FROM ROSENBLATT’S LEARNING MODEL
TO
LEARNING USING PRIVILEGED
INFORMATION

Vladimir Vapnik

Columbia University, NEC-labs
THE ROSENBLATT’S SCHEME:

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

\[(x_1, y_1), \ldots (x_\ell, y_\ell)\]

construct a separating hyperplane in space $Z$

\[
y = \text{sgn}(wz + b)\]

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)
1. There exist two and only two factors responsible for generalization:
   a) The percent of training errors $\nu_{\text{train}}$.
   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

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

where $\ell$ is the number of observations.
2b. When $\nu_{\text{train}} = 0$ the following bounds are valid

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

The bounds are achievable.
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

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

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

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

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 1: Large margin.

Minimize the functional

$$R = (w, w)$$

subject to the constraints

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

The solution $$(w_\ell, b_\ell)$$ has the bound

$$P_{test} \leq O^* \left(\frac{VCdim}{\ell}\right).$$
Generalization 2: Nonseparable case.

Minimize the functional

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

subject to constraints

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

The solution \((w_\ell, b_\ell)\) has the bound

\[ P_{test} \leq \nu_{train} + O^* \left( \sqrt{\frac{VCdim}{\ell}} \right). \]
• In the separable case using $\ell$ examples one estimates $n$ parameters of $w$.

• In the non-separable case one estimates $n + \ell$ parameters ($n$ parameters of vector $w$ and $\ell$ 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 $\delta$ 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 $\ell$ observations.

Can the rate of convergence in this case be faster?
Suppose we are given triplets
\[(x_1, \xi^0_1, y_1), \ldots, (x_\ell, \xi^0_\ell, y_\ell),\]
where \(\xi^0_i = \xi^0(x_i), \ i = 1, \ldots, \ell\) are the slack values with respect to the best hyperplane. Then to find the approximation \((w_{\text{best}}, b_{\text{best}})\) we minimize the functional
\[R(w, b) = (w, w)\]
subject to constraints
\[y_i[(w, x_i) + b] \geq r_i, \ r_i = 1 - \xi^0_i(x_i), \ i = 1, \ldots, \ell.\]

**Proposition 1.** For Oracle SVM the following bound holds
\[P_{\text{test}} \leq \nu_{\text{train}} + O^* \left( \frac{VCdim}{\ell} \right).\]
Sample Training Data

$x_1 \rightarrow$

$x_2 \rightarrow$

class II

class I
ILLUSTRATION — II

K : linear

Error rate vs Training Data Size

- SVM
- Oracle SVM
- Bayes error

Error Rate

Training Data Size

11% 12% 13% 14% 15% 16% 17% 18% 19% 20%

12% 13% 14% 15% 16% 17% 18% 19% 20%

0 5 10 15 20 25 30 35 40 45

Error Rate

Training Data Size
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
  $$\xi_i = \xi(x_i^*, \delta_{\text{best}})$$
  that approximates the oracle slack function $\xi^0 = \xi^0(x^*)$ well.

- During training process supply students with triplets
  $$(x_1, x_1^*, y_1), \ldots, (x_\ell, x_\ell^*, y_\ell)$$
  in order to estimate simultaneously both the correcting (slack) function
  $$\xi = \xi(x^*, \delta_\ell)$$
  and the decision hyperplane (pair $(w_\ell, b_\ell)$).
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 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 $O\left(\frac{1}{\sqrt{\ell}}\right)$ to $O\left(\frac{1}{\ell}\right)$.

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

\[ (x_1, y_1), \ldots, (x_\ell, y_\ell) \]

into the pairs

\[ (z_1, y_1), \ldots, (z_\ell, y_\ell) \]

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

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

• Use inner product in \( Z \) space in the form

\[
(z_i, z_j) = K(x_i, x_j).
\]
The decision function has a form

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

(2)

where \( \alpha_i \geq 0, \ i = 1, \ldots, \ell \) are values which maximize the functional

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

(3)

subject to constraints

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

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 $(x_1, x_1^*, y_1), \ldots, (x_\ell, x_\ell^*, y_\ell)$ into the triplets $(z_1, z_1^*, y_1), \ldots, (z_\ell, z_\ell^*, y_\ell)$ 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
  \[ y_i[(w, z_i) + b] \geq 1 - [(w^*, z_i^*) + b^*], \quad i = 1, \ldots, \ell. \]
- Use inner products in $Z$ and $Z^*$ spaces in the kernel form
  \[ (z_i, z_j) = K(x_i, x_j), \quad (z_i^*, z_j^*) = K^*(x_i^*, x_j^*). \]
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 \( \alpha_i, \ i = 1, \ldots, \ell \) 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

\[ \sum_{i=1}^{\ell} \alpha_i y_i = 0, \quad \sum_{i=1}^{\ell} (\alpha_i - \beta_i) = 0. \]

and the constraints

\[ \alpha_i \geq 0, \quad 0 \leq \beta_i \leq C \]
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.
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+)

Improvement of SVM+ over SVM (error rate decrease)
### Classification of Proteins: Details

#### SVM vs SVM+ vs SVM (3D)

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

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

SVM+ provides significant improvement over SVM (several times)
Time series prediction

Given pairs

\[(x_1, y_1), \ldots, (x_\ell, y_\ell),\]

find the rule

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

where

\[x_t = (x(t), \ldots, x(t - m)).\]

For regression model of time series:

\[y_t = x(t + \Delta).\]

For classification model of time series:

\[y_t = \begin{cases} 
1, & \text{if } x(t + \Delta) > x(t), \\
-1, & \text{if } x(t + \Delta) \leq x(t).
\end{cases}\]
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 \(\tau\) (delay) are parameters.

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

\[
x_t = (x(t), x(t - 1), x(t - 2), x(t - 3))
\]

\[
x_t^* = (x(t + \Delta - 1), x(t + \Delta - 2), x(t + \Delta + 1), x(t + \Delta + 2))
\]
INTERPOLATION AND EXTRAPOLATION

Extrapolation of trends has to face large conditional variance

Interpolation of trends faces small conditional variance
ILLUSTRATION

steps ahead: $\Delta = 1$

steps ahead: $\Delta = 5$

steps ahead: $\Delta = 8$

- SVM
- SVM+
- Oracle SVM
- Bayes rate (SVM for 10,000)
Classification of digit 5 and digit 8 from the NIST database.

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

5

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.

8

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),
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.privileged
RESULTS

X: 10x10 digits; both RBF kernels

Percentage of test errors (test size 1,800)

Training data size

SVM
SVM+
Oracle SVM
Bayes rate (SVM for 10,000)
HOLISTIC SPACE VS. ADVANCED TECHNICAL SPACE
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 \( \alpha_i, \ i = 1, \ldots, \ell \) 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
\[ \sum_{i=1}^{\ell} \alpha_i y_i = 0, \quad \sum_{i=1}^{\ell} (\alpha_i - \beta_i) = 0. \]

and the constraints
\[ \alpha_i \geq 0, \quad 0 \leq \beta_i \leq C \]
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.
THREE CRITICAL POINTS IN PHILOSOPHY OF SCIENCE AND MACHINE LEARNING

1. REALISM and INSTRUMENTALISM in Philosophy of Science and Machine Learning (discussions of 1960).

2. PRINCIPLE OF INDUCTION in Philosophy of Science and Machine Learning (discussions of 1990).

3. REDUCTIONISM and HOLISM in Philosophy of Science and Machine Learning (oncoming discussions).
• 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).


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

Concepts of (Western) LOGOS and (Eastern) HARMONY. Facts that could be important in philosophy of Complex World.