COMPLETE STATISTICAL THEORY OF LEARNING
LEARNING USING STATISTICAL INVARIANTS

Zepu Xi
Followed Prof. Vladimir N. Vapnik

Sun Yat-sen University

June 21, 2024

PART I

VC THEORY OF GENERALIZATION

THE MAIN QUESTION OF LEARNING THEORY

QUESTION:

When in the set of functions $\{f(x)\}$ we can minimize functional

$$R(f) = \int L(y, f(x))\, dP(x, y), \quad f(x) \in \{f(x)\},$$

if measure $P(x, y)$ is unknown but we are given ℓ iid pairs

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

ANSWER:

We can minimize functional $R(f)$ using data *if and only if* the VC-dimension h of set $\{f(x)\}$ is finite.

DEFINITION OF VC DIMENSION

Let $\{\theta(f(x))\}$ be a set of indicator functions (here $\theta(u) = 1$ if $u \geq 0$ and $\theta(u) = 0$ is $u < 0$).

- VC-dimension of set of indicator functions $\theta(f(x))$ is equal h if h is the maximal number of vectors $x_1, \dots, x_h$ that can be shattered (separated into all 2^h possible subsets) using indicator functions from $\{\theta(f(x))\}$. If such vectors exist for any number h the VC dimension of the set if infinite.
- VC-dimension of set of real valued functions $\{f(x)\}$ is the VC-dimension of the set of indicator functions $\{\theta(f(x) + b)\}$.

TWO THEOREMS OF VC THEORY

Theorem

If set $\{f(x)\}$ has VC dimension h, then with probability $1-\eta$ for all functions $f(x)$ the bound

$$R(f) \leq R^{\ell}_{emp}(l) + \sqrt{e^2 + 4eR^{\ell}_{emp}(f)},$$

hold true, where

$$R^{\ell}_{emp} = \frac{1}{\ell}\sum_{i=1}^{\ell} L(y_i, f(x_i)), e = O\left(\frac{h - \ln\eta}{\ell}\right).$$

Theorem

Let $x, w \in R^n$. The VC dimension h of set of linear indicator functions $\{\theta(x^T w) : ||x||^2 \leq 1, ||w||^2 \leq C\}$ is

$$h \leq \min(C, n) + 1.$$

STRUCTURAL RISK MINIMIZATION PRINCIPLE

To find the desired approximation $f_\ell(x)$ in a set $\{f(x)\}$:

FIRST introduce a structure on a set of functions $\{f(x)\}$

$$\{f(x)\}_1 \subset \{f(x)\}_2 \subset \ldots \subset \{f(x)\}_m \subset \{f(x)\}$$

with corresponding VC-dimension h_k

$$h_1 \leq h_2 \leq \ldots \leq h_m \leq \infty.$$

SECOND chose the function $f_\ell(x)$ that minimizes the bound

$$R(f) \leq R^\ell_{\mathrm{emp}}(f) + \sqrt{e^2 + 4eR^\ell_{\mathrm{emp}}(f)},\, e = O\left(\frac{h_k - \ln\eta}{\ell}\right).$$

1. over elements $\{f(x)\}_k$ (with VC-dimension h_k) and
2. the function $f_\ell(x)$ (with the smallest in $\{f(x)\}_k$ loss $R^\ell_{\mathrm{emp}}(f)$).

FOUR QUESTIONS TO COMPLETE LEARNING THEORY

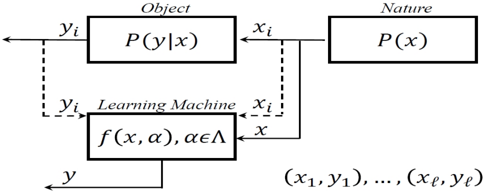
1. How to choose loss function $L(y, f)$ in functional $R(f)$?
2. How to select an admissible set of functions $\{f(x)\}$?
3. How to construct structure on admissible set?
4. How to minimize functional on constructed structure?

The talk answers these questions for pattern recognition problem.

PART II

TARGET FUNCTIONAL FOR MINIMIZATION

SETTING OF PROBLEM: GOD PLAYS DICE


Object
$P(y|x)$
Nature
$P(x)$
y_i
x_i
y_i
Learning Machine
x_i
$f(x,\alpha), \alpha \epsilon \Lambda$
x
y
$(x_1, y_1), \ldots, (x_\ell, y_\ell)$

Given ℓ iid observations

$$(x_1, y_1), \ldots, (x_\ell, y_\ell), x \in X, y \subset \{0, 1\}$$

generated by unknown $P(x, y) = P(y|x)P(x)$, find the rule

$$r(x) = \theta(f_0(x)),$$

which minimizes in a set $\{f(x)\}$ probability if misclassification

$$R_\theta(f) = \int |y - \theta(f(x))| dP(x, y)$$

STANDARD REPLACEMENT OF BASIC SETTING

Using data

$$(x_1, y_1), \ldots, (x_\ell, y_\ell), x \in X, y \subset \{0, 1\}$$

minimize in the set of functions $\{f(x)\}$ the functional

$$R(f) = \int (y - f(x))^2 dP(x, y)$$

(instead of functional $R_\theta(f) = \int |y - \theta(f(x))| dP(x, y)$).
Minimizer $f_0(x)$ of $R(f)$ estimates conditional probability function $f_0(x) = P(y = 1|x)$. Use the classification rule

$$r(x) = \theta(f_0(x) - 0.5) = \theta(P(y = 1|x) - 0.5).$$

PROBLEM WITH STANDARD REPLACEMENT

Minimization of functional $R(f)$ in the set $\{f(x)\}$ is equivalent to minimization of the expression

$$\begin{aligned} R(f) &= \int (y - f(x))^2 \, dP(x,y) \\ &= \int \left[(y - f_0(x)) + (f_0(x) - f(x))\right]^2 dP(x,y) \end{aligned}$$

where $f_0(x)$ minimizes $R(f)$. This is equivalent to minimization

$$R(f) = \int (y - f_0(x)) \, dP(x,y) + \\ \int (f_0(x) - f(x))^2 \, dP(x) + 2 \int (y - f_0(x))(f_0(x) - f(x)) \, dP(x,y).$$

ACTUAL GOAL IS: USING ℓ OBSERVATIONS TO MINIMIZE THE SECOND INTEGRAL, NOT SUM OF LAST TWO INTEGRALS.

