LEARNING WITH TEACHER: LEARNING USING HIDDEN INFORMATION

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

CONTENTS

- What can a teacher do.
- Technical realization of a teacher's effort.
- Three examples of hidden information:
  a) advanced technical models,
  b) future events,
  c) holistic description.
- General discussion.

DEVELOPMENT OF LEARNING ALGORITHMS

- **1966 – 1971.** Development of the theory for Empirical Risk Minimization (ERM) methods. *Any algorithm that minimizes empirical risk in a set of functions with finite capacity (say, finite VC dimension) is consistent.*
- **1968 – 1974.** Development of the theory for Structural Risk Minimization (SRM) principle. *Any algorithm that satisfies the SRM principle converges to the Bayesian solution.*
- **1964, 1992, 1995.** The SVM algorithm that realizes the SRM principle was developed. *SVM with a universal kernel converges to Bayesian solution.*
- In order to improve performance transductive (**1974**) and semi-supervised (**the 1990s**) methods that take into account additional information (not only training data) were introduced.

However, all of these learning models ignore the main driving force in the learning process: THE TEACHER.

WHAT IS THE ROLE OF TEACHER IN LEARNING

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

This information is available only for the training examples. It will not be available (hidden) for the test examples.

Hidden information can play an important role in the learning process.

EXAMPLES OF HIDDEN INFORMATION

Example 1. Suppose our goal is to find a rule that can predict outcome y of a treatment in a year given the current symptoms x of a patient. However at the training stage a teacher can also give an additional information x^* about the development of symptoms in three months, in six months, and in nine months. Can this additional information about the development of symptoms improve a rule that predicts outcome in a year?

Example 2. Suppose that our goal is to find a rule to classify biopsy images into two categories: cancer and non-cancer. Here the problem is given images described in the pixel space find a classification rule in the pixel space. However, along with the picture doctor has a report, written by a pathologist, which describes the pictures using a high level holistic language. The problem is to use pictures along with the pathologist's reports which will not be available at the test stage (in fact, the goal is to make accurate diagnosis without consulting with a pathologist) to find a good classification rule in the pixel space.

EXAMPLES OF HIDDEN INFORMATION

Example 3. Suppose that our goal is to predict the exchange rate of a currency at the moment t in the money exchange problem. In this problem we have observations about the rate before the moment t and the goal is to predict if the rate will go up or down at the moment t. However in the historical data along with observations about the rates before moment t we also have observations about rates after the moment t. This information is hidden for testing (but available for training). Can this hidden information (about future of the past) help one to construct a better predictive rule?

The situation with existence of hidden information is very common. In fact, for almost all machine learning problems there exists some sort of hidden information.

THE BASIC MODELS

The classical learning model:. given a set of iid pairs (training data)

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

generated according to a fixed but unknown probability measure $P(x, y)$, 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 (incorrect values of y).

The LUHI learning model: given a set of iid training triplets

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

generated according to a fixed but unknown probability measure $P(x, x^*, y)$ find among a given set of functions $f(x, \alpha), \alpha \in \Lambda$ the function $y = f(x, \alpha_*)$ that guarantees the smallest probability of incorrect classifications.

THE BASICS OF LEARNING ALGORITHMS

Let $f(x, w, b), w \in W$ be a set of separating hyperplanes

$$f(x, w, b) = \text{sgn}[(\text{w,x})+\text{b}], \quad w \in R^n, \ b \in R^1$$

and

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

be a training set.

Separable case. Minimize the functional

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

subject to constraints

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

The solution (w_ℓ, b_ℓ) defines the hyperplane whose error rate with probability $1 - \eta$ has a bound

$$P_{err}(w_\ell, b_\ell) \leq A \frac{n \ln \frac{\ell}{n} - \ln \eta}{\ell}.$$

Non-separable case. Minimize the functional

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

subject to constraints

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

The solution (w_ℓ, b_ℓ) defines the hyperplane whose error rate with probability $1 - \eta$ has the bound

$$P_{err}(w_\ell, b_\ell) \leq P_{err}(w_{best}, b_{best}) + A\sqrt{\frac{n \ln \frac{\ell}{n} - \ln \eta}{\ell}}.$$

Note that in the separable case using ℓ examples one has to estimate n parameters (of vector w) while in the non-separable case (generally speaking) $n + \ell$ parameters.

