SVM and Kernel machine

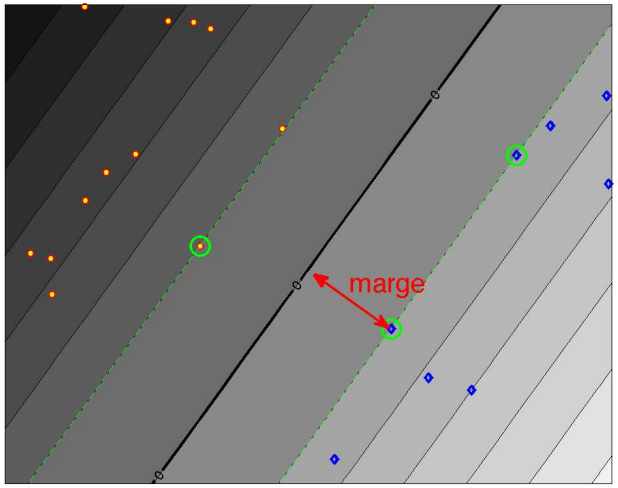
Stéphane Canu
stephane.canu@litislab.eu

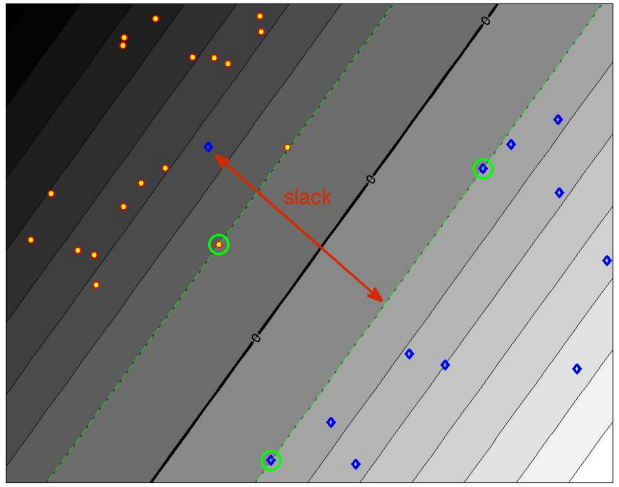
Escuela de Ciencias Informáticas 2011

July 31, 2011

Road map

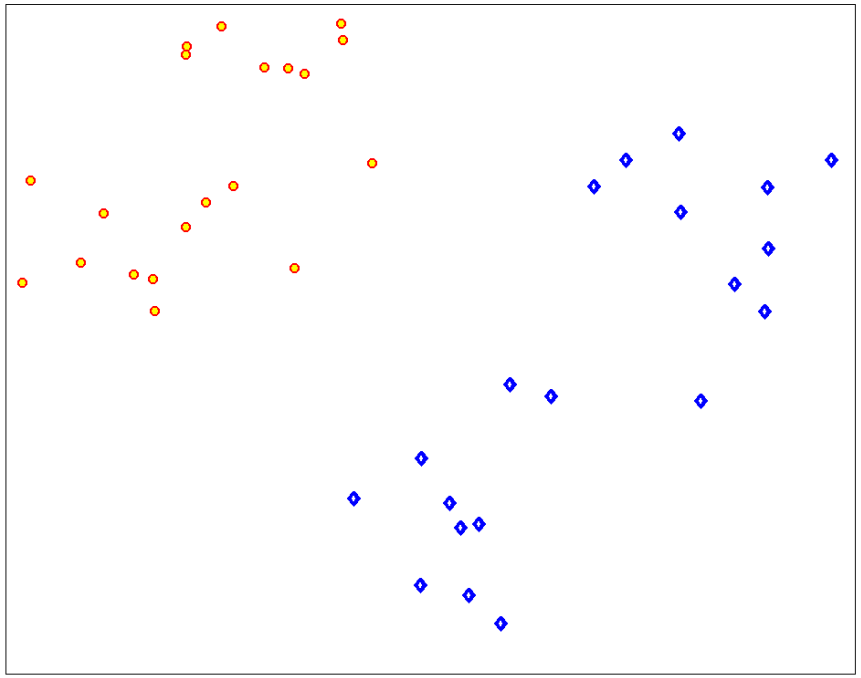
1. Linear SVM
   - Linear classification
   - The margin
   - Linear SVM: the problem
   - Optimization in 5 slides
   - Dual formulation of the linear SVM
   - Solving the dual
   - The non separable case