DIRECT ESTIMATION OF CONDITIONAL PROBABILITY

1. When $y \subset \{0,1\}$ the conditional probability $P(y=1|x)$ is defined by some real valued function $0 \leq f(x) \leq 1$.
2. From Bayesian formula
$$P(y=1|x)p(x) = p(y=1,x)$$
follows that any function $G(x-x') \in L_2$ defines equation
$$\int G(x-x')f(x')\,dP(x') = \int G(x-x')\,dP(y=1,x') \quad (*)$$
which solution is conditional probability $f(x) = P(y=1|x)$.
3. To estimate conditional probability means to solve the equation (*) when $P(x)$ and $P(y=1,x)$ are unknown but iid data,
$$(x_1,y_1),\ldots,(x_1,y_\ell)$$
generated according to $P(y,x)$, are *given*.
4. Solution of equation (*) is ill-posed problem.

MAIN INDUCTIVE STEP IN STATISTICS

Replace the unknown Cumulative Distribution Function (CDF) $P(x), x = (x^1, \ldots, x^n)^T \in R^n$ with it estimate $P_\ell(x)$: The Empirical Cumulative Distribution Function (ECDF)

$$P_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta\{x - x_i\}, \theta\{x - x_i\} = \prod_{k=1}^{n} \theta\{x^k - x_i^k\}$$

obtained from data

$$x_1, \ldots, x_\ell, x_i = (x_i^1, \ldots, x_i^n)^T.$$

The main theorem of statistics claims that ECDF convergences to actual CDF *uniformly* with fast rate of convergence. The following inequality holds true

$$P\{\sup_x |P(x) - P_\ell(x)| > \epsilon\} < 2\exp\{-2\epsilon^2 l\}, \quad \forall \epsilon.$$

TWO CONSTRUCTIVE SETTINGS OF CLASSIFICATION PROBLEM

1. *Standard constructive setting:* Minimization of functional

$$R_{\text{emp}}(f) = \int (y - f(x))^2 \, dP_\ell(x, y),$$

in a set $\{f(x)\}$ using data $(x_1, y_1), \ldots, (x_\ell, y_\ell)$ leads to

$$R_{\text{emp}}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i))^2, \quad f(x) \in \{f(x)\}.$$

---

2. *New constructive setting:* Solution of equation

$$\int G(x - x') f(x') \, dP_\ell(x') = \int G(x - x') \, dP_\ell(y = 1, x'),$$

using data leads to solution in $\{f(x)\}$ the equation

$$\frac{1}{\ell} \sum_{i=1}^{\ell} G(x - x_i) f(x_i) = \frac{1}{\ell} \sum_{j=1}^{\ell} y_j G(x - x_j), \quad f(x) \in \{f(x)\}.$$

NADARAYA-WATSON ESTIMATOR OF CONDITIONAL PROBABILITY

It is known Nadaraya-Watson estimator of $P(y=1|x)$:

$$f(x) = \frac{\sum_{i=1}^{\ell} y_i G(x-x_i)}{\sum_{i=1}^{\ell} G(x-x_i)}$$

where special kernels $G(x-x_i)$ (say, Gaussian) are used.
This estimator is the solution of ”corrupted” equation

$$\frac{1}{\ell}\sum_{i=1}^{\ell} G(x-x_i)f(x) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i G(x-x_i)$$

(which uses special kernel) rather than the obtained equation

$$\frac{1}{\ell}\sum_{i=1}^{\ell} G(x-x_i)f(x_i) = \frac{1}{\ell}\sum_{j=1}^{\ell} y_i G(x-x_j)$$

(which is defined for any kernel $G(x-x')$ from L_2).

WHAT MEANS TO SOLVE THE EQUATION

To solve the equation

$$\frac{1}{\ell}\sum_{i=1}^{\ell} G(x-x_i)f(x_i) = \frac{1}{\ell}\sum_{j=1}^{\ell} y_j G(x-x_j)$$

means to find the function in $\{f(x)\}$ minimizing L_2-distance

$$R(f) = \int \left(\sum_{i=1}^{\ell} G(x-x_i)f(x_i) - \sum_{j=1}^{\ell} y_j G(x-x_j) \right)^2 d\mu(x)$$

Simple algebra leads to expression

$$R_{\mathcal{V}}(f) = \sum_{i,j=1}^{\ell} (y_i - f(x_i))(y_j - f(x_j))\, v(x_i, x_j),$$

where values $v(x_i, x_j)$ are

$$v(x_i, x_j) = \int G(x-x_i)\, G(x-x_j)\, d\mu(x), \quad i, j = 1, \ldots, l.$$

Values $v(x_i, x_j)$ form $\mathcal{V}$-matrix.