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

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

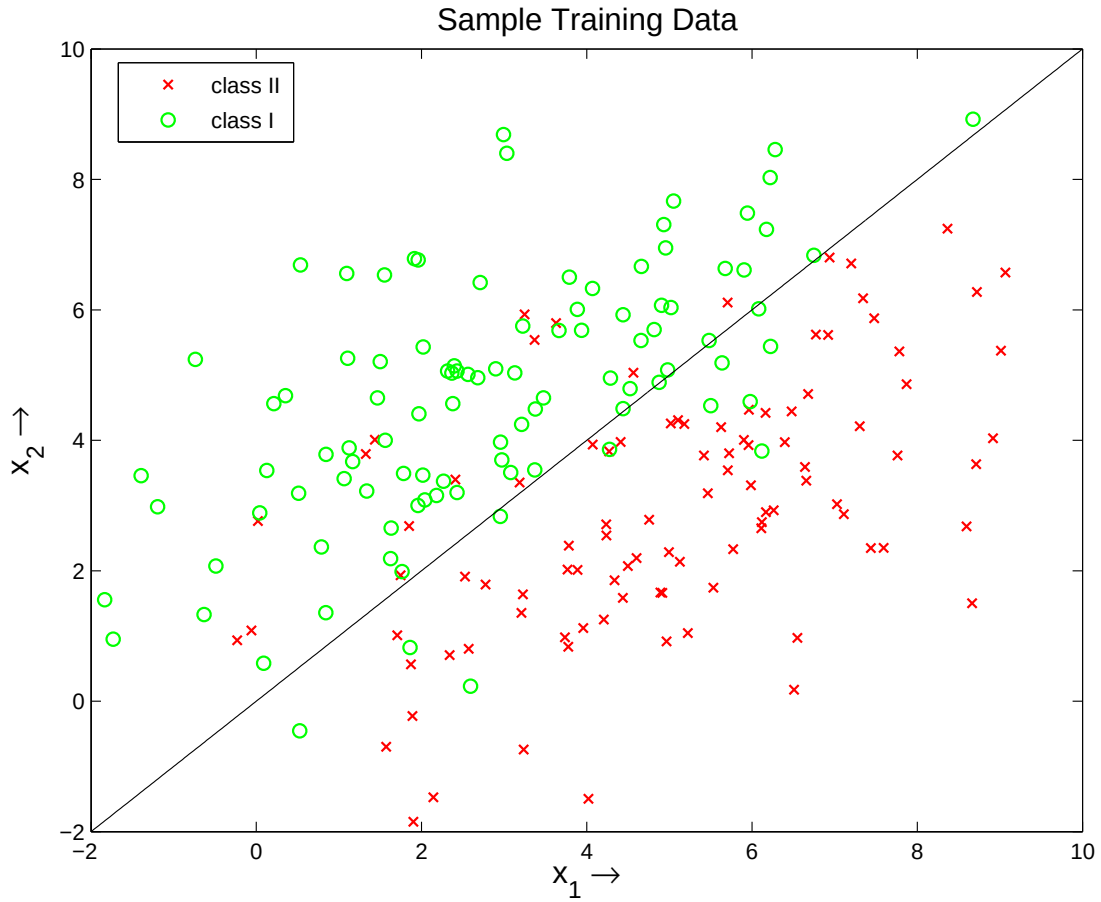
subject to constraints

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

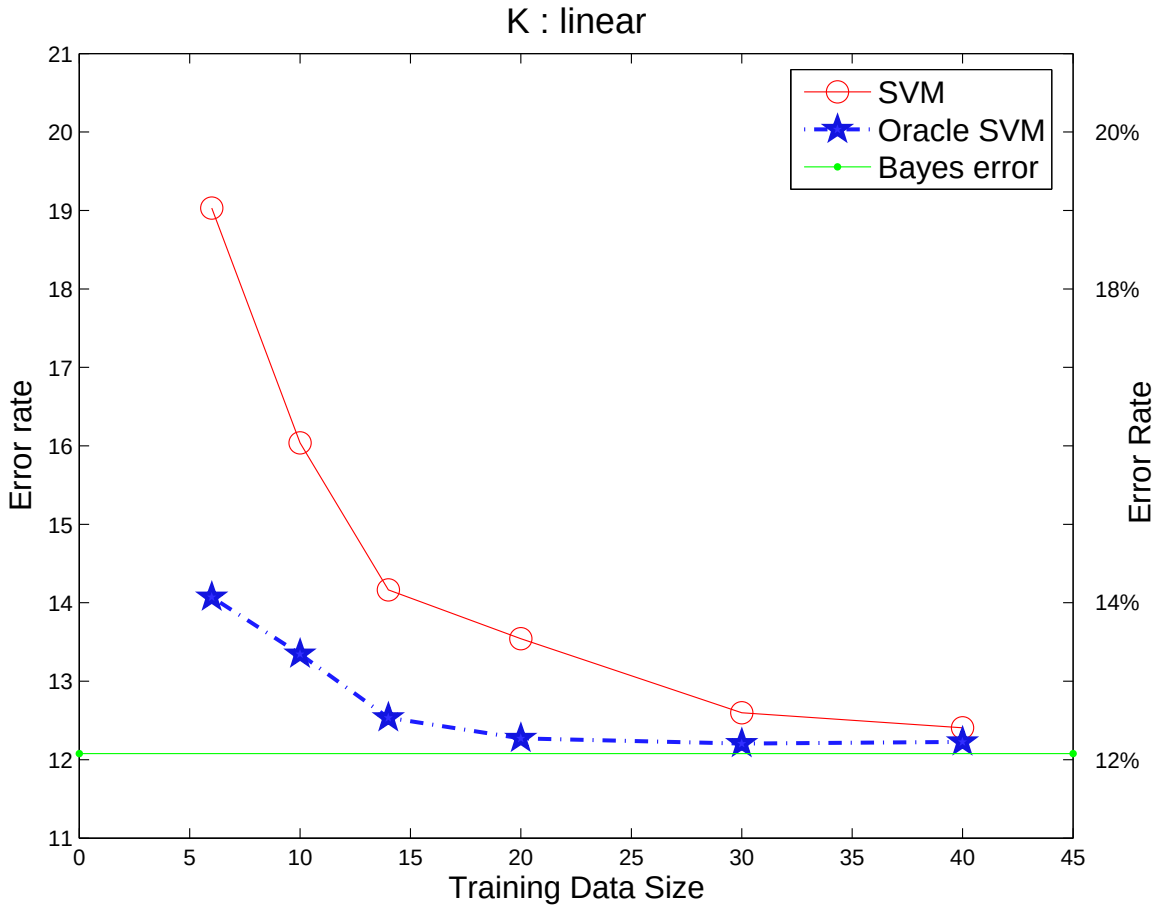
Proposition 1. With probability $1 - \eta$ the following inequality holds

$$P_{err}(w_\ell, b_\ell) \leq P_{err}(w_{best}, b_{best}) + A\frac{n \ln \frac{\ell}{n} - \ln \eta}{\ell}.$$

ILLUSTRATION



ILLUSTRATION



WHAT CAN A TEACHER DO

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

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

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

that approximates the oracle slack function $\xi^0 = \xi^0(x^*)$ well.

