# SVM+

Learning using privileged information
Learning with teacher

content adapted from:

V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information", Neural Networks, 2009, pp.544-557.

Zoya Gavrilov
Nov. 9, 2012
Vision Reading Group Presentation
CSAIL

training data

## SVM

$$(x_1, y_1), \ldots, (x_l, y_l), x_i \in X, y_i \in \{-1, 1\}$$

LUPI (learning using privileged information) paradigm for SVM:

## SVM+

$$(x_1, x_1^*, y_1), \ldots, (x_l, x_l^*, y_l), x_i \in X, \boxed{x_i^* \in X^*}, y_i \in \{-1, 1\}$$

additional information about training instances provided only during training (NOT available during testing), hence "privileged"

# examples of privileged information

1)
y: outcome of a treatment in a year
x: current symptoms of a patient
x*: development of symptoms in 3 months, 6 months, 9 months

2)
y: whether a biopsy image is cancerous or non-cancerous
x: images described in pixel space
x*: report by a pathologist describing the pictures using a high level holistic language
Goal: find a good classification rule in pixel space to make an accurate diagnosis without consulting with a pathologist

3)
y: prediction of whether exchange rate will go up or down at moment t
x: observations about the rate before moment t
x*: (obtained form historical data) observations about rates after moment t

# Motivation: parameter estimation

$$R(w, b, \epsilon) = \frac{1}{2} < w, w > +C \sum_{i=1}^{l} \epsilon_i$$

convergence rate:

in separable case, have to estimate n parameters of w $O(h/l)$

to find optimal hyperplane in non-separable case, one has to estimate extra terms corresponding to the slack variables (as many as the training instances), for a total of n + l parameters $O(\sqrt{h/l})$

h:VC dimension of admissible set of hyperplanes

# Oracle SVM

suppose there exists an oracle function: $\epsilon_i = \epsilon(x_i)$

such that: $y_i[< w, z_i > +b] \geq 1 - \epsilon_i \qquad \forall i = 1, \ldots, l$

in the corresponding SVM+ setting, let the teacher supply us with triplets:

$$(x_1, \epsilon_1, y_1), \ldots, (x_l, \epsilon_l, y_l)$$

then just as in separable case, have to estimate only n parameters

functional to minimize: $R(w, b, \epsilon) = \frac{1}{2} < w, w >$ $\qquad z_i = \phi(x_i) \in Z$

subject to: $y_i[< w, z_i > +b] \geq r_i$ known

$\forall i = 1, \ldots, l$

# Motivation: parameter estimation

convergence rate:

in separable case, have to estimate n parameters of w $O(h/l)$

to find optimal hyperplane in non-separable case, one has to estimate extra terms corresponding to the slack variables (as many as the training instances), for a total of n + l parameters $O(\sqrt{h/l})$

Oracle SVM case: $O(\sqrt{h^*/l})$

h: VC dimension of admissible set of hyperplanes
h*: VC dimension of admissible set of correcting functions

## SVM

functional to minimize: $R(w,b,\epsilon) = \frac{1}{2} < w,w > +C\sum_{i=1}^{l}\epsilon_i$ $\quad z_i = \phi(x_i) \in Z$

subject to: $y_i[< w, z_i > +b] \geq 1 - \epsilon_i$

$\epsilon_i \geq 0, i = 1,\ldots,l$

inner product defined in Z space

## SVM+

$\epsilon_i = [(w^*, z_i^*) + b^*]$

$R(w,w^*,b,b^*) = \frac{1}{2} < w,w > +\frac{\gamma}{2} < w^*,w^* > +C\sum_{i=1}^{l}[< w^*, z_i^* > +b^*]$

$y_i[< w, z_i > +b] \geq 1 - [(w^*, z_i^*) + b^*]$ $\quad z_i^* = \phi^*(x_i^*) \in Z^*$

$[< w^*, z_i^* > +b^*] \geq 0, i = 1,\ldots,l$

## primal

## SVM

$$L(w, b, \epsilon, \alpha, \beta) = \underset{\text{min}}{} \underset{\text{max}}{} \frac{1}{2} < w, w > + C \sum_{i=1}^{l} \epsilon_i - \sum_{i=1}^{l} \alpha_i (y_i [< w, z_i > + b] - 1 + \epsilon_i) - \sum_{i=1}^{l} \beta_i \epsilon_i$$