THE $\mathcal{V}$-MATRIX ESTIMATE

1. For $\mu(x) = P(x)$ elements $v(x_i, x_j)$ of $\mathcal{V}$-matrix are

$$v(x_i, x_j) = \int G(x - x_i)\, G(x - x_j)\, dP(x).$$

   Using empirical estimate $P_\ell(x)$ instead of $P(x)$ we obtain the following estimates of elements of $\mathcal{V}$-matrix

$$v(x_i, x_j) = \frac{1}{\ell} \sum_{s=1}^{\ell} G(x_s - x_i)\, G(x_s - x_j)$$

2. For $\mu(x) = x, x \in (-1, 1)$ and $G(x - x') = \exp\{-0.5\delta^2 (x - x')^2\}$,

$$v(x_i, x_j) =$$
$$\exp\{-\delta^2 (x_i - x_j)^2\}$$
$$\{erf[\delta(1 + 0.5(x_i + x_j))] + erf[\delta(1 - 0.5(x_i + x_j))]\}$$

LEAST $\mathcal{V}$-QUADRATIC FORM METHOD AND LEAST SQUARES METHOD

Let $(x_1, y_1), \ldots, (x_\ell, y_\ell)$ be training data. Using notations:

$$Y = (y_1, \ldots, y_\ell)^T, \quad F(f) = (f(x_1), \ldots, f(x_\ell))^T, \ldots, \mathcal{V} = ||v(x_i, x_j)||$$

we can rewrite functional

$$R_{\mathcal{V}}(f) = \sum_{i,j=1}^{\ell} (y_i - f(x_i))(y_j - f(x_j))v(x_i, x_j),$$

in matrix form

$$R_{\mathcal{V}}(f) = (y - F(f))^T \mathcal{V}(y - F(f)).$$

We call this functional *Least $\mathcal{V}$-quadratic* functional.
Identity matrix $\mathcal{I}$ instead of $\mathcal{V}$ forms *Least Squares* functional

$$R_I(f) = (Y - F(f))^T (Y - F(f)).$$

PART III

SELECTION OF ADMISSIBLE SET OF FUNCTIONS

STRONG AND WEAK CONVERGENCE

Functions $f_\ell(x) \in L_2$ have two modes of convergence:

1. Strong mode of convergence (convergence of functions)

$$\lim_{l\to\infty} \int (f_\ell(x) - f_0(x))^2 \, d\mu(x) = 0.$$

2. Weak mode of convergence (convergence of functionals)

$$\lim_{l\to\infty} \int f_\ell(x)\phi(x)\, d\mu(x) = \int f_0(x)\phi(x)\, d\mu(x), \quad \forall \phi(x) \in L_2$$

(convergence for all possible functions$\phi(x) \in L_2$)

- Strong mode of convergence implies weak convergence:

$$\left(\int (f_\ell(x) - f_0(x))\phi(x)\, d\mu(x)\right)^2 \leq \int (f_\ell(x) - f_0(x))^2 \, d\mu(x) \int \phi^2(x)\mu($$

- For functions $f_\ell(x)$ belonging to compact ewak mode of convergence implies strong mode of convergence.

WEAK CONVERGENCE TO CONDITIONAL PROBABILITY FUNCTIONS $P(y=1|x)$

Weak mode convergence of sequence of functions $f_\ell(x)$ to function $f_0(x) = P(y=1|x)$ means equalities

$$\begin{aligned}\lim_{l\to\infty} \int \phi(x) f_\ell(x)\, dP(x) &= \int \phi(x) P(y=1|x)\, dP(x) \\ &= \int \phi(x)\, dP(y=1,x)\end{aligned}$$

for all $\phi(x) \in L_2$.

Let us call set of m functions $\phi_1(x), \ldots, \phi_m(x)$ from L_2 the *chosen predicates*. Let us call subset of functions $\{f(x)\}$ for which the following m equalities hold true

$$\int \phi_k(x) f(x)\, dP(x) = \int \phi_k(x)\, dP(y=1,x), \quad k = 1, \ldots, m,$$

the *admissible set of functions (defined by the predicates)*.

ADMISSIBLE SUBSETS FOR ESTIMATION CONDITIONAL PROBABILITY FUNCTION

Replacing $P(x), P(y = 1, x)$ with $P_\ell(x), P_\ell(y = 1, x)$ we obtain

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \phi_k(x_i) f(x_i) = \frac{1}{\ell} y_i \phi_k(x_i), \quad k = 1, \dots, m.$$

In the matrix notations

$$Y = (y_1, \dots, y_\ell)^T, F(f) = (f(x_1), \dots, f(x_\ell))^T, \Phi_k = (\phi_k(x_1), \dots, \phi_k(x_\ell))^T.$$

we obtain that:
The admissible set of functions $\{f(x)\}$ *satisfies equalities*

$$\Phi_k^T F(f) = \Phi_k^T Y, \quad k = 1, \dots, m.$$

We call these equalities *statistical invariants for* $P(y = 1|x)$.

DUCK TEST,STATISTICAL INVARIANTS, PREDICATES, AND FEATURES

THE DUCK TEST LOGIC

"If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck." (English proverb.)

STATISTICAL INVARIANTS

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\phi_k(x_i)f(x_i) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i\phi_k(x_i), \quad k = 1, \dots, m,$$

(or $\Phi_k^T F(f) = \Phi_k^T Y, k = 1, \dots, m$ in vector notations) collect set of admissible functions $\{f(x)\}$ which "identify" animal as a duck if it "looks, swims, and quacks like a duck".

PREDICATES AND FEATURES

Concepts of predicates and features are very different"

- With increasing number of predicates the VC-dimension of admissible set of functions $\{f(x)\}$ DECREASES.
- With increasing number of features the VC-dimension of admissible set of functions $\{f(x)\}$ INCREASES.