Linear classification

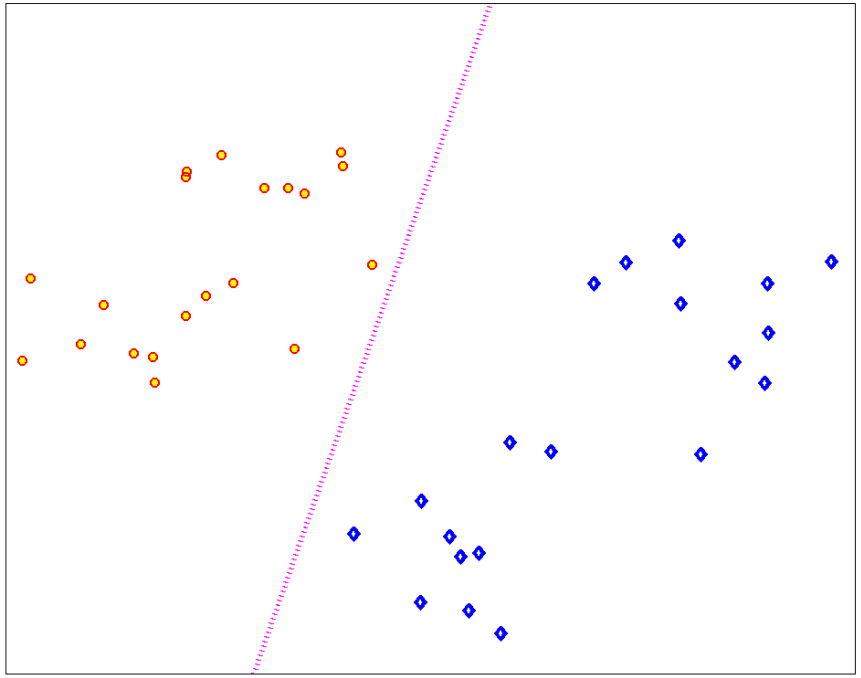
Find a line to separate (classify) blue from red



$$D(x) = \text{sign}\left(\mathbf{v}^\top \mathbf{x} + a\right)$$

Linear classification

Find a line to separate (classify) blue from red



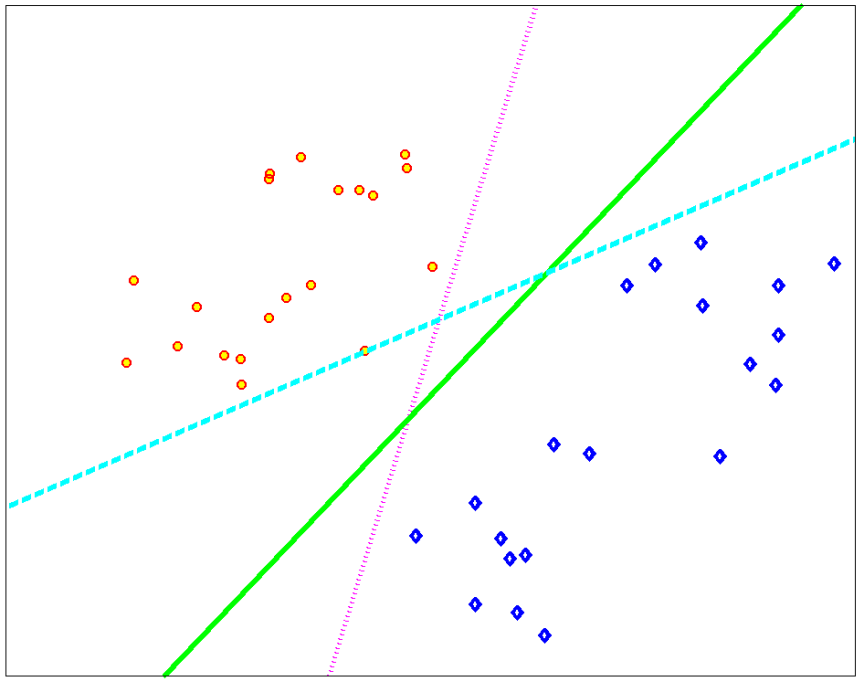
$$D(x) = \text{sign}\left(\mathbf{v}^\top \mathbf{x} + a\right)$$

the decision border:

$$\mathbf{v}^\top \mathbf{x} + a = 0$$

Linear classification

Find a line to separate (classify) blue from red



$$D(x) = \text{sign}\left(\mathbf{v}^\top \mathbf{x} + a\right)$$

the decision border:

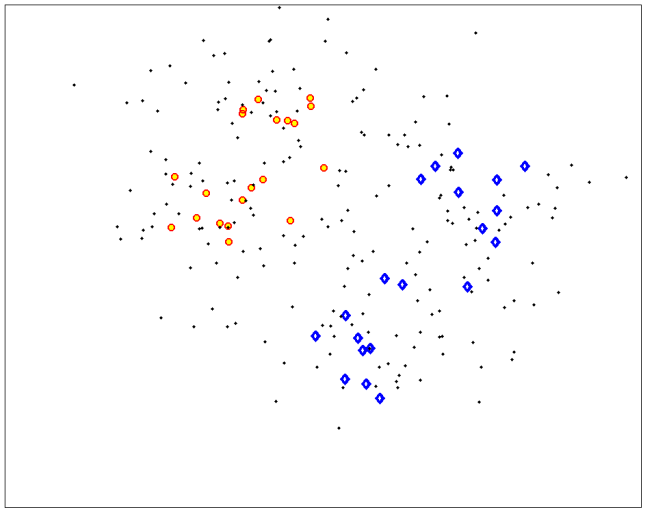
$$\mathbf{v}^\top \mathbf{x} + a = 0$$

there are many solutions...
The problem is ill posed

How to choose a solution?

