Optimization and Support Vector Machines

ELAVIO 1–5 July 2019, LLeida

Jordi Castro
GNOM - Mathematical Optimization Group

Dept. of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona

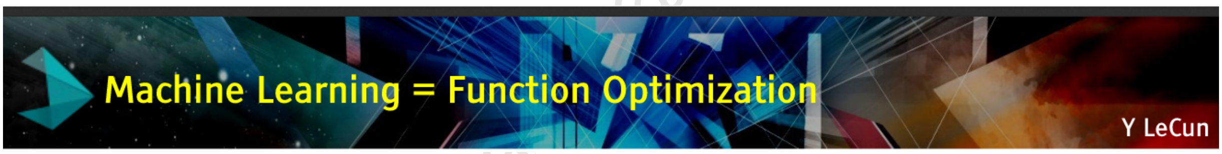
Contents

Multivariate data classification/regression by Support Vector Machines (SVM)

- Learning methodology: synthesize models from examples.
- Supervised learning: learning from input/output (x/y) pairs.
- Learning algorithm: finds a function relating outputs to inputs ($f(x) = y$).
- Depending on type of output we have:
  - Binary classification: two classes of output ($0/1$, $+1/-1$).
  - Multiclass classification: > 2 classes.
  - Regression: continuous output.
- SVM: supervised method based on hyperplanes for binary/multiclass classification or regression.

Optimization is instrumental in machine learning!

This is from a presentation at CERN (Switzerland) on 24 March 2016 by Yann Le Cun:

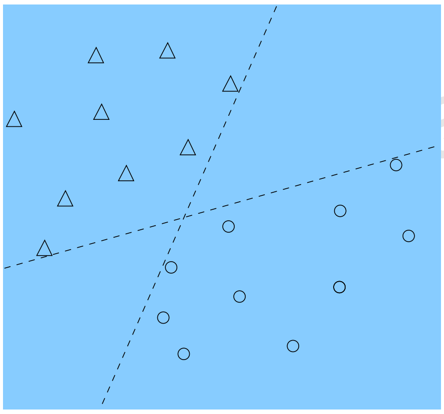


Yann Le Cun, Facebook AI Research Director, Center for Data Science, NYU Courant Institute of Mathematical Sciences, NYU. http://yann.lecun.com.

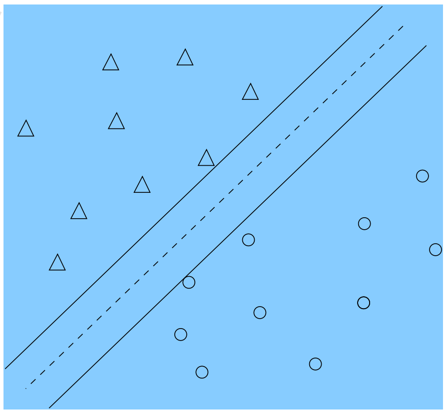
Purpose of SV Classifiers

- To find two parallel hyperplanes separating two classes such that we both minimize the classification error and maximize the margin between the two separating hyperplanes:

Bad classifiers



Good classifier

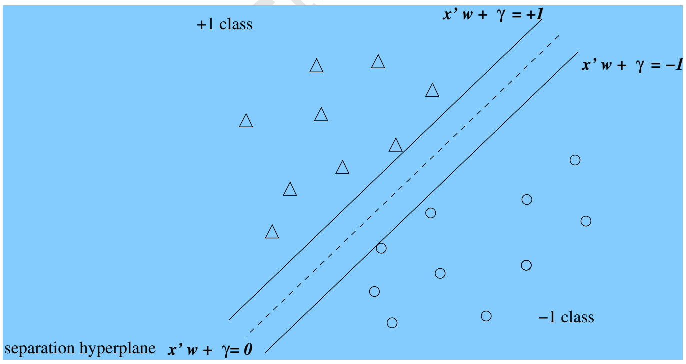


Modelling SVMs: linearly separable data

- We want to classify m points $x_i \in \mathbb{R}^n,\ i = 1, \ldots, m$.
- Every point x belongs to one of two classes, linearly separated by the hyperplane $w^\top x + \gamma = 0$, such that

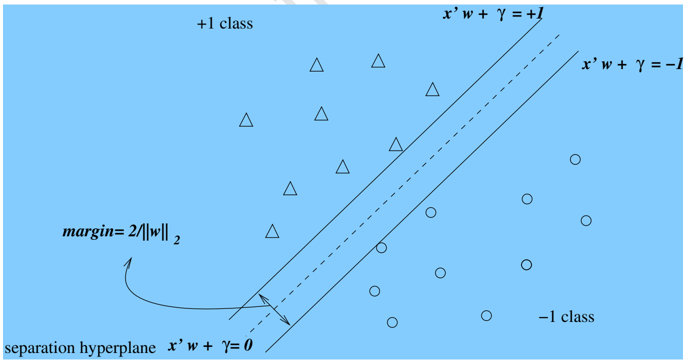
$$\begin{array}{ll} \text{if } x \text{ is of class } \triangle \text{ then} & w^\top x + \gamma \geq +\delta \\ \text{if } x \text{ is of class } \bigcirc \text{ then} & w^\top x + \gamma \leq -\delta \end{array}$$

- We can normalize assuming $\delta = 1$ (always possible dividing w and γ by $\delta > 0$). Then every point belongs to the class +1 o -1.



Separation margin

- Given the m pairs $(x_i, y_i) \in \mathbb{R}^n \times \{+1, -1\},\ i = 1, \ldots, m$ we look for the hyperplane defined by (w, γ) with the maximum distance between parallel planes $w^\top x + \gamma \geq +1$ and $w^\top x + \gamma \leq -1$.
- w is the normal to the separation hyperplane, γ determines its location with respect to the origin.
- We'll show next that the margin between planes is $\frac{2}{||w||_2}$.