EXACT SETTING OF COMPLETE LEARNING PROBLEM

- The complete solution of classification problem requires: *in a given set of functions $\{f(x)\}$ to minimize functional*

$$R_{\mathcal{V}}(f) = (\,Y - F(f))^T \mathcal{V}(\,Y - F(f)),$$

subject to constraints (statistical invariants)

$$\Phi_k^T F(f) = \Phi_k^T Y, \quad k = 1, \ldots, m.$$

We call this conditional minimization model of learning *Learning Using Statistical Invariants*(LUSI)

- Classical methods require in a given (*specially constructed*) subset of functions $\{f(x)\}$ to minimize the functional

$$R_{\mathcal{I}}(f) = (\,Y - F(f))^T(\,Y - F(f)).$$

APPROXIMATE SETTING OF COMPLETE LEARNING PROBLEM

In this setting minimization of the functional

$$R_{\mathcal{V}}(f) = (Y - F(f))^T \mathcal{V} (Y - F(f)),$$

on the set of functions $\{f(x)\}$ satisfying m constraints

$$\Phi_s^T F(f) = \Phi_s^T Y, \quad s = 1, \ldots, m$$

is replaced with minimization of the functional

$$R_{\mathcal{V}\mathcal{P}}(f) = \hat{\tau}(Y - F(f))^T \mathcal{V} (Y - F(f)) + \frac{\tau}{m} \sum_{s=1}^{m} (\Phi_s^T F(f) - \Phi_s^T Y)^2,$$

where $\hat{\tau}, \tau \geq 0, \hat{\tau} + \tau = 1$. This functional can be rewritten as

$$R_{\mathcal{V}\mathcal{P}}(f) = (Y - F(f))^T (\hat{\tau}\mathcal{V} + \tau\mathcal{P})(Y - F(f)),$$

where $(l \times l)$ matrix $\mathcal{P}$ defines predicates convariance

$$\mathcal{P} = \frac{1}{m} \sum_{s=1}^{m} \Phi_s \Phi_s^T.$$

PART IV

COMPLETE SOLUTION IN REPRODUCING KERNEL HILBERT SPACE (RKHS)

IMPORTANT FACTS FROM RKHS 1.

1. RKHS is set of functions $\{f(x)\}$ for which

$$(K(x,x'), f(x')) = f(x), \quad (K(x,x') \text{ is Mercer kernel}).$$

2. Mercer kernel is defined by orthonormal functions $\psi_k(x)$

$$K(x,x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x'), \quad \lambda_i > 0, \lambda_t \rightarrow_{t\rightarrow\infty} = 0.$$

3. Set of functions

$$f_c(x) = \sum_{i=1}^{\infty} c_i \psi_i(x)$$

with inner product (and norm)

$$(f_c(x), f_{c^*}(x)) = \sum_{i=1}^{\infty} \frac{c_i c_i^*}{\lambda_i} \left(||f_c(x)||^2 = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i} \right)$$

forms RKHS of kernel $K(x,x')$.

IMPORTANT FACTS FROM RKHS 2.

4. **REPRESENTER THEOREM.** Minimum of functional

$$R_{\mathcal{V}}(f) = (Y - F(f))^T \mathcal{V} (Y - F(f))$$

in subset of RKHS with $||f(x)||^2 \leq C$ has representation

$$f_0(x) = \sum_{i=1}^{\ell} a_t K(x_i, x) = A^T \mathcal{K}(x), \quad (*)$$

where $A = (a_1, \ldots, a_\ell)^T$, $\mathcal{K}(x) = (K(x_1, x), \ldots, K(x_\ell, x))^T$.

5. Square of norm of function $f(x)$ in form (*) is

$$||f(x)||^2 = A^T K A, \quad K = ||K(x_i, x_j)||, \quad F(f) = KA.$$

6. Sunset of functions from RKHS with bounded norm $A^T K A \leq C$ has finite VC-dimension (the smaller C, the smaller is VC-dimension). By controlling C, one control both: the VC dimension of subset of functions and their smoothness.

Structure defined by C is the key in implementation SRM principle for functions belonging to RKHS.

CONDITIONAL MINIMIZATION IN RKHS: EXACT LUSI SOLUTION

For RKHS we have $F(f) = KA$. Minimum of the functional

$$R_{\mathcal{V}}(f) = (KA - Y)^T \mathcal{V}(KA - Y),$$

subject to m constraints

$$\Phi_k^T KA = \Phi_k^T Y, \quad k = 1, \ldots, m$$

and constraint

$$A^T KA \leq C$$

has unique solution of the form $f_\ell(x) = A_{\text{LUSI}}^T \mathcal{K}(x)$, where

$$A_{\text{LUSI}} = A_{\mathcal{V}} - \sum_{s=1}^{m} \mu_s A_s, A_{\mathcal{V}} = (\mathcal{V}K + \gamma_c I)^{-1} \mathcal{V} Y, A_s = (\mathcal{V}K + \gamma_c I)^{-1} \Phi_s$$

Parameters μ_s are solution of linear equations

$$\sum_{s=1}^{m} \mu_s A_s^T K \Phi_s = (KA_{\mathcal{V}} - Y)^T \Phi_s, \quad s = 1, \ldots, m.$$

UNCONDITIONAL MINIMIZATION IN RKHS (SOLUTION OF APPROXIMATE SETTING)

Minimum of functional

$$R_{\mathcal{V}\mathcal{P}}(f) = (KA - Y)^T(\hat{\tau}\mathcal{V} + \tau\mathcal{P})(KA - Y)$$

in the set of functions $\{f(x)\}$ belonging to RKHS of kernel $K(x, x')$ with bounded norm

$$A^T KA \leq C$$

has unique solution of the form

$$f_0(x) = A^T_{\mathcal{V}\mathcal{P}}\mathcal{K}(x),$$

where

$$A_{\mathcal{V}\mathcal{P}} = ((\hat{\tau}\mathcal{V} + \tau\mathcal{P})K + \gamma_c I)^{-1}(\hat{\tau}\mathcal{V} + \tau\mathcal{P})\,Y.$$