## SVM+

$$\epsilon_i = [(w^*, z_i^*) + b^*]$$

$$L(w, b, w^*, b^*, \alpha, \beta) = \underset{\text{min}}{} \underset{\text{max}}{} \frac{1}{2} < w, w > + \frac{\gamma}{2} < w^*, w^* > + C \sum_{i=1}^{l} [< w^*, z_i^* > + b^*]$$
$$- \sum_{i=1}^{l} \alpha_i (y_i [< w, z_i > + b] - 1 + [< w^*, z_i^* > + b^*]) - \sum_{i=1}^{l} \beta_i [< w^*, z_i^* > + b^*]$$

very similar mathematically: quadratic optimization problem with similar constraints, but requires tuning 4 hyperparameters (instead of 2)

# dual

## SVM

dual: $\underset{\text{max}}{R(\alpha)} = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$

subject to: $\sum_{i=1}^{l} \alpha_i y_i = 0 \quad 0 \le \alpha_i \le C$

## SVM+

$$\underset{\text{max}}{R(\alpha, \beta)} = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \frac{1}{2\gamma} \sum_{i,j=1}^{l} (\alpha_i + \beta_i - C)(\alpha_j + \beta_j - C) K^*(x_i^*, x_j^*)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \sum_{i=1}^{l} (\alpha_i + \beta_i - C) = 0 \qquad \alpha_i \ge 0, \beta_i \ge 0$$

# admissible SVM+ solutions contain SVM solution

$$R(\alpha,\beta) = \underset{\text{max}}{} \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) - \frac{1}{2\gamma}\sum_{i,j=1}^{l}(\alpha_i+\beta_i-C)(\alpha_j+\beta_j-C)K^*(x_i^*,x_j^*)$$

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

consider case: $\gamma \downarrow 0$ reject privileged information → similarity measures in correcting space not appropriate

then max of: $R(\alpha,\beta)$

occurs when: $(\alpha_i+\beta_i-C)=0, \forall i = 1,..,l$

back to: $\sum_{i=1}^{l}\alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C$

# decision and correcting functions

2 different kernels define similarity measures between objects in 2 different spaces

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

$x_i \in X, i = 1,\ldots,l$ — decision space

$x_i^* \in X^*, i = 1,\ldots,l$ — correcting space

decision function: $$f(x) = \sum_{i=1}^{l} y_i\alpha_i \boxed{K(x_i,x)} + b$$

depends directly only on kernel in decision space

BUT alpha depends on similarity measures in both spaces

correcting function: $$f^*(x^*) = \frac{1}{\gamma}\sum_{i=1}^{l}(\alpha_i+\beta_i-C)K^*(x_i^*,x^*) + b^*$$

## Extension: non-smooth model for slacks

$$\xi_i = [(w^*, z_i^*) + b^*] + \xi_i^*, \quad i = 1, \ldots, \ell,$$

$$(w^*, z_i^*) + b^* \geq 0, \quad \xi_i^* \geq 0, \ i = 1, \ldots, \ell.$$

Our goal is to minimize the functional

$$R(w, w^*, b, b^*, \xi^*) = \frac{1}{2}[(w, w) + \gamma(w^*, w^*)]$$

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

subject to constraints

$$y_i[(w, z_i) + b] \geq 1 - [(w^*, z_i^*) + b^*] - \xi_i^*,$$

$$[(w^*, z_i^*) + b^*] \geq 0,$$

$$\xi_i^* \geq 0.$$

## Extension: privileged information not available for all examples

$$R(w, w^*, b, b^*, \xi) = \frac{1}{2}[(w, w) + \gamma(w^*, w^*)] + C\sum_{i=1}^{n}\xi_i + C^*\sum_{i=n+1}^{\ell}[(w^*, z_i^*) + b^*]$$

subject to constraints

$$y_i[(w, z_i) + b] \geq 1 - \xi_i, \quad i = 1, \ldots, n,$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n,$$

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

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

## Extension: multi-space privileged information

$$R(w, w^*, w^{**}, b, b^*, b^{**}) = \frac{1}{2}[(w, w) + \gamma((w^*, w^*) + (w^{**}, w^{**}))] + C\sum_{i=1}^{n}[(w^*, z_i^*) + b^*] + C\sum_{i=n+1}^{\ell}[(w^{**}, z_i^{**}) + b^{**}]$$

subject to the constraints

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

# dSVM+

**Step 1** Consider the conjugate problem of finding the decision rule in the space $X^*$ by minimizing the functional

$$R(w^*, b^*, \xi^*) = \frac{1}{2}(w^*, w^*) + C\sum_{i=1}^{\ell}\xi_i^*$$

subject to the constraints

$$y_i[(w^*, x_i^*) + b^*] \geq 1 - \xi_i^*, \qquad \xi_i^* \geq 0,\ i = 1, \ldots, \ell$$

(using the classical SVM approach in $X^*$ space). Let $w_\ell^*$ and $b_\ell^*$ be the solution to this problem.