(2) 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

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

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

WHAT CAN A TEACHER DO

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

$$R(w, b, \delta) = \sum_{i=1}^{\ell} I(\phi(x_i^*, \delta) - 1)$$

subject to constraints

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

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

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

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

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 the space 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 kernel form

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

DUAL SPACE SOLUTION FOR SVM

The decision function has a form

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

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

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

subject to constraints

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

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

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

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

- 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

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

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

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

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

subject to constraints

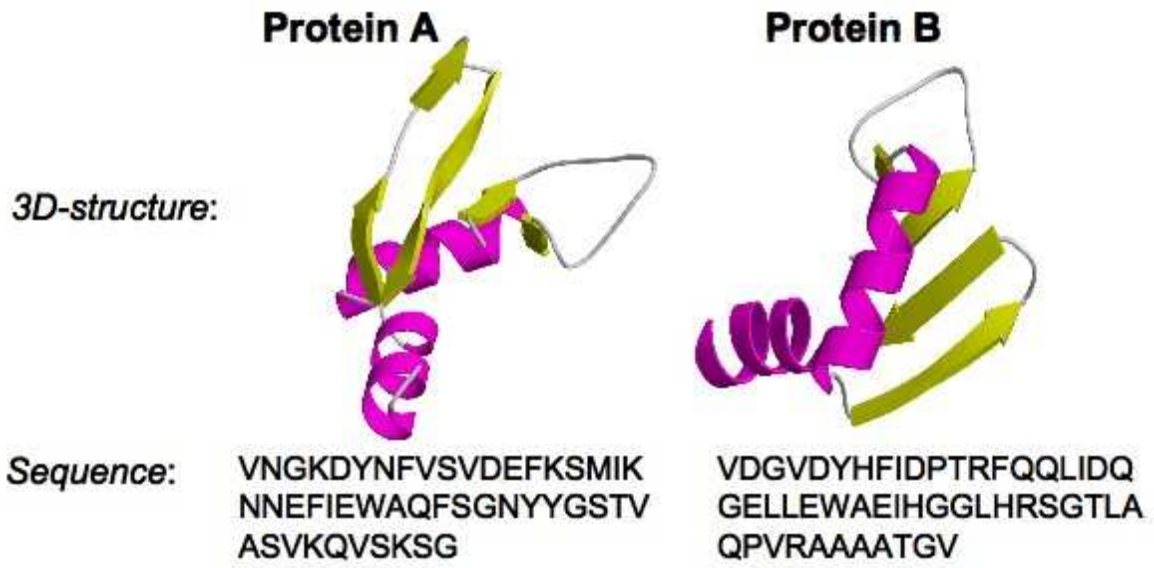
$$\sum_{i=1}^{\ell} \alpha_i y_i = 0,$$

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

ADVANCED TECHNICAL MODEL AS HIDDEN INFORMATION

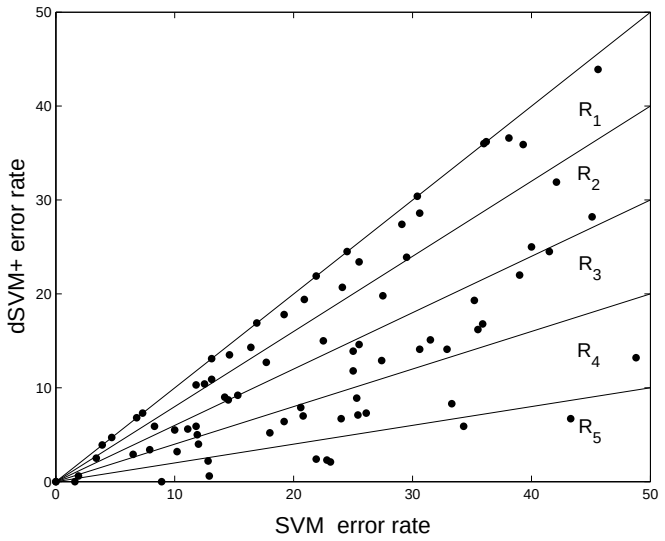
Classification of proteins into families

The problem is : Given amino-acid sequences of proteins classify into families of proteins. The decision space X is the space of amino-acid sequences. The correcting (hidden) space X^* is the space of 3D structure of the proteins.