SVM AND LUSI-SVM ESTIMATIONS IN RKHS

- **SVM:** Given data

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

  find in RKHS the function $f(x) = A^T \mathcal{K}(x)$ with norm

  $$||f(x)||^2 = A^T K A \leq C \quad (*)$$

  that minimizes losses

  $$L(A) = \sum_{i=1}^{\ell} |y_i - A^T \mathcal{K}(x_i)|$$

- **LUSI-SVM:** Given data find in RKHS the function $f(x) = A^T \mathcal{K}(x)$ with bounded norm (*) that minimizes

  $$L(A) = \tau \sum_{s=1}^{m} |A^T K \Phi_s - Y^T \Phi_s| + \hat{\tau} \sum_{i=m+1}^{m+l} |y_i - A^T \mathcal{K}(x_i)|,$$

  where $\tau + \hat{\tau} = 1, \tau > 0, \hat{\tau} > 0$.

LUSI-SVM ESTIMATOR

LUSI-SVM method selects in set $A^TKA \leq C$ the function

$$f(x) = \sum_{i=1}^{\ell} a_i K(x_i, x) = A^T \mathcal{K}(x),$$

where $A = \sum_{t=1}^{m} \delta_t \Phi_t + \sum_{t=m+1}^{m+l} \delta_t \Phi_t$.
To find δ_t one has to maximize the functional

$$R(\delta) = \sum_{i=1}^{m+l} \delta_i \Phi_s^T Y - \frac{1}{2} \sum_{r,s=1}^{m+l} \delta_r^T K \Phi_s \delta_s$$

subject to constraints

$$-\hat{\tau}\gamma_c^* \leq \delta_t \leq \hat{\tau}\gamma_c^*, \quad t = (m+1), \ldots, (m+l),$$
$$-\tau\gamma_c^* \leq \delta_t \leq \tau\gamma_c^*, \quad t = 1, \ldots, m, \hat{\tau} + \tau = 1,$$

where we denoted

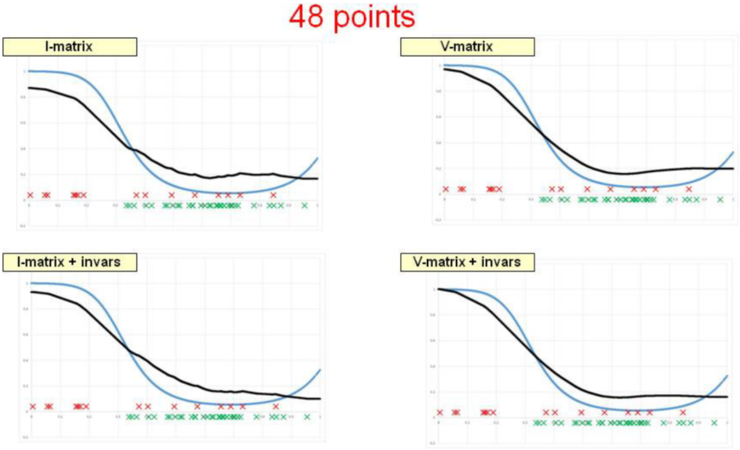
$$\Phi_{m+t} = (0, \ldots, 1, \ldots, 0)^T, t = m+1, \ldots, m+l.$$

LEARNING DOES NOT REQUIRE BIG DATA

According to Representer Theorem, the optimal solution of learning problem in RKHS have properties:

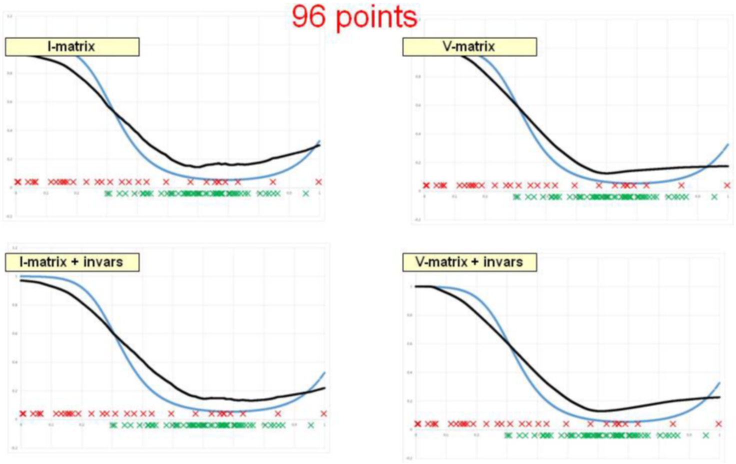
1. It is defined linear parametric functions in form of expansion on kernel functions (i.e optimal solution belongs to one layer network, not multi-layer network).
2. Observation vectors $x_1, \ldots, x_\ell$ and kernel $K(x, x')$ define basis of linear expansion for optimal ℓ parametric solution.
3. SVM: to control VC-dimension uses data to find both the basis of expansion and the parameters of solution.
4. LUSI-SVM: to estimate unknown parameters of solution, adds to ℓ training pairs m pairs $(K\Phi_s,\, Y^T\Phi_s)$ obtained using predicates. When $\tau \approx 1$ it uses just these m pairs.
5. Since any functions from Hilbert space can be used as predicates $\phi_s(x)$, there exist one or several ”smart” (actually ”sophia”) predicates defining pairs $(K\Phi_s,\, Y^T\Phi_s)$ to form optimal solution.