This is not the problem we want to solve

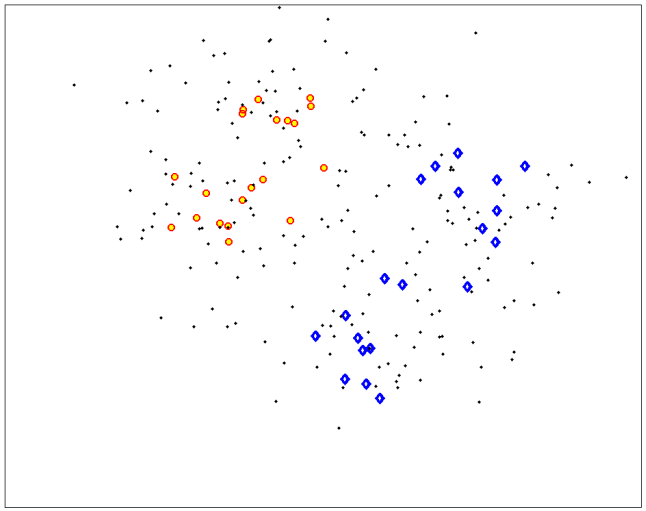
$\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ a training sample, i.i.d. drawn according to $\mathbb{P}(\mathbf{x}, y)$
unknown



we want to be able to classify new observations: minimize $\mathbb{P}(\text{error})$

This is not the problem we want to solve

$\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ a training sample, i.i.d. drawn according to $\mathbb{P}(\mathbf{x}, y)$
unknown



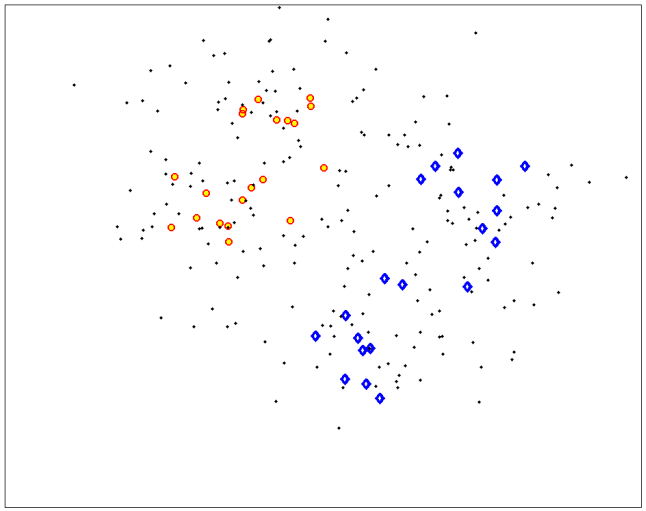
we want to be able to classify new observations: minimize $\mathbb{P}(\text{error})$

Looking for a universal approach

- use training data: (a **few** errors)
- prove $\mathbb{P}(\textit{error})$ remains small
- non linear
- scalable - algorithmic complexity

This is not the problem we want to solve

$\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ a training sample, i.i.d. drawn according to $\mathbb{P}(\mathbf{x}, y)$
unknown



we want to be able to classify new observations: minimize $\mathbb{P}(\text{error})$

Looking for a universal approach

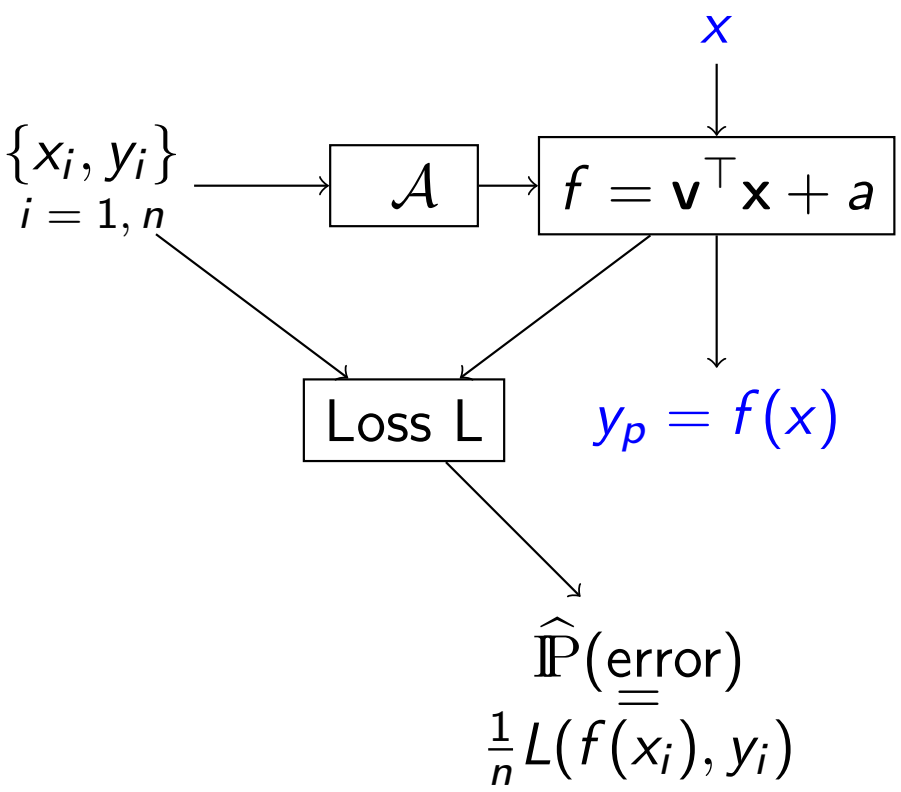
- use training data: (a **few** errors)
- prove $\mathbb{P}(\textit{error})$ remains small
- non linear
- scalable - algorithmic complexity

with large probability:

$$\mathbb{P}(\text{error}) < \underbrace{\widehat{\mathbb{P}}(\text{error})}_{=0 \text{ here}} + \varphi\Big(\underbrace{\|\mathbf{v}\|}_{\frac{1}{\text{margin}}} \Big)$$

