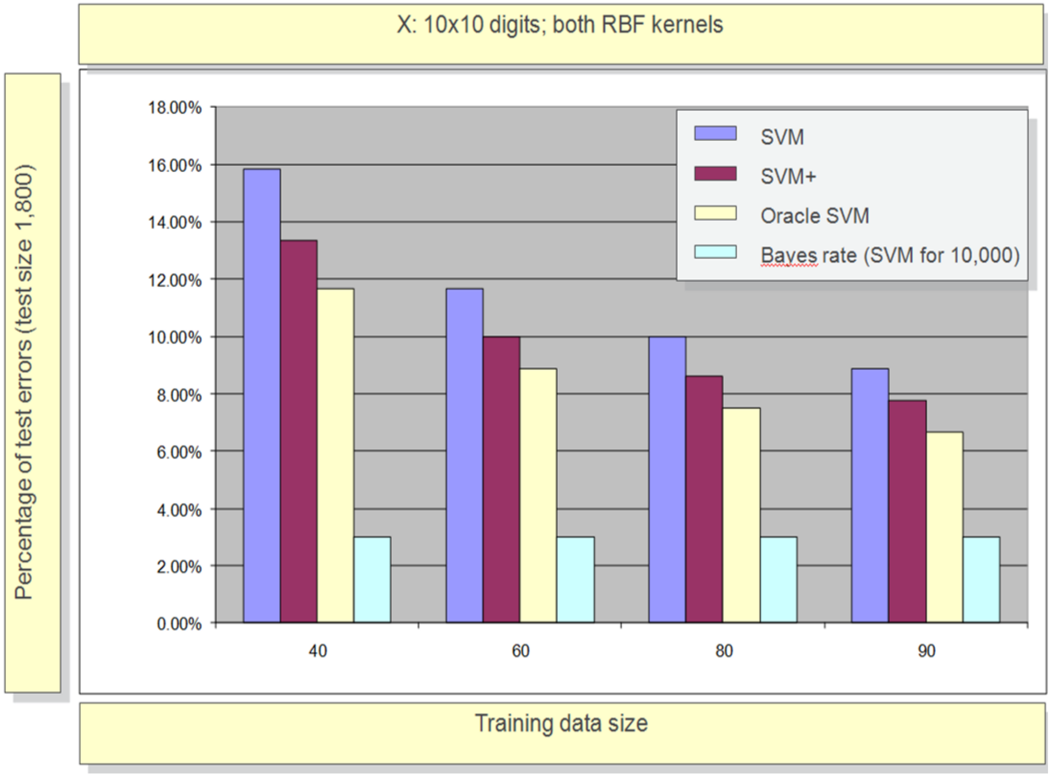
 A young man is energetic and seriously absorbed in his career. He is not absolutely precise and accurate. He seems a bit aggressive mostly due to lack of sense of humor. He is too busy with himself to be open to the world. He has simple mind and evident plans connected with everyday needs. He feels good in familiar surroundings. Solid soil and earth are his native space. He is upward seeking but does not understand air.