ILLUSTRATION



I: 0.3756 V:0.1432
I&I: 0.2166 V&I: 0.1017

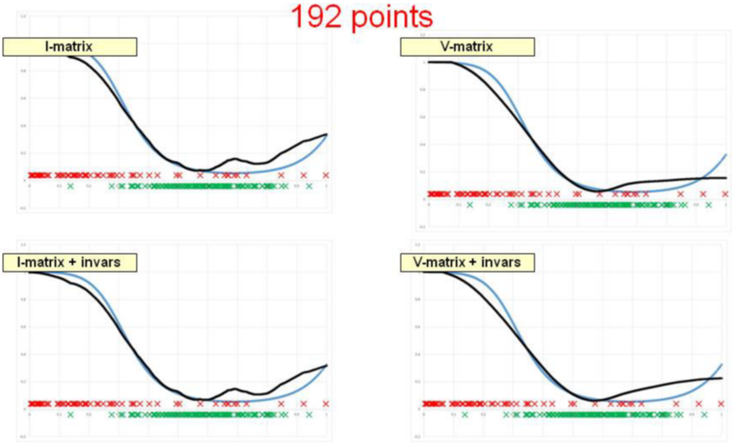
ILLUSTRATION



I: 0.3212 V:0.1207

I&I: 0.1808 V&I: 0.0778

ILLUSTRATION



I: 0.1672 V:0.0689
I&I: 0.1072 V&I: 0.0609

MULTIDIMENSIONAL EXAMPLES

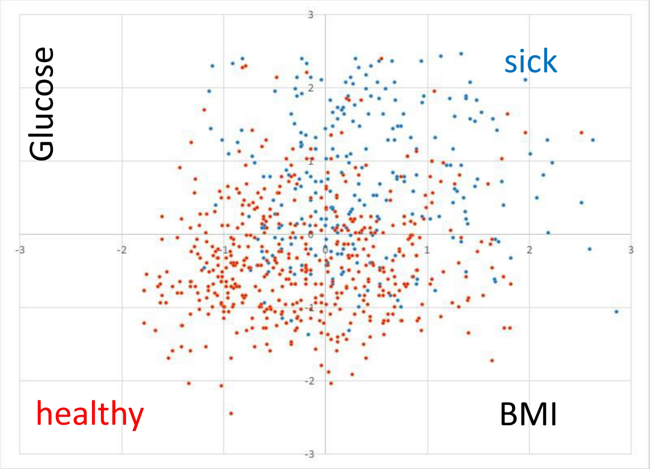
TABLE 1

Data set	Training	Features	SVM	V&$\mathcal{I}$
Diabetes	562	8	25.94%	22.73%
MAGIC	1005	10	19.03%	15.10%
WPBC	134	33	25.48%	23.02%
Bank Marketing	445	16	12.06%	10.58%

TABLE 2

Diabetes			MAGIC		
Training	SVM	V&$\mathcal{I}_9$	Training	SVM	V&$\mathcal{I}$
71	28.42	27.52%	242	20.51	17.35%
151	26.97%	24.56%	491	20.93%	15.91%
304	26.35%	23.78%	955	18.89%	15.19%
612	25.43%	22.60%	1903	18.03%	14.25%

NEW INVARIANT FOR DIABETES



$I\&\mathcal{I}_{+*}$

decreases errors rate from 22.73% to 22.07%.

WAY TO FIND NEW INVARIANT

Find a situation (the box $\mathcal{B}$ in Fig.), where the existing solution (the approximation $P_\ell(y=1|x)$) contradicts the evidence (contradicts invariant for predicate $\phi(x)=1$ inside the box) and then modify the solution (obtain a new approximation $P_{n+1}(y=1|x)$) which resolves this contradiction.

This is the same principle that used in Physics to discover the laws of Nature. To discover laws of the Nature physicists first trying to find a situation where existing theory contradicts observations (The invariants fail. Theoretical predictions do not supported by experiments). Then they trying to reconstruct theory to remove the contradictions. They construct a new approximation of theory which does not contradict the observed reality —— keeps all invariants.

The most important (and most difficult) part in scientific discovery is to find contradictive situation.

PART V

LUSI APPROACH IN NEURAL NETWORKS

$\mathcal{VP}$-BACK PROPAGATION ALGORITHM

- Neural Networks searching for minimum of functional
$$R_{\mathcal{I}}(f) = (F(f) - Y)^T (F(f) - Y),$$
in the set of piece-wise linear functions $\{f\}$ realized by neural network. It uses gradient descent procedure of minimization (called *Back Propagation*). Procedure has three steps: 1. Forward propagation. 2. Backward propagation. 3. Updates of parameters.
- To minimize in the same set of functions the $\mathcal{VP}$-form
$$R_{\mathcal{VP}}(f) = (F(f) - Y)^T (\hat{\tau}\mathcal{V} + \tau\mathcal{P})(F(f) - Y),$$
using back propagation technique, one has to modify just backward step: instead of vector
$E = (y_1 - u_1), \ldots, (y_\ell - u_\ell)^T$, (where $u_1, \ldots, u_\ell$ are outputs of the last layer (last unit) on vector $x_1, \ldots, x_\ell$) one has back propagate modified vector
$$\hat{E} = (\hat{\tau}\mathcal{V} + \tau\mathcal{P})E.$$

SCHEME OF $\mathcal{VP}$-BACK PROPAGATION ALGORITHM

1. *Forward propagation step.* Given initial weights w of Net, propagate training vectors x_i through all hidden layers.
2. *Border conditions for back propagation.* Let u_i be value corresponding to vector x_i propagated on the last layer (unit) and $e_i = (y_i - u_i)$ be difference between target value y_i and obtained value u_i. Consider vector $E = (e_1, \dots, e_\ell)^T$.
3. *Back propagation step.* Back propagation vector

$$\hat{E} = (\hat{\tau}\mathcal{V} + \tau\mathcal{P})E, \mathrm{where} E = (e_1, \dots, e_\ell)^T.$$