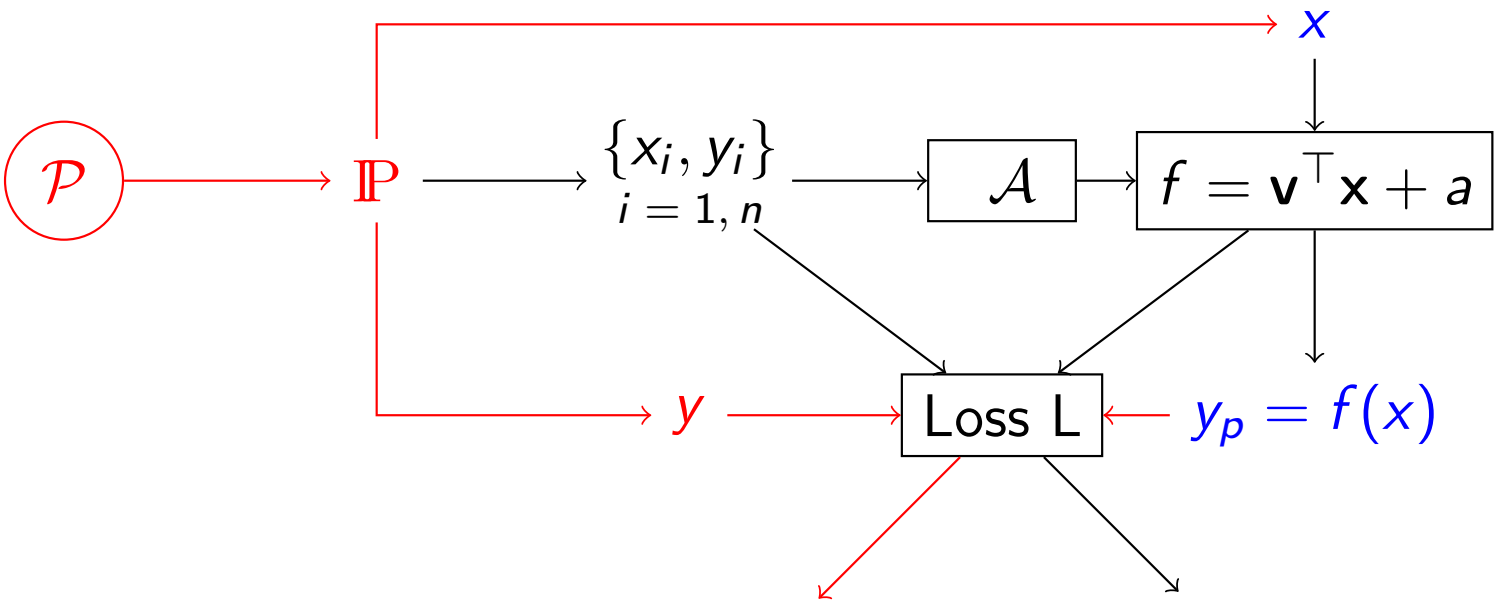
Vapnik's Book, 1982

Statistical machine learning – Computation learning theory (COLT)



Vapnik's Book, 1982

Statistical machine learning – Computation learning theory (COLT)

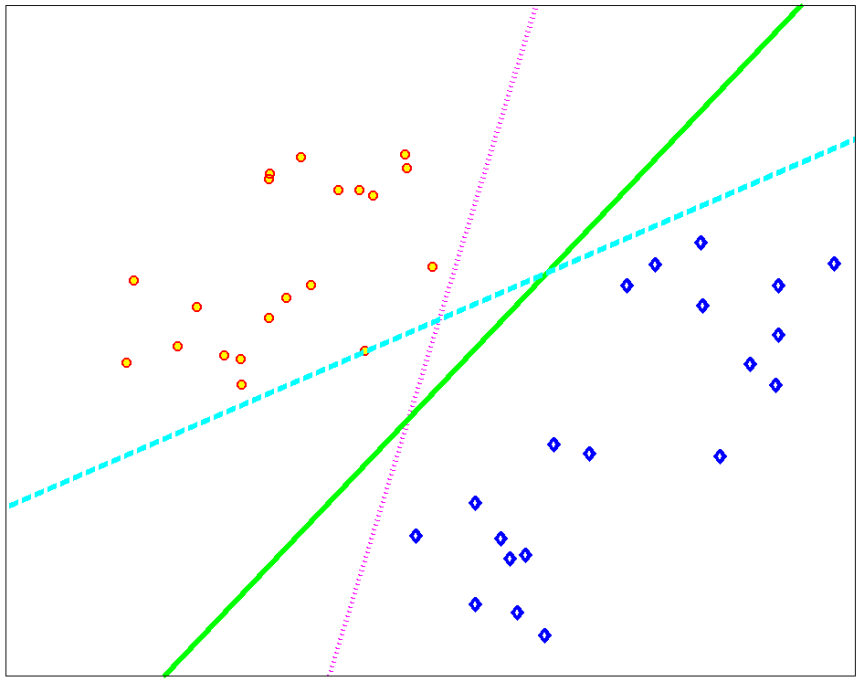


$$\forall \mathbb{P} \in \mathcal{P} \qquad \text{Prob}\Big(\underset{\mathbb{E}(L)}{\underset{=}{\mathbb{P}(\text{error})}} \leq \underset{\frac{1}{n} L(f(x_i), y_i)}{\underset{=}{\widehat{\mathbb{P}}(\text{error})}} + \varphi(\|\mathbf{v}\|) \Big) \geq \delta$$