(Collaboration with Akshay Vashist.)

CLASSIFICATION OF PROTEIN FAMILIES



In 11 cases LUHI does not improve performance (points on the diagonal).
In 15 cases the improvement was small ($\theta < 1.25$ times, region R_1).
In 12 cases the improvement was big ($1.25 \leq \theta \leq 1.6$ times, region R_2).
In 17 cases the improvement was significant ($1.6 \leq \theta \leq 2.5$ times; region R_3).
In 13 cases the improvement was major ($2.5 \leq \theta \leq 5$ times; region R_4).
In 9 cases the improvement was dramatic ($\theta \geq 5$ times; region R_5).

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:

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

(Collaboration with Akshay Vashist.)

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

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

are defined as follows

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

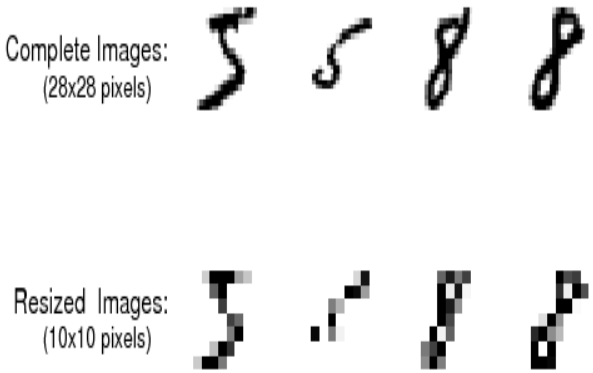
$$x_t^* = (x(t+\Delta-1), x(t+\Delta-2), x(t+\Delta+1), x(t+\Delta+2))$$

ILLUSTRATION

steps ahead, $\Delta =$ training size		1	5	8	steps ahead, $\Delta =$ training size	1	5	8
SVM	100	5.6	11.4	17.6	400	2.4	8.6	14.0
SVM+	100	4.6	8.8	15.9	400	1.8	6.9	12.1
Oracle SVM	100	2.7	8.0	13.2	400	1.5	6.8	8.6
SVM	200	3.0	11.7	15.5	500	2.3	7.9	13.0
SVM+	200	1.8	7.9	12.3	500	1.8	6.3	12.0
Oracle SVM	200	1.8	7.2	12.0	500	1.4	6.2	8.4
$\approx$Bayes	10000	1.1	2.2	4.8	10000	1.1	2.2	4.8

HOLISTIC DESCRIPTION AS HIDDEN 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.

(Collaboration with Akshay Vashist and Natalya Pavlovitch.)

HOLISTIC DESCRIPTION OF THE DIGIT 5

Not absolute two-part creature. Looks more like one impulse. As for two-partness the head is a sharp tool and the bottom is round and flexible. As for tools it is a man with a spear ready to throw it. Or a man is shooting an arrow. He is firing the bazooka. He swung his arm, he drew back his arm and is ready to strike. He is running. He is flying. He is looking ahead. He is swift. He is throwing a spear ahead. He is dangerous. It is slanted to the right. Good snaked-ness. The snake is attacking. It is going to jump and bite. It is free and absolutely open to anything. It shows itself, no kidding. Its bottom only slightly (one point!) is on earth. He is a sportsman and in the process of training. The straight arrow and the smooth flexible body. This creature is contradictory - angular part and slightly roundish part. The lashing whip (the rope with a handle). A toe with a handle. It is an outside creature, not inside. Everything is finite and open. Two open pockets, two available holes, two containers. A piece of rope with a handle. Rather thick. No loops, no saltire. No hill at all. Asymmetrical. No curlings.