SVM formulations | SV Classifiers for linearly separable data

The margin between planes is $\frac{2}{||w||_2}$

- Given two parallel hyperplanes $x^\top w = a$ and $x^\top w = b$, and x_1 point of the first plane ($x_1^\top w = a$), the closest point to x_1 in the second hyperplane, named x_2, ($x_2^\top w = b$) can be written as $x_2 = x_1 + \alpha w$.
- Then:

$$\begin{aligned} x_2^\top &= x_1^\top + \alpha w^\top \\ x_2^\top w &= x_1^\top w + \alpha w^\top w \\ b &= a + \alpha ||w||_2^2 \\ \alpha &= \frac{b-a}{||w||_2^2} \end{aligned}$$

- The margin between hyperplanes is the 2-norm of αw:

$$||\alpha w||_2 = |\alpha| \cdot ||w||_2 = \frac{|b-a|}{||w||_2^2} ||w||_2 = \frac{|b-a|}{||w||_2}$$

- For the SVM, $a = -\gamma - 1$ and $b = -\gamma + 1$, thus the margin is $\frac{2}{||w||_2}$.

jordi.castro@upc.edu

SVM is a constrained quadratic optimization problem

- The SVM is an optimization problem in variables w, γ:

$$\max_{(w,\gamma)\in\mathbb{R}^{n+1}} \quad \frac{2}{||w||_2}$$
$$\text{s. to} \quad y_i(w^\top x_i + \gamma) \geq 1 \quad i = 1, \dots, m$$

- Constraints impose:

$$w^\top x_i + \gamma \geq +1, \text{ for } y_i = +1$$
$$w^\top x_i + \gamma \leq -1, \text{ for } y_i = -1$$

- The objective is equivalent to $\min \frac{1}{2}||w||_2$, which is equivalent to $\min \frac{1}{2}||w||_2^2 \equiv \min \frac{1}{2}w^\top w$.
- Denoting by $A \in \mathbb{R}^{m\times n}$ the matrix storing rowwise the vectors x_i, by $Y = diag(y_1, \dots, y_m)$, and by e a vector of m 1's, the problem is formulated in compact matrix form as:

$$\min_{(w,\gamma)\in\mathbb{R}^{n+1}} \quad \frac{1}{2}w^\top w$$
$$\text{s. to} \quad Y(Aw + \gamma e) \geq e.$$

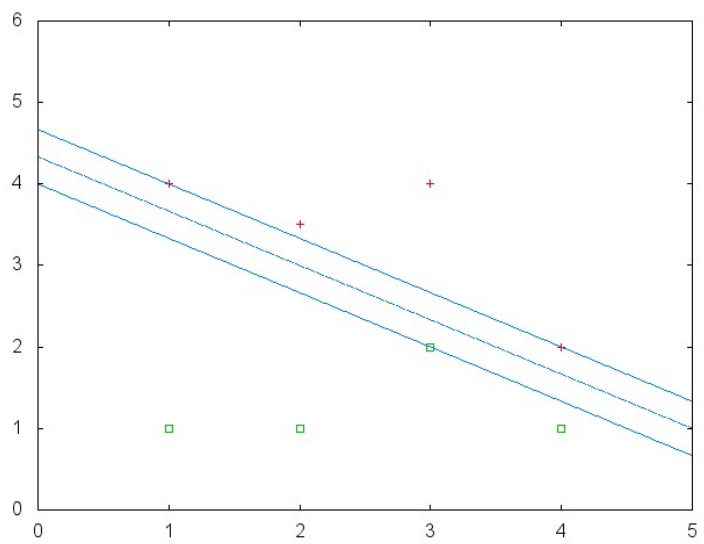
Example

For this AMPL data...

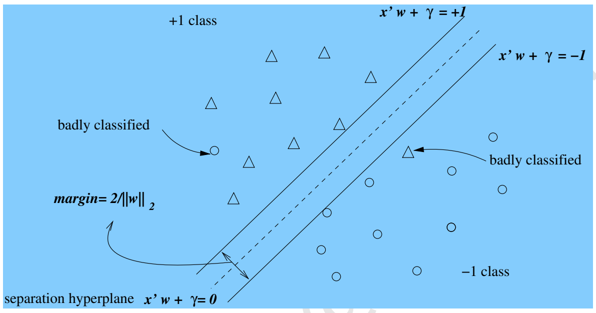
```
# Initial values for n, m
param m:=8;
param n:=2;
# y and A
param y :=
  1     1
  2     1
  3     1
  4     1
  5     -1
  6     -1
  7     -1
  8     -1;

param A :  1  2   :=
  1        3  4
  2        4  2
  3        2  3.5
  4        1  4
  5        1  1
  6        3  2
  7        2  1
  8        4  1 ;
```

...we get this separation plane by solving the optimization problem



Modelling SVMs: linearly inseparable data, soft margin



- We consider artificial variables $s_i \geq 0,\ i = 1, \dots, m$, named slacks, one for each point, to account for errors in the classification. The resulting constraints are named soft constraints:

$$y_i(w^\top x_i + \gamma) \geq 1 - s_i \quad i = 1, \dots, m$$

- Constraints impose:

$$\begin{aligned} w^\top x_i + \gamma + s_i &\geq +1, \text{ for } y_i = +1 \\ w^\top x_i + \gamma - s_i &\leq -1, \text{ for } y_i = -1 \end{aligned}$$

The constrained quadratic optimization problem

- The objective is to maximize the margin, and at the same time to minimize the classification errors $\sum_{i=1}^{m} s_i = e^\top s$. These two opposite objectives are weighted by parameter $\nu \in \mathbb{R}$.
- The SVM is an optimization problem in variables w, γ, s:

$$\begin{array}{lll} \min\limits_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} & \frac{1}{2} w^\top w + \nu \sum_{i=1}^{m} s_i & \\ \text{s. to} & y_i(w^\top x_i + \gamma) + s_i \geq 1 & i = 1, \dots, m \\ & s_i \geq 0 & i = 1, \dots, m \end{array}$$

or equivalently in matrix form

$$\begin{array}{ll} \min\limits_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} & \frac{1}{2} w^\top w + \nu e^\top s \\ \text{s. to} & Y(Aw + \gamma e) + s \geq e \\ & s \geq 0 \end{array}$$