Vapnik's Book, 1982

linear discrimination

Find a line to classify blue and red



$$D(x) = \text{sign}\left(\mathbf{v}^\top \mathbf{x} + a\right)$$

the decision border:

$$\mathbf{v}^\top \mathbf{x} + a = 0$$

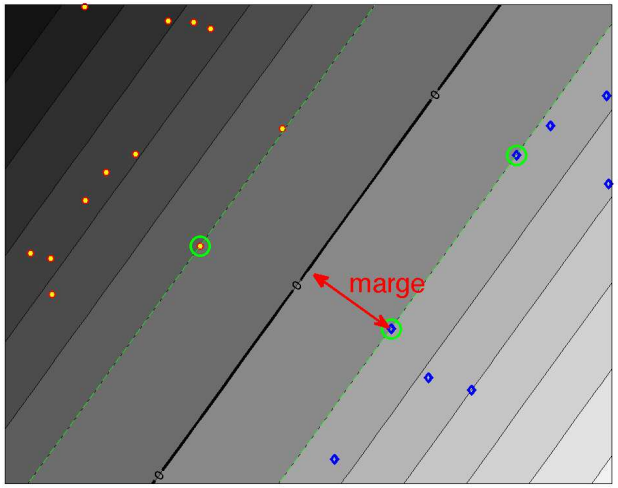
there are many solutions...
The problem is ill posed

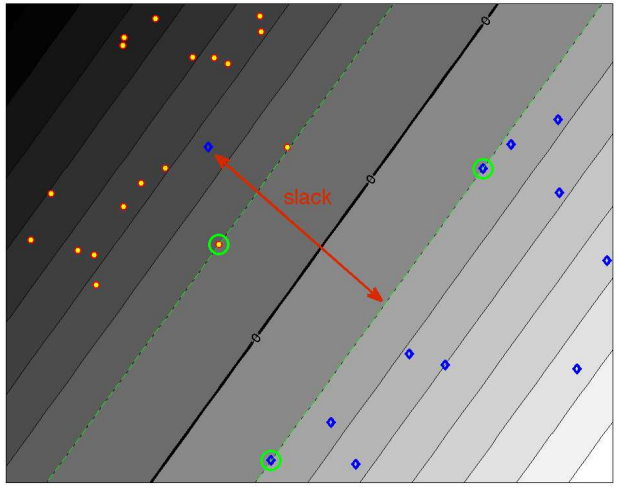
How to choose a solution ?

$\Rightarrow$ choose the one with larger margin

Road map

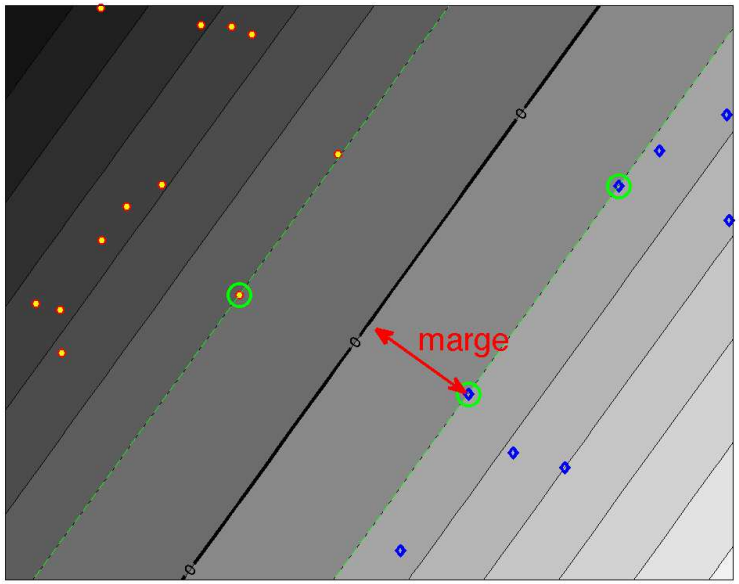
1. Linear SVM
   - Linear classification
   - The margin
   - Linear SVM: the problem
   - Optimization in 5 slides
   - Dual formulation of the linear SVM
   - Solving the dual
   - The non separable case





Maximize our *confidence* = maximize the margin

the decision border: $\Delta(\mathbf{v}, a) = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{v}^\top \mathbf{x} + a = 0\}$



maximize the margin

$$\max_{\mathbf{v},a} \underbrace{\min_{i\in[1,n]} \mathrm{dist}(\mathbf{x}_i, \Delta(\mathbf{v}, a))}_{\text{margin: } m}$$

Maximize the margin

$$\begin{cases} \max\limits_{\mathbf{v},a} & m \\ \text{with} & \min\limits_{i=1,n} \dfrac{|\mathbf{v}^\top \mathbf{x}_i + a|}{\|\mathbf{v}\|} \geq m \end{cases}$$

the problem is still ill posed

if $(\mathbf{v}, a)$ is a solution, $\forall\, 0 < k \;\; (k\mathbf{v}, ka)$ is also a solution...