**Step 2** Using the solution to this problem define the so-called *deviation values*

$$d_i = 1 - y_i[(w_\ell^*, x_i^*) + b_\ell^*].$$

**Step 3** Construct a new set of triplets of training data

$$(x_1, d_1, y_1), \ldots, (x_\ell, d_\ell, y_\ell)$$

(use deviation value $d$ as privileged information instead of vector $x^*$).

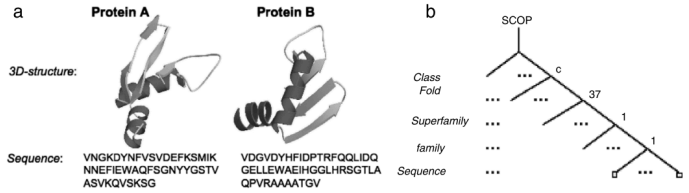
# examples: advanced technical model as privileged information

**problem statement:** classification of proteins (hierarchical scheme of organization, to define evolutionary relations)
**input:** amino-acid sequences
**output:** classification (position in hierarchy)

note: human experts construct hierarchies according to 3D protein structures

**privileged information:** 3D-structures
- obtaining this is a very hard and time consuming problem

**similarity measures:** profile kernel for matching amino-acid sequences; MAMMOTH measure for matching 3D structures



# examples: future events as privileged information

**problem statement:** time series prediction
**input:** historical information about the values of time series up to moment t
**output:** (quantitative prediction - regression framework) value of time series at moment t + del(t); OR (qualitative prediction - pattern recognition) whether time series at moment t + del(t) will be larger/smaller than at moment t

note: human experts construct hierarchies according to 3D protein structures

**privileged information:** future events

**Table 1**
Error rates of SVM, X*SVM+, dSVM+, and Oracle SVM on qualitatively predicting the Mackey–Glass series.

| Steps ahead, $T$ = Training size | | 1 | 5 | 8 | Steps ahead, $T$ = Training size | 1 | 5 | 8 |
|---|---|---|---|---|---|---|---|---|
| SVM | 100 | 2.7 | 5.2 | 8.2 | 400 | 2.2 | 3.5 | 5.2 |
| X*SVM+ | 100 | 2.4 | 5.0 | 7.8 | 400 | 1.8 | 3.1 | 4.7 |
| dSVM+ | 100 | 2.0 | 4.8 | 7.2 | 400 | 1.7 | 2.9 | 4.3 |
| Oracle SVM | 100 | 1.6 | 2.9 | 5.3 | 400 | 1.2 | 1.8 | 2.8 |
| SVM | 200 | 2.5 | 4.9 | 7.3 | 500 | 2.1 | 3.2 | 5.0 |
| X*SVM+ | 200 | 2.1 | 4.6 | 6.8 | 500 | 1.7 | 3.1 | 4.5 |
| dSVM+ | 200 | 1.9 | 3.8 | 6.5 | 500 | 1.7 | 2.7 | 4.2 |
| Oracle SVM | 200 | 1.2 | 2.2 | 4.6 | 500 | 1.1 | 1.5 | 2.7 |
| ≈Bayes (SVM with 10,000 training examples) | | 0.3 | 0.5 | 0.6 | | 0.3 | 0.5 | 0.6 |

# examples: holistic description as privileged information

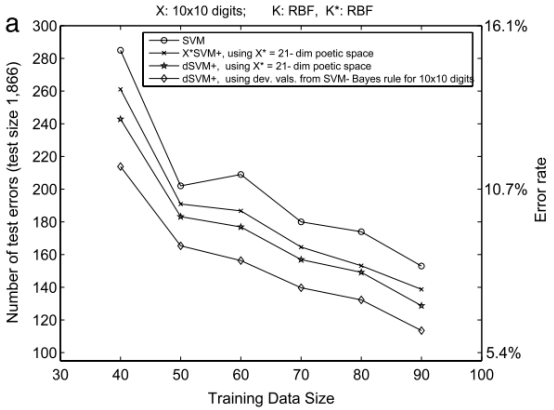
**problem statement:** MNIST digit recognition (5 vs 8)
**input:** 10 by 10 pixel images
**output:** classification

*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.*

**privileged information:** holistic (poetic) descriptions provided by an independent expert

translated into 21-dimensional feature vectors, with entries like: two-part-ness (0-5), tilting to the right (0-3), aggressiveness (0-2), stability (0-3), uniformity (0-3), etc.

# examples: holistic description as privileged information



## final remarks by Vapnik

"We considered a new learning paradigm, the LUPI paradigm which allows one to introduce in the machine learning process, human elements of teaching: teacher's remarks, explanations, analogy, and so on."

"These sort of ideas lead to an integration, in learning techniques, of elements of an exact science and humanities, an exact science and emotions..."