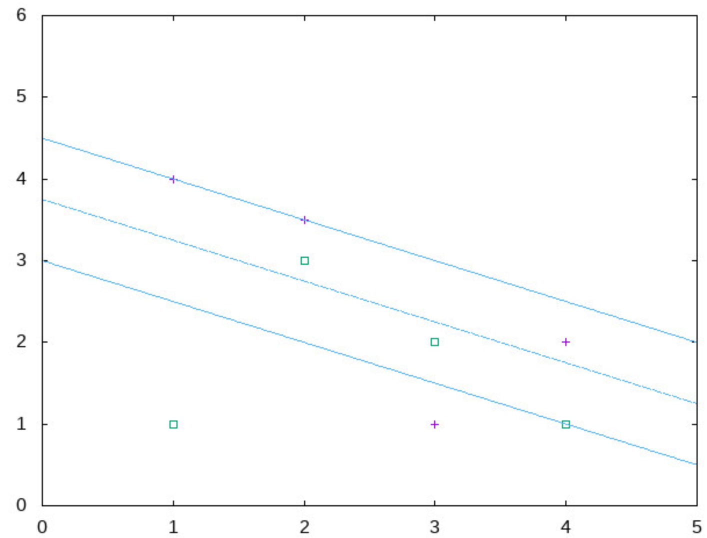
Example

For this AMPL data...

```
# Initial values for n, m and nu
param nu:=30;
param m:=8;
param n:=2;
# Generated y and A
param y :=
  1      1
  2      1
  3      1
  4      1
  5      -1
  6      -1
  7      -1
  8      -1;

param A :  1  2  :=
  1        3  1
  2        4  2
  3        2  3.5
  4        1  4
  5        1  1
  6        3  2
  7        2  3
  8        4  1;
```

... we get this separation plane by solving the optimization problem



Input and feature spaces

- The space of the training set $x \in X \subseteq \mathbb{R}^n$ is named input space.
- If data are linearly inseparable, we can consider a mapping function

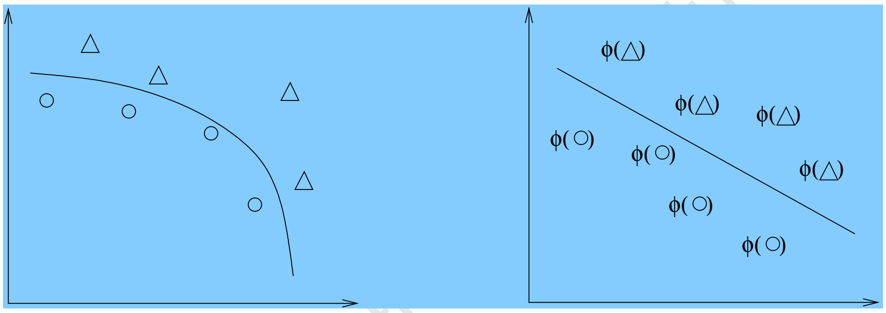
$$\phi : X \subseteq \mathbb{R}^n \rightarrow F \subseteq \mathbb{R}^N$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \phi(x) = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_N(x) \end{pmatrix}$$

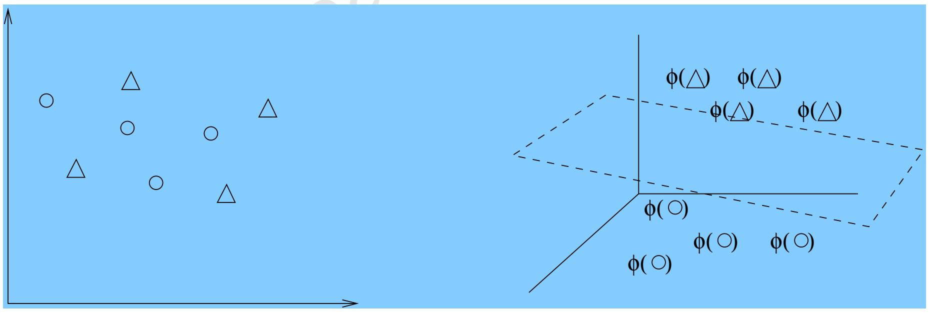
- F is named feature space.
- The dimension of the input space is n. The dimension of the feature space is N. N can be different from n.
- We expect that using $\phi(x)$ instead of x the SVM will perform better: the hyperplane will separate $\phi(x)$ better than x.

Input and feature spaces: examples

Case $n = N = 2$



Case $3 = N > n = 2$



SVM formulation in feature space

- Same formulation than in input space, replacing x by $\phi(x)$.
- The mapping only affects to input data, not to optimization model.
- Linearly separable case:

$$\begin{array}{ll} \min\limits_{(w,\gamma)\in\mathbb{R}^{N+1}} & \frac{1}{2}w^\top w \\ \text{s. to} & y_i(w^\top \phi(x_i) + \gamma) \geq 1 \quad i = 1,\dots,m \end{array}$$

- Soft margin case:

$$\begin{array}{lll} \min\limits_{(w,\gamma,s)\in\mathbb{R}^{N+1+m}} & \frac{1}{2}w^\top w + \nu \sum_{i=1}^{m} s_i & \\ \text{s. to} & y_i(w^\top \phi(x_i) + \gamma) + s_i \geq 1 & i = 1,\dots,m \\ & s_i \geq 0 & i = 1,\dots,m \end{array}$$

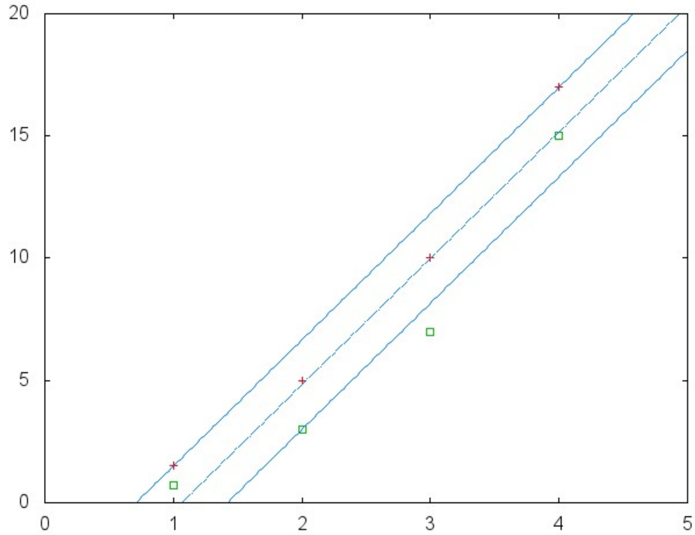
Example in input space

For this AMPL data...

```
# Initial values for n, m and nu
param nu:=30;
param m:=8;
param n:=2;
# Generated y and A
param y :=
  1     1
  2     1
  3     1
  4     1
  5    -1
  6    -1
  7    -1
  8    -1;

param A :  1   2   :=
  1        3  10
  2        4  17
  3        2  5
  4        1  1.5
  5        1  0.7
  6        3  7
  7        2  3
  8        4  15 ;
```