Margin and distance: details

$$\text{dist}(\mathbf{x}_i, \Delta(\mathbf{v}, a)) \quad = \quad \underset{\mathbf{s}\in\Delta}{\arg\min} \ \|\mathbf{x}_i - \mathbf{s}\|$$

$$\begin{cases} \underset{\mathbf{s}\in\mathbf{R}^d}{\min} & \frac{1}{2}\|\mathbf{x}_i - \mathbf{s}\|^2 \\ \text{with} & \mathbf{v}^\top \mathbf{s} + a = 0 \end{cases}$$

$$L(\mathbf{s}, \lambda) = \frac{1}{2}\|\mathbf{x}_i - \mathbf{s}\|^2 + \lambda(\mathbf{v}^\top \mathbf{s} + a)$$

$$\nabla_{\mathbf{s}} L(\mathbf{s}, \lambda) = -\mathbf{x}_i + \mathbf{s} + \lambda \mathbf{v}$$

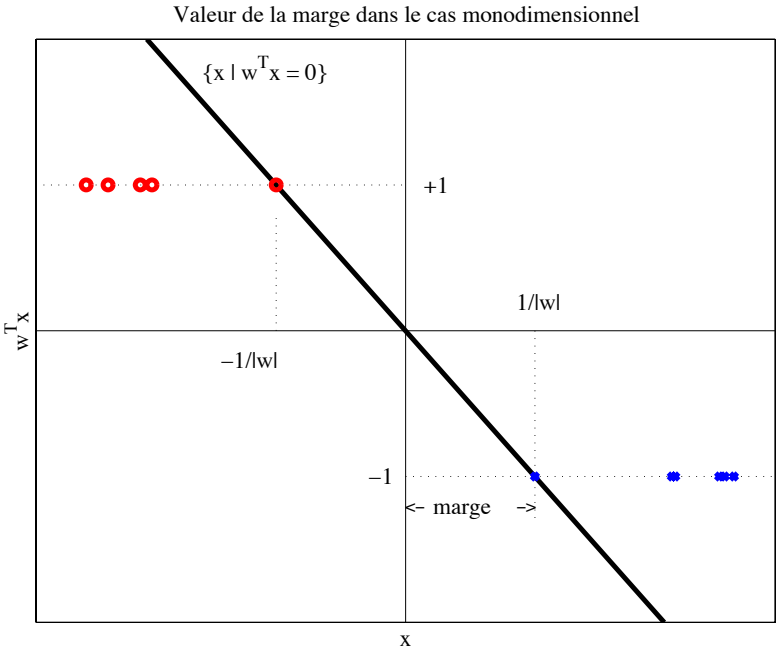
$$\nabla_{\mathbf{s}} L(\mathbf{s}, \lambda) = 0 \iff \mathbf{s} = \mathbf{x}_i - \lambda \mathbf{v}$$

$$\begin{cases} \mathbf{v}^\top \mathbf{s} + a = 0 \\ \mathbf{s} = \mathbf{x}_i - \lambda \mathbf{v} \end{cases} \quad \Longrightarrow \quad \mathbf{v}^\top (\mathbf{x}_i - \lambda \mathbf{v}) + a = 0 \quad \Longrightarrow \quad \lambda = \frac{\mathbf{v}^\top \mathbf{x}_i + a}{\mathbf{v}^\top \mathbf{v}}$$

$$\mathbf{s} = \mathbf{x}_i - \frac{\mathbf{v}^\top \mathbf{x}_i + a}{\mathbf{v}^\top \mathbf{v}} \mathbf{v} \quad \Longrightarrow \quad \|\mathbf{x}_i - \mathbf{s}\|^2 = \left(\frac{\mathbf{v}^\top \mathbf{x}_i + a}{\mathbf{v}^\top \mathbf{v}} \right)^2 \mathbf{v}^\top \mathbf{v}$$

$$\underset{\mathbf{s}\in\Delta}{\arg\min} \ \|\mathbf{x}_i - \mathbf{s}\| = \frac{|\mathbf{v}^\top \mathbf{x}_i + a|}{\|\mathbf{v}\|}$$

Margin, norm and regularity



Maximize the margin

$$\begin{cases} \max\limits_{\mathbf{v},a} & m \\ \text{with} & \min\limits_{i=1,n} |\mathbf{v}^\top \mathbf{x}_i + a| \geq m \\ & \|\mathbf{v}\|^2 = 1 \end{cases}$$

if the min is greater, everybody is greater ($y_i \in \{-1, 1\}$)

$$\begin{cases} \max\limits_{\mathbf{v},a} & m \\ \text{with} & y_i(\mathbf{v}^\top \mathbf{x}_i + a) \geq m, \quad i = 1, n \\ & \|\mathbf{v}\|^2 = 1 \end{cases}$$

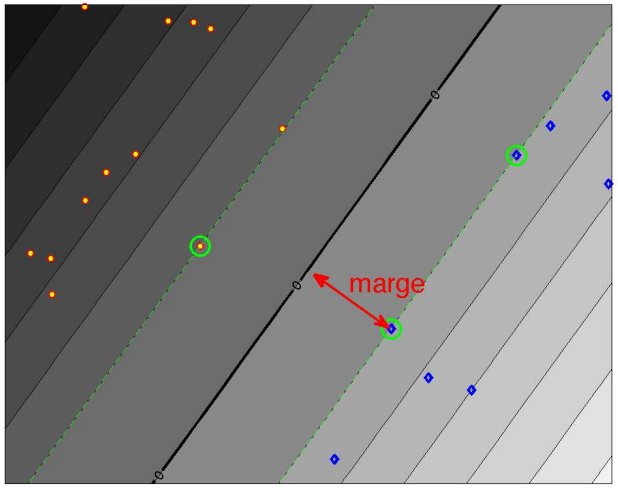
change variable: $\mathbf{w} = \frac{\mathbf{v}}{m}$ and $b = \frac{a}{m} \implies \|\mathbf{w}\| = \frac{\|\mathbf{v}\|}{m}$