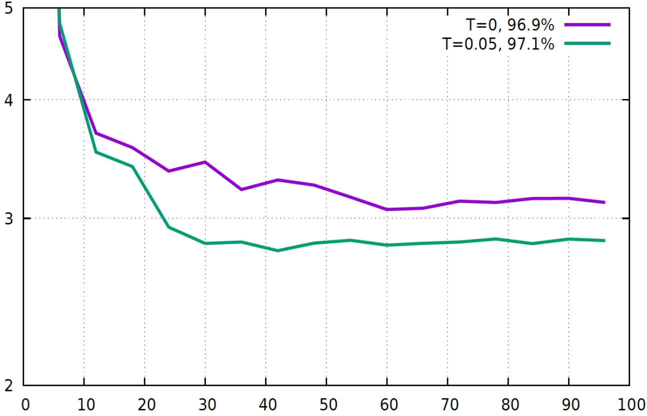
4. *Weights updating step.* Compute gradient of weights and update the weights of the network.

EXAMPLE: MNIST DIGIT RECOGNITION

Minimization of $R(f) = (Y - F(f))^T(\hat{\tau}V + \tau P)(Y - F(f))$.
2D image of digit $\mathbf{u_i}(x^1, x^2)$.

$$\text{Predicate: } \phi(\mathbf{u_i}) = 1.$$

Experiment settings: $\mathcal{V} = I, \ell = 1000$ (100 per class). Batch 6.



Error rate: DNNet - 3.1%, $\mathcal{VP}$-NNet - 2.9%

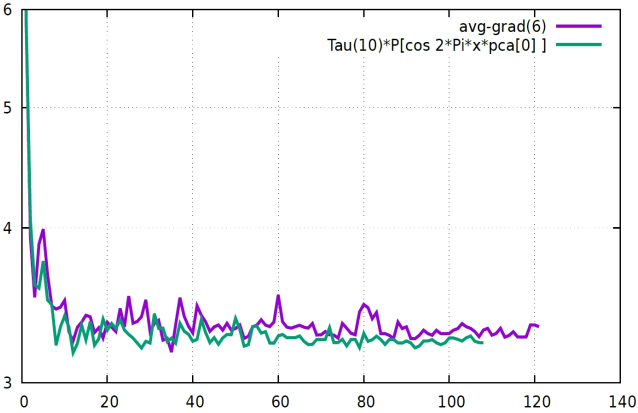
EXAMPLE: MNIST DIGIT RECOGNITION

Minimization of $R(f) = (Y - F(f))^T(\hat{\tau} V + \tau P)(Y - F(f))$.

2D image of digit $u_i(x^1, x^2)$.

Predicate: $\phi(u_i) = \int_0^1 \mathbf{u_i}(x^1, x^2) \cos 2\pi x^1 \, dx^1 \, dx^2$.

Experiment settings: $\mathcal{V} = I, \ell = 1000$ (100 per class). Batch 6.



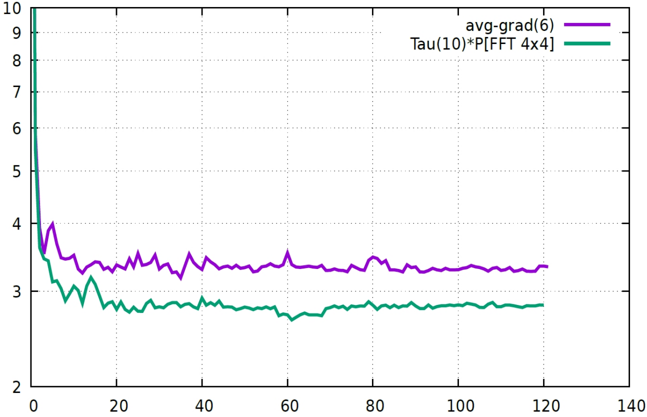
Error rate: DNNet - 3.4%, $\mathcal{VP}$-NNet - 3.3%

EXAMPLE: MNIST DIGIT RECOGNITION

Minimization of $R(f) = (Y - F(f))^T(\hat{\tau} V + \tau P)(Y - F(f))$.

Predicate: $\phi(u_i) = \int_0^1 \mathbf{u_i}(x^1, x^2) \cos \mathbf{m}\pi x^1 \cos \mathbf{n}\pi x^2 \, dx^1 \, dx^2$, m, n = 1, 2, 3, 4.

Experiment settings: $\mathcal{V} = I, \ell = 1000$ (100 per class). Batch 6.



Error rate: DNNet - 3.4%, $\mathcal{VP}$-NNet - 2.8%

STATISTICAL PART OF LEARNING THEORY IS COMPLETED

Theory found that:

1. The functional for minimization defines $\mathcal{V}$-quadratic form

$$R(f) = (Y - F(f))^T \mathcal{V} (Y - F(f)). \quad (1)$$

2. In RKHS, where $F(f) = KA$, the admissible set of functions is defined by invariants for given m predicates function ϕ_k:

$$\Phi_k^T KA = \Phi_k^T Y, k = 1, \ldots, m. \quad (2)$$

3. For RHKS the structure in SRM method is defined by the values of norm of functions from RKHS

$$A^T KA \leq C, \quad (3)$$

   which satisfies (2).

4. There exist unique (closed form) solution for the problem of minimization (1) subject to constraints (2) and (3).

The only question left is "*How to choose set of predicates*"? Answer to this question forms intelligent content of learning.