...we get this separation plane by solving the optimization problem

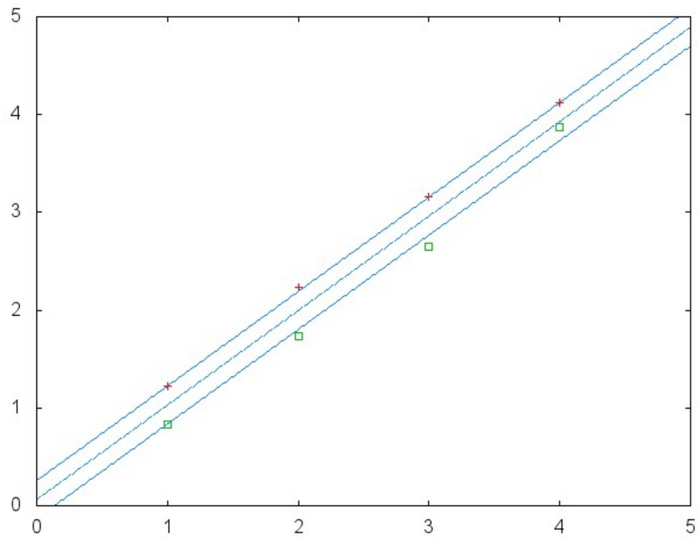


Example in feature space

For the previous data using the mapping

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix} = \begin{pmatrix} x_1 \\ \sqrt{x_2} \end{pmatrix}$$

we get this separation plane by solving the optimization problem



Definition of kernel

- Consider for simplicity the input space X is finite with m points: $X = \{x_1, \ldots, x_m\}$, $x_i \in \mathbb{R}^n, i = 1, \ldots, m$.
- The representation of x_i in the feature space is $\phi(x_i)$.
- If we formulate the dual problem of the primal SVM problem (to be seen later in the course) we will obtain inner products:

$$K_{ij} = K(x_i, x_j) = \phi(x_i)^\top \phi(x_j) = \sum_{l=1}^{N} \phi_l(x_i)\phi_l(x_j) \quad i, j = 1, \ldots, m$$

Definition

A kernel is a function $K : X \times X \to \mathbb{R}$ such that for all $x, y \in X$

$$K(x, y) = \phi(x)^\top \phi(y)$$

where ϕ is a mapping from the input space X to the (inner product) space F.

Some properties of kernels

- Kernels are symmetric functions:

$$K(x, y) = \phi(x)^\top \phi(y) = \phi(y)^\top \phi(x) = K(y, x)$$

- If X is finite with m points then we then get a symmetric matrix

$$K = (K(x_i, x_j))_{i,j=1,\ldots,m} = (K_{ij})_{i,j=1,\ldots,m} = K^\top$$

- The kernel can be seen a function that measures similarities between pairs of inputs in the feature space.
- If we have matrix K we even don't need to know the mapping $\phi(x)$ (using the dual formulation of the SVM), even if $N = \infty$.
- The only requirement is that matrix K has to be positive semidefinite ($K \succeq 0$).

$K \succeq 0$ is sufficient and necessary to be a kernel

Proposition

Let X be a finite input space $X = \{x_1, \ldots, x_m\}, x_i \in \mathbb{R}^n$, and $K(x, x')$ a symmetric function ($K(x, x') = K(x', x)$) on $X \times X$. Then $K(x, x')$ is a kernel function if and only if the matrix

$$K = (K(x_i, x_j))_{i,j=1,\ldots,m}$$

is positive semidefinite.

RBF or Gaussian kernel

- There are several available kernel functions, for instance the radial basis or Gaussian kernel:

$$K(x, y) = e^{-\frac{||x-y||^2}{\sigma^2}}$$

Purpose of SV regression

- Given m points $(x_i, y_i), i = 1, \ldots, m$, where $x_i \in \mathbb{R}^n, y \in \mathbb{R}$
- Find affine model $y = w^\top x + \gamma$, $w \in \mathbb{R}^n, \gamma \in \mathbb{R}$, such that small errors $(< \varepsilon)$ are neglected.

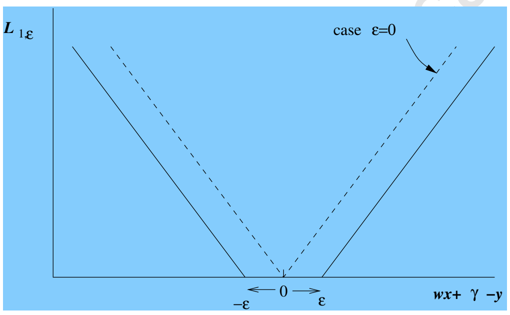
We consider two ε-insensitive loss functions:

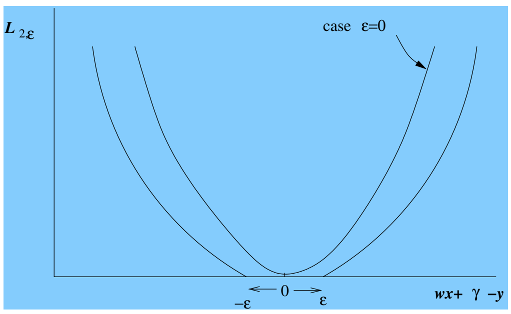
- ε-insensitive linear function

$$\begin{aligned} L_{1,\varepsilon} &= \max(0, |w^\top x + \gamma - y| - \varepsilon) \\ &= \begin{cases} 0 & \text{if } -\varepsilon \le w^\top x + \gamma - y \le \varepsilon \\ |w^\top x + \gamma - y| - \varepsilon & \text{otherwise} \end{cases} \end{aligned}$$

- ε-insensitive quadratic function

$$\begin{aligned} L_{2,\varepsilon} &= \max(0, (|w^\top x + \gamma - y| - \varepsilon)^2) \\ &= \begin{cases} 0 & \text{if } -\varepsilon \le w^\top x + \gamma - y \le \varepsilon \\ (|w^\top x + \gamma - y| - \varepsilon)^2 & \text{otherwise} \end{cases} \end{aligned}$$