HOLISTIC DESCRIPTION OF THE DIGIT 8

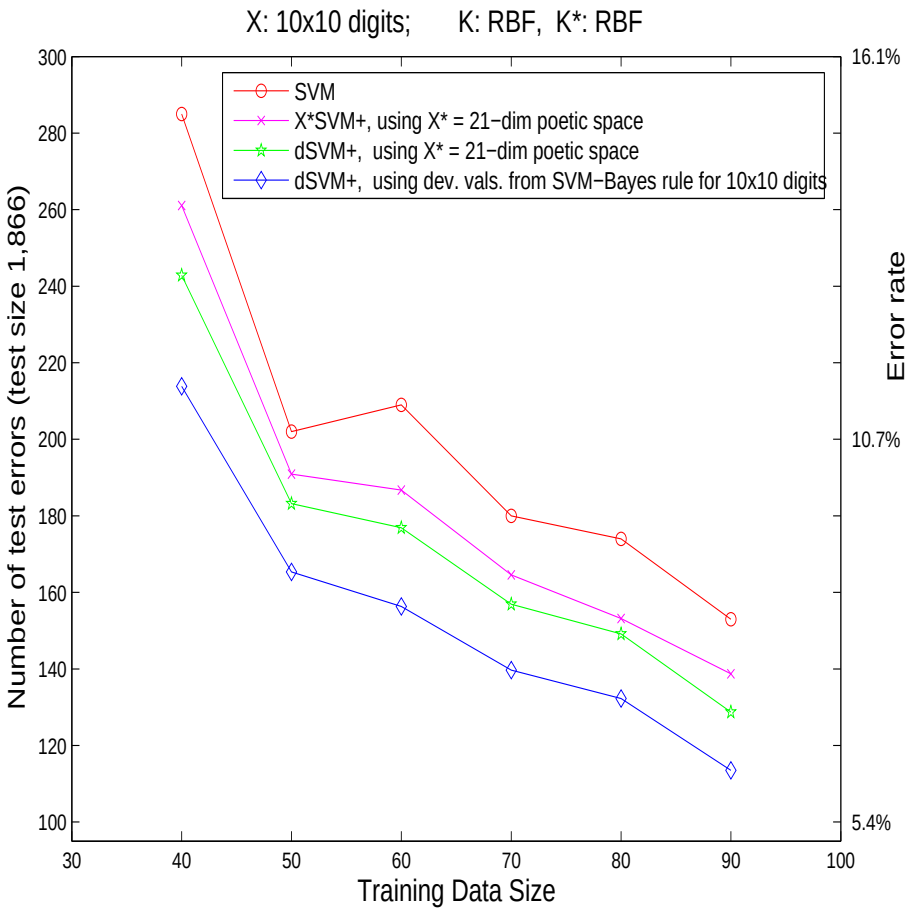
Two-part creature. Not very perfect infinite way. It has a deadlock, a blind alley. There is a small right-hand head appendix, a small shoot. The right-hand appendix. Two parts. A bit disproportionate. Almost equal. The upper one should be a bit smaller. The starboard list is quite right. It is normal like it should be. The lower part is not very steady. This creature has a big head and too small bottom for this head. It is nice in general but not very self-assured. A rope with two loops which do not meet well. There is a small upper right-hand tail. It does not look very neat. The rope is rather good - not very old, not very thin, not very thick. It is rather like it should be. The sleeping snake which did not hide the end of its tail. The rings are not very round - oblong - rather thin oblong. It is calm. Standing. Criss-cross. The criss-cross upper angle is rather sharp. Two criss-cross angles are equal. If a tool it is a lasso. Closed absolutely. Not quite symmetrical (due to the horn).