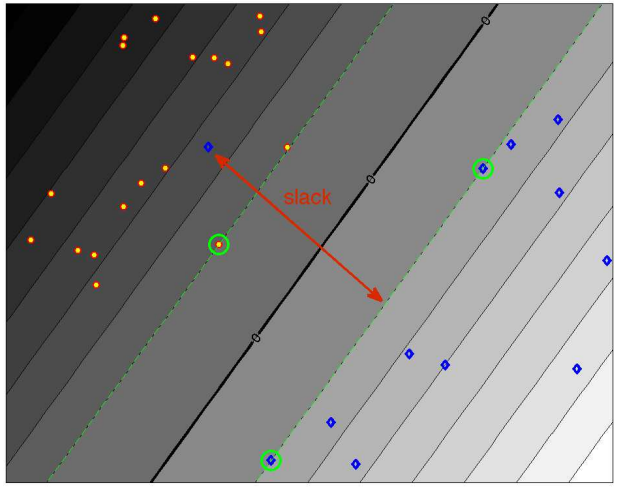
$$\begin{cases} \max\limits_{\mathbf{w},b} & m \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad ; \; i = 1, n \\ \text{and} & m = \frac{1}{\|\mathbf{w}\|} \end{cases}$$

$$\begin{cases} \min\limits_{\mathbf{w},b} & \|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

Road map

1. Linear SVM
   - Linear classification
   - The margin
   - Linear SVM: the problem
   - Optimization in 5 slides
   - Dual formulation of the linear SVM
   - Solving the dual
   - The non separable case





Linear SVM: the problem

Linear SVM are the solution of the following problem (called primal)

Let $\{(\mathbf{x}_i, y_i);\ i = 1 : n\}$ be a set of labelled data with $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, -1\}$.
A support vector machine (SVM) is a linear classifier associated with the following decision function: $D(x) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$ where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ a given thought the solution of the following problem:

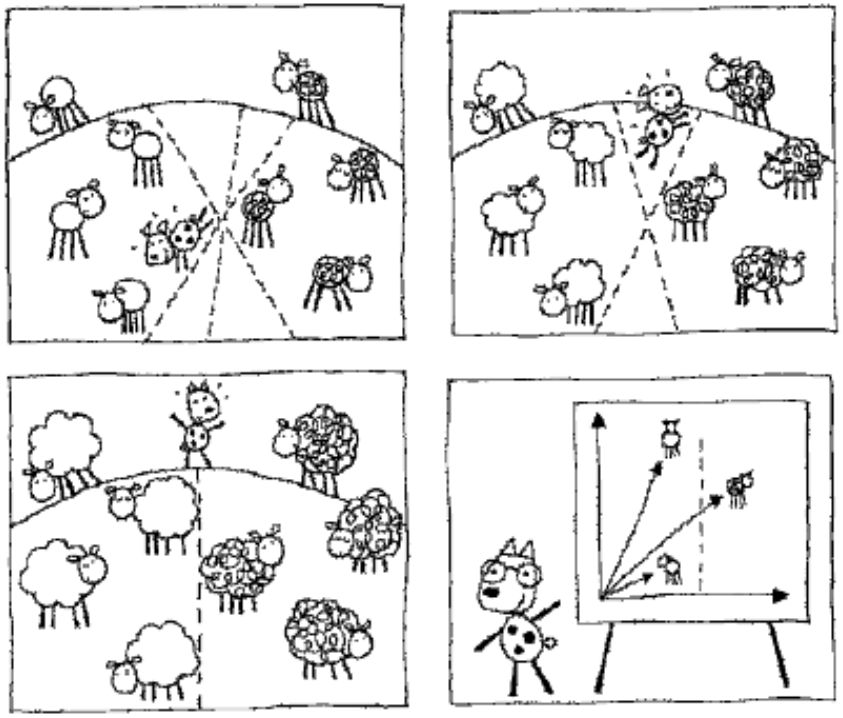
$$\begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w}^\top \mathbf{w} \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

This is a quadratic program (QP):

$$\begin{cases} \min\limits_{\mathbf{z}} & \frac{1}{2}\mathbf{z}^\top A\mathbf{z} - \mathbf{d}^\top \mathbf{z} \\ \text{with} & B\mathbf{z} \leq \mathbf{e} \end{cases}$$

$\mathbf{z} = (\mathbf{w}, b)^\top$, $\mathbf{d} = (0, \ldots, 0)^\top$, $A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$, $B = -[\mathbf{y}X, \mathbf{y}]$ et $\mathbf{e} = -(1, \ldots, 1)^\top$