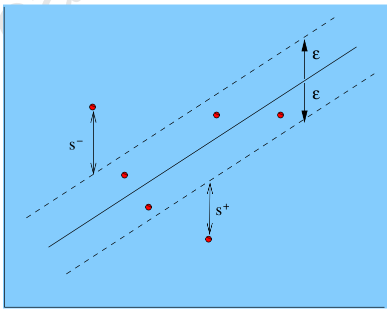
The SV regression models I

- In ideal situation all points (x_i, y_i) should be within distance ε of plane $w^\top x + \gamma = y$, that is:

$$-\varepsilon \leq (w^\top x_i + \gamma) - y_i \leq \varepsilon \qquad i = 1, \ldots, m$$

- Since this is not possible, we need slacks $s_i^+ \geq 0$ and $s_i^- \geq 0$ such that:

$$-s_i^- - \varepsilon \leq (w^\top x_i + \gamma) - y_i \leq \varepsilon + s_i^+ \qquad i = 1, \ldots, m$$



The SV regression models II

- The L_1 SV regression model:

$$\begin{array}{lll} \min\limits_{w,\gamma,s^+,s^-} & \frac{1}{2}\|w\|_2^2 + \nu \sum_{i=1}^{m} (s_i^+ + s_i^-) & \\ \text{s. to} & -s_i^- - \varepsilon \leq (w^\top x_i + \gamma) - y_i \leq \varepsilon + s_i^+ & i = 1, \ldots, m \\ & s_i^+ \geq 0, s_i^- \geq 0 & i = 1, \ldots, m \end{array}$$

- The L_2 SV regression model:

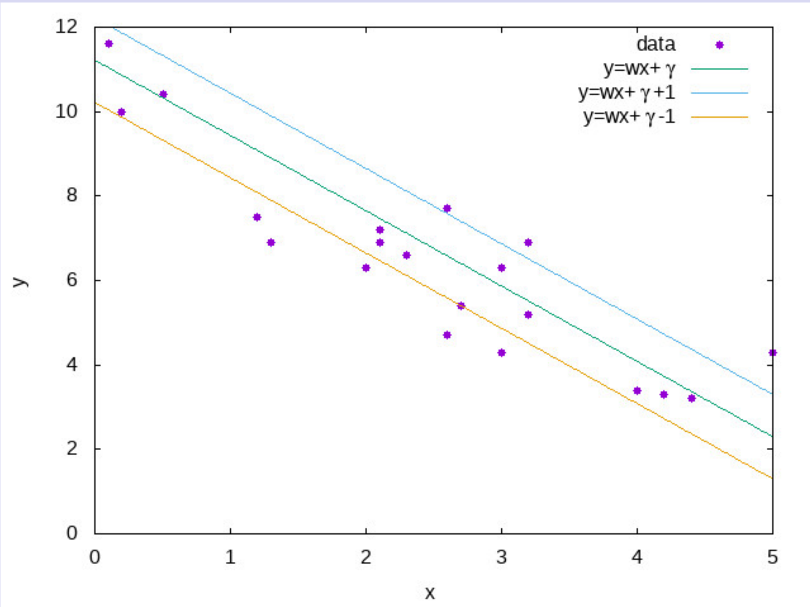
$$\begin{array}{lll} \min\limits_{w,\gamma,s^+,s^-} & \frac{1}{2}\|w\|_2^2 + \nu \sum_{i=1}^{m} ((s_i^+)^2 + (s_i^-))^2 & \\ \text{s. to} & -s_i^- - \varepsilon \leq (w^\top x_i + \gamma) - y_i \leq \varepsilon + s_i^+ & i = 1, \ldots, m \\ & s_i^+ \geq 0, s_i^- \geq 0 & i = 1, \ldots, m \end{array}$$

The SV regression models III

- Why term $\|w\|_2^2$ is minimized (equivalent to maximize margin between planes)? Just a short explanation:
  - According to theory, the prediction error of $w^\top x + \gamma = y$ decreases with $\|w\|_2$: the smaller $\|w\|_2$, the smaller the prediction error.
  - ε is vertical distance between planes; $\|w\|$ is associated to margin. For fixed ε we find planes with widest margin. Wider margins make more general the SVM and reduce overfitting.
  - The term $\|w\|_2$ convexifies the problem: it guarantees a unique solution for L_1.

The SV regression models IV

Example of L_1 SV regression ($\varepsilon = 1$, $\nu = 1$)



SVMs are convex optimization problems

Definition

The optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s. to} & x \in \Omega \end{array}$$

is convex if f is convex function and Ω is convex set.

SVM most general formulation: soft margin in feature space

$$\begin{array}{lll} \min\limits_{(w,\gamma,s)\in\mathbb{R}^{N+1+m}} & \frac{1}{2}w^\top w + \nu \sum\limits_{i=1}^{m} s_i & \\ \text{s. to} & y_i(w^\top \phi(x_i) + \gamma) + s_i \geq 1 & i = 1, \ldots, m \\ & s_i \geq 0 & i = 1, \ldots, m \end{array}$$

- The objective functions is convex quadratic.
- The feasible set is a convex polyhedron

Summary of necessary optimality conditions

- Let

$$\begin{array}{lll} \min & f(x) & \\ \text{s. to} & h(x) = 0 & [h_i(x) = 0 \quad i = 1, \ldots, m] \\ & g(x) \leq 0 & [g_j(x) \leq 0 \quad j = 1, \ldots, p] \end{array}$$

  and its Lagrangian

$$L(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x)$$

- **Necessary conditions** If x^* regular point and a local minimizer then there exist $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that:

  First-order conditions (KKT)

  (i) $h(x^*) = 0$, $g(x^*) \leq 0$

  (ii) $\nabla_x L(x^*, \lambda^*, \mu^*) = \nabla f(x^*) + \nabla h(x^*)\lambda^* + \nabla g(x^*)\mu^* = 0$

  (iii) $\mu^* \geq 0$ and $\mu^{*\top} g(x^*) = 0$ (if $g_j(x^*)$ is inactive then $\mu_j^* = 0$)

  Second-order conditions

  (iv) $d^\top \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d \geq 0$, for all $d \in M = \{d : \nabla h(x^*)^\top d = 0,\ \nabla g_j(x^*)^\top d = 0\ j \in \mathcal{A}(x^*)\}$.