CODES FOR HOLISTIC DESCRIPTION

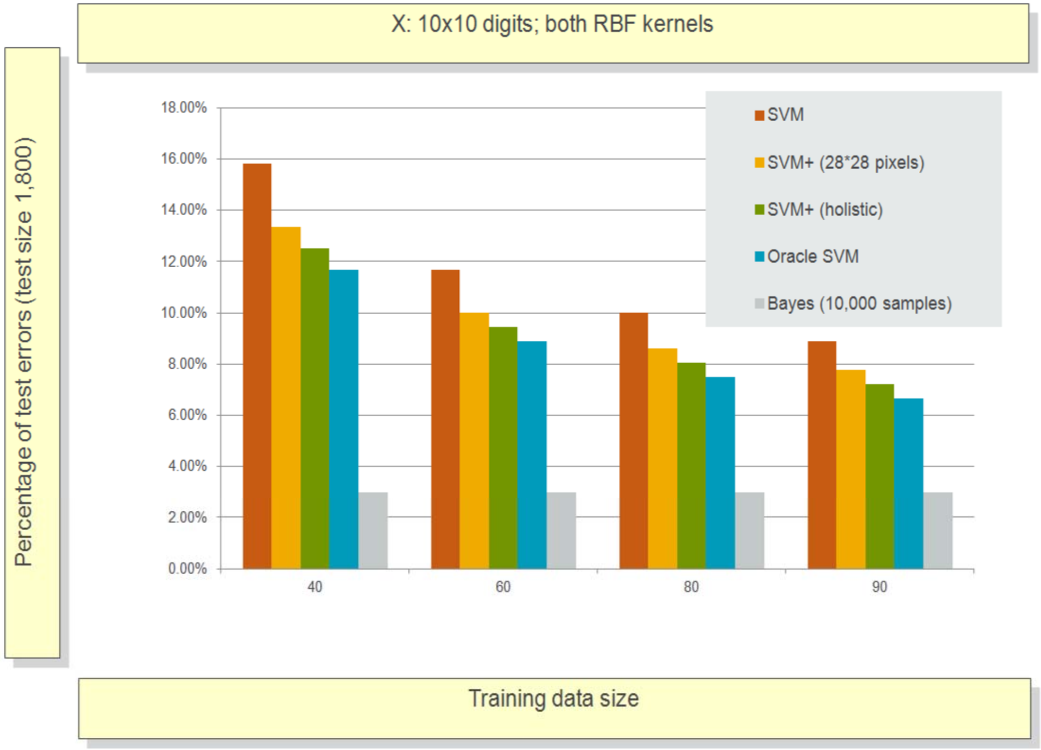
1.Active (0 — 5), **2.Passive** (0 — 5), **3.Feminine** (0 — 5),
4.Masculine (0 — 5), **5.Hard** (0 — 5), **6.Soft** (0 — 5),
7.Occupancy (0 — 3), **8.Strength** (0 — 3), **9.Hot** (0 — 3),
10.Cold (0 — 3), **11.Aggressive** (0 — 3), **12.Controlling** (0 — 3),
13.Mysterious (0 — 3), **14.Clear** (0 — 3), **15.Emotional** (0 — 3),
16.Rational (0 — 3), **17.Collective** (0 — 3), **18.Individual** (0 — 3),
19.Serious (0 — 3), **20.Light-minded** (0 —3), **21.Hidden** (0 — 3),
22.Evident (0 — 3), **23.Light** (0 — 3), **24.Dark** (0 — 3),
25.Upward-seeking (0 — 3), **26.Downward-seeking** (0 — 3),
27.Water flowing (0 — 3), **28.Solid earth** (0 — 3),
29.Interior (0 — 2), **30.Surface** (0 — 2), **31.Air** (0—3).

`http://ml.nec-labs.com/download/data/svm+/mnist.priviledged`

RESULTS


X: 10x10 digits; both RBF kernels
Percentage of test errors (test size 1,800)
18.00%
16.00%
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
40
60
80
90
SVM
SVM+
Oracle SVM
Bayes rate (SVM for 10,000)
Training data size

HOLISTIC SPACE VS. ADVANCED TECHNICAL SPACE



IMAGES CLASSIFICATION I

Dog/cat classification (Pascal) 2006 (Geng, Qi, Tian, Yu) Images were resized to be gray 80×100 pixels. (50+50 training set)

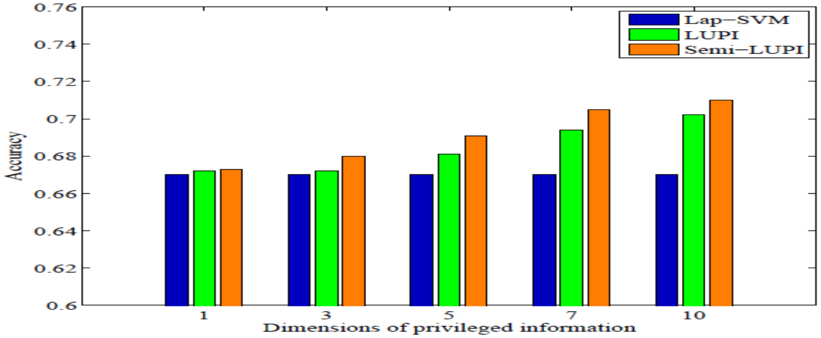
Privileged information

The ear is small in proportion to its face; the mouth is narrow and non-prominent; the nose is small and its color is light; short and rounded head; hardly see its lip on the face; the color of the whole body is very bright and rich; see the whole body; several cats in the picture; the image is clear. A holistic description for some dog is as follows (see Fig. 6 (b)): The ear is large in proportion to its face; the mouth is wide and prominent; the nose is large and black; the face is very long; the lip is also long and just like a zipper on the face; the color of the whole body is very dark and lacks diversity; only see the part of the body; only a dog in the picture; the image is clear.