PART VI

EXAMPLES OF PREDICATES

EXAMPLES OF GENERAL TYPE PREDICATES

$$\sum_{i=1}^{\ell} P_\ell(y=1|x_i)\phi(x_i) = \sum_{i=1}^{\ell} y_i\phi(x_i) \quad (*)$$

1. Predicate $\phi(x) = 1$ in (*) collects functions for which *Expected number of elements of class $y = 1$ computed using $P_\ell(y = 1|x)$ equal to the number of training examples of the first class.*
2. Predicate $\phi(x) = x$ in (*) collects functions for which *Expected center of mass of vectors x of class $y = 1$ computed using $P_\ell(y = 1|x)$ coincides with center of mass of training examples of the first class.*
3. Predicate $\phi(x) = xx^T, x \in R^n$ collects functions for which *Expected $0.5n(n+1)$ values of covariance matrix computed using $P_\ell(y = 1|x)$ coincide with values of covariance matrix computed for vectors x of the first class.*

EXAMPLE OF PREDICATES FOR 2D IMAGES $\{u(x^1, x^2)\}$

Let $2D$ functions

$$u(x^1, x^2), 0 \leq x^1, x^2 \leq \pi$$

describe images and let ℓ pairs

$$(\mathbf{u_1}(x^1, x^2), y_1), \ldots, (\mathbf{u}_\ell(x^1, x^2), y_\ell),$$

from the training set.

1. Predicates

   $$\phi_{r,s}(\mathbf{u_i}) = \int_0^\pi \int_0^\pi \mathbf{u_i}(x^1, x^2) \cos \mathbf{r} x^1 \cos \mathbf{s} x^2 \, dx^1 \, dx^2, r, s = 1, \ldots, N$$

   define coefficients $a_{r,s}$ of cosines expansion of image $\mathbf{u_i}(x^1, x^2)$.
2. For a given function $g(x^1, x^2)$, predicate

   $$\phi(\mathbf{u_i}, x_\mu, x_\nu) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{u_i}(x^1, x^2) g(x^1 - x^1_\mu, x^2 - x^2_\nu) \, dx^1 \, dx^2,$$

   defines value of convolution at point (x^1_μ, x^2_ν).

INSTRUMENTS FOR SPECIAL PREDICATES

LIE DERIVATIVES

Let image is defined by differentiable 2D function $\mathbf{u}(x^1, x^2)$. Consider small linear transformations of 2D space $(x^1, x^2) \in R^2$:

$$\mathbf{t}_\alpha \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} \Longrightarrow \begin{pmatrix} x^1 + a_1 x^1 + a_2 x^2 + a_3 \\ x^2 + a_4 x^2 + a_5 x^1 + a_6 \end{pmatrix}$$

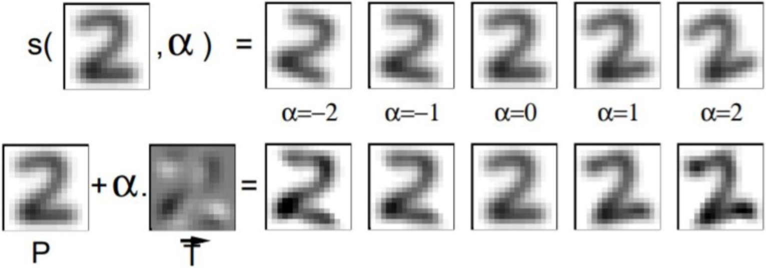
For small a_k, function $\mathbf{u}(\mathbf{t}_a(x^1, x^2))$ in space $\mathbf{t}(x^1, x^2)$ has the following representation in non-transformed space (x^1, x^2):

$$\mathbf{u}(\mathbf{t}_\alpha(x^1, x^2)) \approx \mathbf{u}(x^1, x^2) + \sum_{k=1}^{6} \mathbf{a_k L_k u}(x^1, x^2),$$

where $L_k\mathbf{u}(x^1, x^2)$ are the so-called ***Lie derivatives***. [1]

[1] Please make sense for the real meaning of this statements.

ILLUSTRATION



LIE OPERATIONS

ILLUSTRATION

INVARIANTS WITH RESPECT TO LINEAR TRANSFORMATIONS

TANGENT DISTANCE

EXAMPLES OF PREDICATES THAT DEFINE DEGREE OF SYMMETRIES

CONCLUSIVE REMARKS

- Complete statistical methods of learning require, using structural risk minimization principle, in a given set of functions $\{f(x)\}$ minimize functional

  $$R_{\mathcal{V}}(f) = (Y - F(f))^T \mathcal{V} (Y - F(f))$$

  subject to invariant constraints

  $$\Phi_s^T F(f) = \Phi_s^T Y.$$

- LUSI method provides unique solution of this problem for functions from RKHS and approximation for Neural Nets.
- Further progress in learning theory goes beyond statistical reasoning. It goes in the direction of search of predicates which form basis for understanding of problems existing in the World (see Plato-Hegel-Wigner line of philosophy).
- Predicates are abstract ideas, while invariants that are built using them from elements of solution. These two concepts reflect essence of intelligence, not just its imitation.

THE CHALLENGE

PLATO-HEGEL-WIGNER LINE OF PHILOSOPHY

In 1928 Vladimir Propp published book ”Morphology of the Folktale” where he described 31 predicates that allow to synthesize Russian folk tales. Later his morphology has been successfully applied to other types of narrative, be it in *literature, theater, film, television series, games, etc.* (although Propp applied it only to the wonder or fairy tale). (See Wikipedia: Vladimir Propp.)

The idea is that *World of Ideas* contains small amount of ideas (predicates) that can be translated in *Worlds of Things* by many different invariants.

Propp found 31 predicates which describe different actions of people in Real World. Probably there exist a small amount of predicates that describe 2D *Real World* images. The challenge is to find them (to understand *World of 2D images*).