Summary of sufficient optimality conditions

- **Sufficient optimality conditions** The point x^* is local minimizer if:

  First-order conditions (KKT)

  (i) $h(x^*) = 0$, $g(x^*) \le 0$

  (ii) $\nabla_x L(x^*, \lambda^*, \mu^*) = \nabla f(x^*) + \nabla h(x^*)\lambda^* + \nabla g(x^*)\mu^* = 0$

  (iii) $\mu^* \ge 0$ and ${\mu^*}^\top g(x^*) = 0$ (if $g_j(x^*)$ is inactive then $\mu_j^* = 0$)

  Second-order conditions

  (iv) $d^\top \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d > 0$, for all $d \in M' = \{d : \nabla h(x^*)^\top d = 0,\ \nabla g_j(x^*)^\top d = 0\ j \in \mathcal{A}(x^*) \cap \{j : \mu_j^* > 0\}\}$.

- Necessary and sufficient conditions differ in:
  - No need x^* is regular;
  - condition (iv):

$$
\begin{array}{lll}
d^\top \nabla^2_{xx} L(x^*, \lambda, \mu) d > 0 & d \in M' & \text{[sufficient]} \\
d^\top \nabla^2_{xx} L(x^*, \lambda, \mu) d \ge 0 & d \in M & \text{[necessary]}
\end{array}
$$

Optimality conditions of convex problems

KKT conditions are sufficient and necessary also for convex problems, no need to check second order conditions

Theorem

Given

$$
\begin{array}{lll}
\min & f(x) & \\
\text{s.to} & h(x) = 0 & [h_i(x) = 0 \quad i = 1, \ldots, m] \\
 & g(x) \le 0 & [g_j(x) \le 0 \quad j = 1, \ldots, p]
\end{array}
$$

$f, h, g \in \mathcal{C}^1$, f and g_j convex, and $h(x) = Ax - b$ affine function. If first-order KKT conditions are satisfied at x^*, then x^* is a global optimum.

The SVM formulation considered

SVM general formulation: soft margin in feature space

$$\begin{array}{rll} \min\limits_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} & \frac{1}{2}w^\top w + \nu\sum_{i=1}^m s_i & \\ \text{s. to} & y_i(w^\top\phi(x_i)+\gamma)+s_i \geq 1 & i=1,\ldots,m \\ & s_i \geq 0 & i=1,\ldots,m \end{array}$$

Defining $A = \begin{bmatrix} \phi(x_1)^\top \\ \vdots \\ \phi(x_m)^\top \end{bmatrix} \in \mathbb{R}^{m\times n}$, $Y = \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_m \end{bmatrix}$, and $e = \begin{bmatrix} 1_1 \\ \vdots \\ 1_m \end{bmatrix}$:

SVM general formulation in matrix form

$$\begin{array}{rl} \min\limits_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} & \frac{1}{2}w^\top w + \nu e^\top s \\ \text{s. to} & Y(Aw+\gamma e)+s \geq e \\ & s \geq 0 \end{array}$$

Lagrange multipliers and Lagrangian function of SVM

- Standard form $\min f(x)$ s. to $g(x)\leq 0$ and Lagrange multipliers:

$$\begin{array}{rll} \min\limits_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} & \frac{1}{2}w^\top w + \nu e^\top s & \\ \text{s. to} & -Y(Aw+\gamma e)-s+e \leq 0 & [\lambda\in\mathbb{R}^m] \\ & -s \leq 0 & [\mu\in\mathbb{R}^m] \end{array}$$

- Lagrangian function

$$\begin{aligned} L(w,\gamma,s,\lambda,\mu) &= \frac{1}{2}w^\top w + \nu e^\top s + \lambda^\top(-Y(Aw+\gamma e)-s+e)+\mu^\top(-s) \\ &= \frac{1}{2}\sum_{i=1}^n w_i^2 + \nu\sum_{i=1}^m s_i + \sum_{i=1}^m \lambda_i(-y_i(\phi(x_i)^\top w+\gamma)-s_i+1) - \sum_{i=1}^m \mu_i s_i \end{aligned}$$

KKT optimality conditions of SVM

KKT conditions are necessary and sufficient (SVM is convex problem)

Primal feasibility

$$(i) \quad Y(Aw + \gamma e) + s \geq e, \quad s \geq 0$$

$\nabla L() = 0$

$$(ii.w) \quad \nabla_w L() = w - (\lambda^\top YA)^\top = w - \sum_{i=1}^{m} \lambda_i y_i \phi(x_i) = 0 \quad (n \text{ equations})$$

$$(ii.\gamma) \quad \nabla_\gamma L() = -\lambda^\top Ye = -\sum_{i=1}^{m} \lambda_i y_i = 0 \quad (1 \text{ equation})$$

$$(ii.s) \quad \nabla_s L() = \nu e - \lambda - \mu = 0 \quad (m \text{ equations})$$

Complementarity

$$(iii.\lambda) \quad \lambda_i \geq 0, \quad \lambda_i(y_i(\phi(x_i)^\top w + \gamma) + s_i - 1) = 0 \quad i = 1, \ldots, m$$

$$(iii.\mu) \quad \mu_i \geq 0, \quad \mu_i s_i = 0 \quad i = 1, \ldots, m$$

Dual problem

- Primal problem

$$\begin{array}{lll} \min_x & f(x) & \\ \text{s.to} & h(x) = 0 & [h_i(x) = 0 \quad i = 1, \ldots, m] \\ & g(x) \leq 0 & [g_j(x) \leq 0 \quad j = 1, \ldots, p] \\ & x \in X & \end{array}$$

- Lagrangian function:

$$L(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x)$$

- Dual function $q(\lambda, \mu)$ is

$$q(\lambda, \mu) = \min_x \; L(x, \lambda, \mu) \quad x \in X$$

Constraints $h(x) = 0$ and $g(x) \leq 0$ dualized, preserving $x \in X$. Depending what is dualized, different formulations obtained.

- Dual problem

$$\max_{\lambda, \mu} \; q(\lambda, \mu) \quad \mu \geq 0$$

NOTE: although inf and sup preferred we will use min and max $q()$.