IMAGES CLASSIFICATION II

Feature space. Holistic descriptions is translated into 10-dimensional feature vectors: the length of the ear in proportion to its face(0-5); the width of the mouth (0-5); the prominent extent of the mouth (0-6); the size of the noise (0-6); the color of the noise (0-4); the length of the head (0-5); the appearance of the head (0-4); the length of the lip (0-6); if see the whole body (1-2); the number of the animal (0-6); the clearness of the image (0-5) .

The results



DUAL SPACE SOLUTION FOR SVM+

The decision function has a form

$$f(x, \alpha) = \operatorname{sgn}\left[\sum_{i=1}^{\ell} \alpha_i y_i K(x_i, x) + b\right]$$

where $\alpha_i, \ i = 1, ..., \ell$ are values that maximize the functional

$$R(\alpha, \beta) =$$

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

subject to the constraints

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0, \qquad \sum_{i=1}^{\ell} (\alpha_i - \beta_i) = 0.$$

and the constraints

$$\alpha_i \geq 0, \qquad 0 \leq \beta_i \leq C$$

TWO EXAMPLES OF POSSIBLE PRIVILEGED INFORMATION

- **Semi-scientific models (say, use Elliott waves, informal human-machine inference) as privileged information to improve formal models.**
- **Alternative theory to improve the theory of interest (say, use Eastern medicine as privileged information to improve rules of Western medicine).**

HOLISTIC (YING-YANG) DESCRIPTIONS OF PULSE

- **Shallow pulse (Yang).** Shallow pulse flows in the surface. You press it and it seems full, you press stronger - it becomes weak. It is like slight breeze whirling up bird's tuft, like wind swaying leaves, like water which sways a chip of wood when the wind is blowing.
- **Deep pulse (Ying).** The deep pulse is similar to a stone wrapped in cotton wool: it is soft from the outside and it is hard inside. It lies in the bottom like a stone thrown in the water.
- **Free pulse (Ying).** Such pulse is irregular. It reminds of a pearl rolling in a plate. It flows like a drop after a drop, sliding like a pearl after a pearl.
- **String pulse (Ying in Yang).** This pulse makes an impression of a tight violin string. Its beating is direct and long like a string.
- **Skin pulse (Ying).** Its beating is elastic and resilient like a drum. The pulse is shallow and reminds touching drum skin.
- **Inconspicuous pulse.** The beating is exceptionally soft and gentle as well as shallow and thin. It reminds of a silk cloth flowing in the water.

RELATION TO DIFFERENT BRANCHES OF SCIENCE

- **Statistics:** Non-symmetric models in predictive statistics (advanced and future events as privileged information in regression and time series analysis).
- **Cognitive science:** Role of right and left parts of the brain (existence and unity of two different information spaces: *analytic* and *holistic*).
- **Psychology:** Emotional logics in inference problems.
- **Philosophy of Science:** Difference in analysis Simple World and Complex World (unity of analytic and holistic models of complex worlds).

LIMITS OF THE CLASSICAL MODELS OF SCIENCE

- WHEN THE 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.

THE BOTTOM LINE

- Machine Learning science is not only about computers. It is also science about humans: unity of their logics, emotions, and cultures.
- Machine Learning is the discipline that can produce and analyze facts that lead to understanding of model of science for Complex World which is based not enterally on logic (let us call it the Soft Science).

LITERATURE

1. Vladimir Vapnik. Estimation of Dependencies Based on Empirical Data, 2-nd Ed.: Empirical Inference Science, Springer, 2006

2. Vapnik V., Vashist A., Pavlovitch N.: Learning using hidden information: Master class learning. In *Proceedings of NATO workshop on mining massive data sets for security* (pp. 3-14) IOS Press, 2008.

3. Vapnik V., Vashist A.,Pavlovitch N.: Learning using hidden information: (learning with teacher). In *Proceedings of IJCNN* (pp. 3188 –3195), 2009

4. Vladimir Vapnik, Akshay Vashist: A new learning paradigm: Learning using privileged information. Neural Networks 22(5-6): (pp. 544-557), 2009

5. D.Pechyony, R.Izmailov, A.Vashist, V.Vapnik, SMO-style algorithms for learning using privileged information, in Proceedings of the 2010 International Conference on Data Mining (DMIN), 2010.

6. Dmitri Pechyony, Vladimir Vapnik, On the theory of learning with privileged information, NIPS –2010.

Digit database (with Poetic and Ying-Yang descriptions by N. Pavlovitch): http://ml.nec-labs.com/download/data/svm+/mnist.priviledged/