CODES FOR HOLISTIC DESCRIPTION

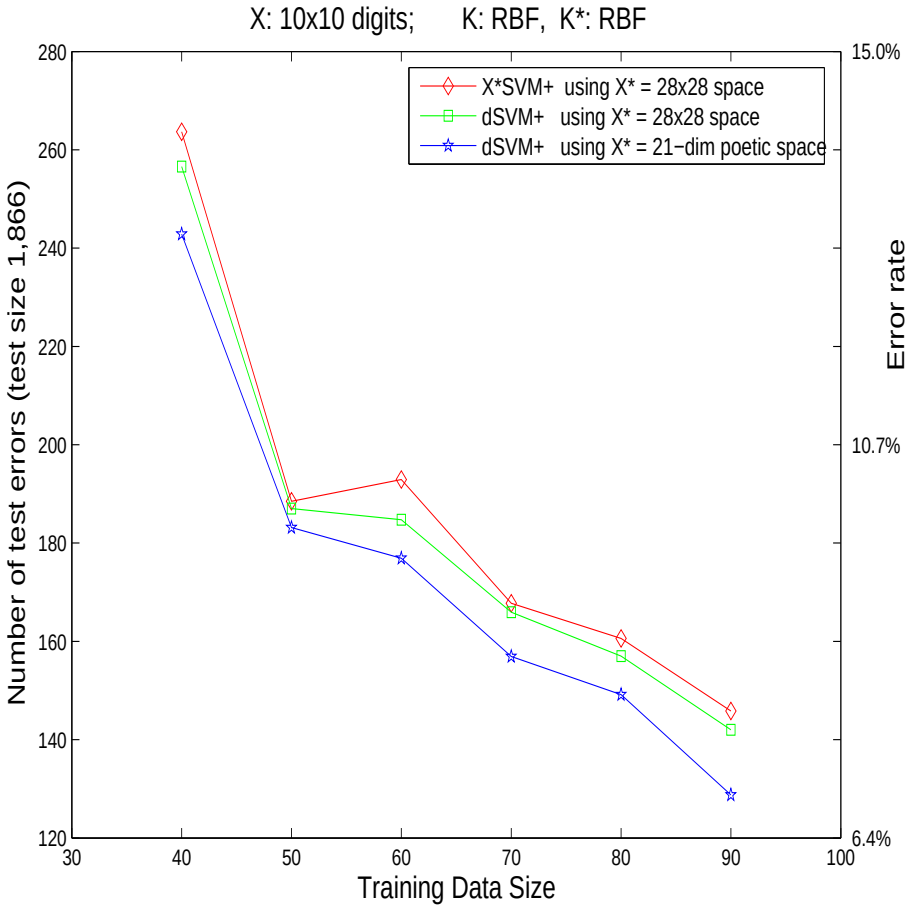
Holistic descriptions were translated into 21-dimensional feature vectors. A subset of these features (with range of possible values) is:

- `two-part-ness` (0 - 5);
- `tilting to the right` (0 - 3);
- `aggressiveness` (0 - 2);
- `stability` (0 - 3);
- `uniformity` (0 - 3),
- `and so on`.

RESULTS


X: 10x10 digits; K: RBF, K*: RBF
SVM
X*SVM+, using X* = 21−dim poetic space
dSVM+, using X* = 21−dim poetic space
dSVM+, using dev. vals. from SVM−Bayes rule for 10x10 digits
Number of test errors (test size 1,866)
Error rate
Training Data Size
300
280
260
240
220
200
180
160
140
120
100
16.1%
10.7%
5.4%
30
40
50
60
70
80
90
100

TECHNICAL SPACE VS. HOLISTIC SPACE



WHAT COULD BE SOURCE OF HIDDEN INFORMATION

- Semi-scientific models (say, use Elliott waves as hidden information to improve methods of stock values prediction).
- Alternative models (say, use Eastern medicine as hidden information to improve rules of Western medicine).
- Non human intelligence (say, use animal intelligence as hidden information to improve human intelligence).

RELATION TO DIFFERENT BRANCHES OF SCIENCE

- **Statistics:** New type of statistical models (say, advanced and future events models in prediction).
- **Cognitive science:** Role of right and left parts of the brain (existence of two different information spaces).
- **Philosophy of Science:** Relation between Simple World and Complex World (unity of exact and metaphoric models).