Dual problem: example

$$\begin{array}{lll} \min & x_1^2 + x_2^2 & \\ \text{s.to} & x_1 + x_2 \geq 4 & \equiv \quad 4 - x_1 - x_2 \leq 0 \\ & x_1 \geq 0, x_2 \geq 0 & \equiv \quad -x_1 \leq 0, -x_2 \leq 0 \end{array}$$

Solution is $x_1^* = x_2^* = 2$, $f(x^*) = 8$.
Dual function is:

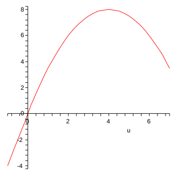
$$q(\mu) = \min_{x_1 \geq 0, x_2 \geq 0} x_1^2 + x_2^2 + \mu(4 - x_1 - x_2) = \min_{x_1 \geq 0, x_2 \geq 0} (x_1^2 - \mu x_1) + (x_2^2 - \mu x_2) + 4\mu$$

Problem is separable, with solution:

$$\begin{cases} x_1 = x_2 = 0 & \text{if } \mu < 0 \quad \text{since } x_i^2 - \mu x_i \geq 0 \\ x_1 = x_2 = \mu/2 & \text{if } \mu \geq 0 \quad \text{since solution of } \min x_i^2 - \mu x_i \end{cases}$$

Then $q(\mu)$ is the concave function:

$$q(\mu) = \begin{cases} 4\mu & \mu < 0 \\ -\mu^2/2 + 4\mu & \mu \geq 0 \end{cases}$$



Solution of dual problem $\max_{\mu \geq 0} q(\mu)$ is $\mu^* = 4$ and $q(\mu^*) = f(x^*) = 8$.

Some duality theorems

Theorem (Concavity of $q(\lambda, \mu)$)

The dual function

$$q(\lambda, \mu) = \min_{x \in X} L(x, \lambda, \mu) = \min_{x \in X} f(x) + \lambda^\top h(x) + \mu^\top g(x)$$

is concave (in the region where it is finite, that is, the minimum exists).

Theorem (Weak duality)

Let x be a feasible point of primal problem (i.e. $h(x) = 0$, $g(x) \leq 0$, $x \in X$) and (λ, μ) a feasible point of dual problem (i.e., $\mu \geq 0$), then

$$q(\lambda, \mu) \leq f(x)$$

Theorem (Strong duality)

If X is a convex set, $f(x)$ and $g(x)$ are convex functions, $h(x) = Ax - b$ (affine function), under certain constraints qualifications (Slater condition) then:

$$q(\lambda^*, \mu^*) = f(x^*)$$

Wolfe duality

- Lagrangian duality does not require differentiability. Wolfe duality assumes differentiability.
- If $f(x)$, $h(x)$ and $g(x)$ are convex and differentiable functions, a necessary and sufficient condition of optimality of the dual function

$$q(\lambda, \mu) = \min_x L(x, \lambda, \mu)$$

  is

$$\nabla_x L(x, \lambda, \mu) = 0$$

- The dual problem

$$\begin{array}{ll} \max_{\lambda,\mu} & q(\lambda, \mu) \\ & \mu \geq 0 \end{array}$$

  can thus be recast as

$$\begin{array}{ll} \max_{x,\lambda,\mu} & L(x, \lambda, \mu) \\ & \nabla_x L(x, \lambda, \mu) = 0 \\ & \mu \geq 0 \end{array}$$

- This allows a simpler formulation of some problems: LP, QP

SVM formulation in standard form

Defining $A = \begin{bmatrix} \phi(x_1)^\top \\ \vdots \\ \phi(x_m)^\top \end{bmatrix} \in \mathbb{R}^{m\times n}$, $Y = \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_m \end{bmatrix}$, and $e = \begin{bmatrix} 1_1 \\ \vdots \\ 1_m \end{bmatrix}$:

SVM in matrix standard form with Lagrange multipliers

$$\begin{array}{lll} \min_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} & \frac{1}{2} w^\top w + \nu e^\top s & \\ \text{s. to} & -Y(Aw + \gamma e) - s + e \leq 0 & [\lambda \in \mathbb{R}^m] \\ & -s \leq 0 & [\mu \in \mathbb{R}^m] \end{array}$$

Lagrangian function

$$\begin{aligned} L(w, \gamma, s, \lambda, \mu) &= \frac{1}{2} w^\top w + \nu e^\top s + \lambda^\top(-Y(Aw + \gamma e) - s + e) - \mu^\top s \\ &= \frac{1}{2}\sum_{i=1}^n w_i^2 + \nu \sum_{i=1}^m s_i + \sum_{i=1}^m \lambda_i(-y_i(\phi(x_i)^\top w + \gamma) - s_i + 1) - \sum_{i=1}^m \mu_i s_i \end{aligned}$$

Dual problem formulation (I)

SVM dual problem

$$\begin{array}{lll} \max_{w,\gamma,s,\lambda,\mu} & L(w,\gamma,s,\lambda,\mu) & \\ & \nabla_w L(w,\gamma,s,\lambda,\mu) = 0 & n \text{ constraints} \\ & \nabla_{\gamma} L(w,\gamma,s,\lambda,\mu) = 0 & 1 \text{ constraint} \\ & \nabla_s L(w,\gamma,s,\lambda,\mu) = 0 & m \text{ constraints} \\ & \lambda \geq 0, \quad \mu \geq 0 & \end{array}$$

Detailed SVM dual problem

$$\begin{array}{ll} \max_{w,\gamma,s,\lambda,\mu} & \frac{1}{2} w^\top w + \nu e^\top s + \lambda^\top(-Y(Aw+\gamma e) - s + e) - \mu^\top s \\ & w - (\lambda^\top YA)^\top = 0 \\ & \lambda^\top Ye = 0 \\ & \nu e - \lambda - \mu = 0 \\ & \lambda \geq 0, \quad \mu \geq 0 \end{array}$$

Dual problem formulation (II)

Replacing $w = (\lambda^\top YA)^\top$ in the objective we get (details in the blackboard)

$$\begin{array}{ll} \max_{\lambda,\mu} & \lambda^\top e - \frac{1}{2}\lambda^\top YAA^\top Y\lambda \\ & \lambda^\top Ye = 0 \\ & \nu e - \lambda - \mu = 0 \\ & \lambda \geq 0, \quad \mu \geq 0 \end{array}$$

Using $\mu \geq 0$ and $\mu = \nu e - \lambda$ we finally have the QP (convex because matrix $AA^\top = K \succeq 0$):

Dual of SVM (matrix form)

$$\begin{array}{ll} \max_{\lambda} & \lambda^\top e - \frac{1}{2}\lambda^\top YAA^\top Y\lambda \\ & \lambda^\top Ye = 0 \\ & 0 \leq \lambda \leq \nu \end{array}$$

Dual of SVM (scalar form)

$$\begin{array}{ll} \max_{\lambda} & \sum_{i=1}^m \lambda_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \lambda_i y_i \lambda_j y_j K_{ij} \\ & \sum_{i=1}^m \lambda_i y_i = 0 \\ & 0 \leq \lambda_i \leq \nu \quad i = 1, \ldots, m \end{array}$$

Retrieving normal vector w from the dual solution λ

- From $w = A^\top Y\lambda$ we have

$$w = \sum_{i=1}^{m} \lambda_i y_i \phi(x_i)$$

- Considering the partition (MC, SV, NB) of points $\{1, \ldots, m\}$:
  - $MC \subseteq \{1, \ldots, m\}$: set of misclassified points.
  - $SV \subseteq \{1, \ldots, m\}$: set of support vectors (on planes $w^\top x + \gamma = \pm 1$).
  - $NB = \{1, \ldots, m\} \setminus (MC \cup SV)$: non-binding points.

  it can be shown that

$$\begin{cases} s_i > 0,\ \lambda_i = \nu & \text{if } i \in MC \\ s_i = 0,\ \lambda_i \geq 0 & \text{if } i \in SV \\ s_i = 0,\ \lambda_i = 0 & \text{if } i \in NB \end{cases}$$

  and then

$$w = \sum_{i=1}^{m} \lambda_i y_i \phi(x_i) = \sum_{i \in MC} \nu y_i \phi(x_i) + \sum_{i \in SV} \lambda_i y_i \phi(x_i)$$

(proof in the blackboard)

Retrieving "intercept" γ from the dual solution λ

We have to compute γ of $w^\top x + \gamma = \pm 1$. This is the procedure:

- Choose some point x_i such that $s_i = 0$ and $\lambda_i > 0$ (that is, $i \in SV$).
- Since x_i is a support vector we know that

$$y_i(w^\top \phi(x_i) + \gamma) - 1 = 0 \qquad y_i = \pm 1$$

- Then we compute γ as

$$\gamma = \frac{1}{y_i} - w^\top \phi(x_i).$$

Best SVM packages in machine learning community

LIBSVM for linear/nonlinear kernels

- Solves the dual SVM formulation, with several kernels (e.g. Gaussian)
- Uses the SMO algorithm, specific for the dual SVM problem.

LIBLINEAR for linear kernels

- Transforms the problem to a "similar" unconstrained one without γ.
- It either solves the primal

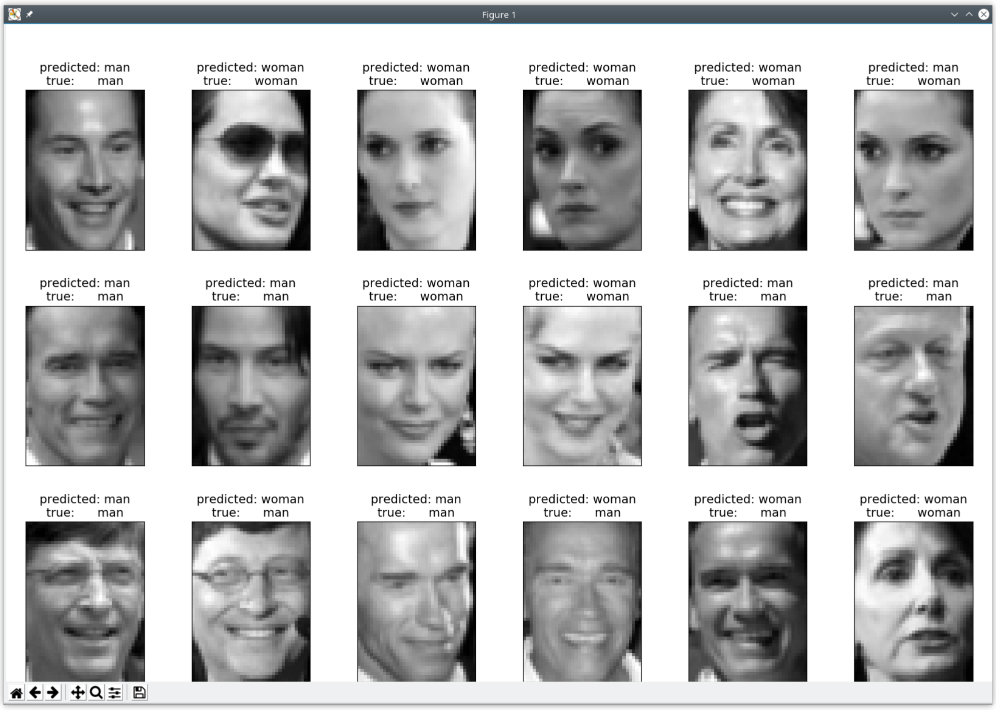
$$\min_{w} \frac{1}{2} w^\top w + \nu \sum_{i=1}^{m} max(0, 1 - y_i w^\top x_i)^2$$

or the dual

$$\begin{array}{ll} \max_{\lambda} & \lambda^\top e - \frac{1}{2}\lambda^\top YAA^\top Y\lambda \\ & 0 \le \lambda \le \nu \end{array}$$

using a trust-region CG Newton method or a coordinate descent algorithm.
- Meaning of optimality tolerances different (looser) than with other approaches.

Example libsvm/liblinear: gender recognition by face



Bibliography

- M.S. Bazaraa, H.D. Sherali, C.M. Shetty, *Nonlinear Programming. Theory and Algorithms, 3rd Ed.*, Wiley, 2006.
- D.P. Bertsekas, *Nonlinear Programming, 2nd Ed.*, 1999, Athena Scientific, Belmont, USA.
- N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*, 2000, Cambridge University Press, Cambridge, UK.
- D.G. Luenberger, Y. Ye., *Linear and Nonlinear Programming, 4th Ed.*, 2016, Springer, New York, USA.
- J. Nocedal, S.J. Wright, *Numerical Optimization, 2nd Ed.*, 2006, Springer, New York, USA.