The linear discrimination problem



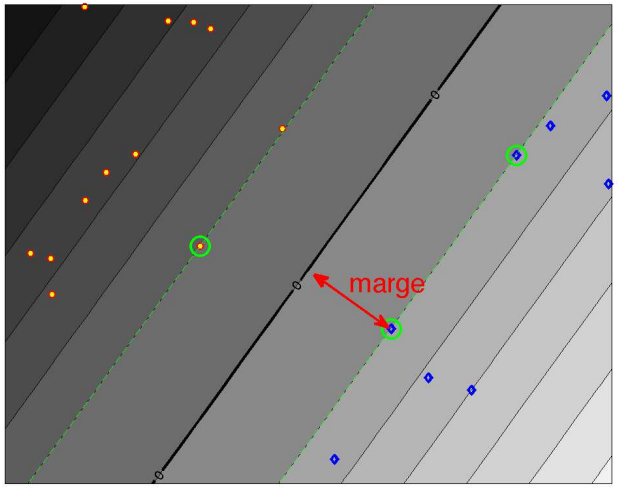
...the story of the sheep dog who was herding his sheep, and serendipitously invented the large margin classification and Sheep Vectors ...

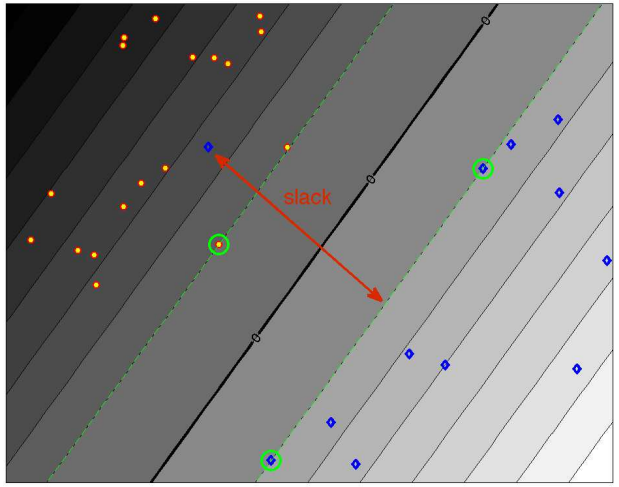
(drawing by Ana Martin Larranaga)

from *Learning with Kernels*, B. Schölkopf and A. Smolla, MIT Press, 2002.

Road map

1. Linear SVM
   - Linear classification
   - The margin
   - Linear SVM: the problem
   - Optimization in 5 slides
   - Dual formulation of the linear SVM
   - Solving the dual
   - The non separable case





First order optimality condition (1)

$$\text{problem } \mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, \ldots, p \\ \text{and} & g_i(x) \leq 0 \; i = 1, \ldots, q \end{cases}$$

Definition: Karush, Kuhn and Tucker (KKT) conditions

$$\begin{array}{rll} \text{stationarity} & \nabla J(x^\star) + \sum_{j=1}^{p} \lambda_j \nabla h_j(x^\star) + \sum_{i=1}^{q} \mu_i \nabla g_i(x^\star) = 0 & \\ \text{primal admissibility} & h_j(x^\star) = 0 & j = 1, \ldots, p \\ & g_i(x^\star) \leq 0 & i = 1, \ldots, q \\ \text{dual admissibility} & \mu_i \geq 0 & i = 1, \ldots, q \\ \text{complementarity} & \mu_i g_i(x^\star) = 0 & i = 1, \ldots, q \end{array}$$

λ_j and μ_i are called the Lagrange multipliers of problem $\mathcal{P}$

First order optimality condition (2)

Theorem (12.1 Nocedal & Wright pp 321)

If a vector $x^\star$ is a stationary point of problem $\mathcal{P}$
Then there exists[a] Lagrange multipliers such that $\left(x^\star, \{\lambda_j\}_{j=1:p}, \{\mu_i\}_{i=1:q}\right)$ fulfill KKT conditions

[a] under some conditions *e.g.* linear independence constraint qualification

If the problem is convex, then a stationary point is the solution of the problem

A quadratic program (QP) is convex when...

$$(QP) \quad \begin{cases} \min\limits_{\mathbf{z}} & \frac{1}{2}\mathbf{z}^\top A\mathbf{z} - \mathbf{d}^\top \mathbf{z} \\ \text{with} & B\mathbf{z} \leq \mathbf{e} \end{cases}$$

...when matrix A is positive definite

KKT condition - Lagrangian (3)

$$\text{problem } \mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, \ldots, p \\ \text{and} & g_i(x) \leq 0 \; i = 1, \ldots, q \end{cases}$$

Definition: Lagrangian

The lagrangian of problem $\mathcal{P}$ is the following function:

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = J(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x)$$

The importance of being a lagrangian

- the stationarity condition can be written: $\nabla \mathcal{L}(\mathbf{x}^\star, \lambda, \mu) = 0$
- the lagrangian saddle point $\max\limits_{\lambda,\mu} \min\limits_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \mu)$

Primal variables: x and dual variables λ, μ (the Lagrange multipliers)

Duality – definitions (1)

Primal and (Lagrange) dual problems

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbf{R}^n} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, p \\ \text{and} & g_i(x) \leq 0 \quad i = 1, q \end{cases} \qquad \mathcal{D} = \begin{cases} \max\limits_{\lambda\in\mathbf{R}^p, \mu\in\mathbf{R}^q} & Q(\lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \end{cases}$$

Dual objective function:

$$\begin{aligned} Q(\lambda, \mu) &= \inf_x \ \mathcal{L}(\mathbf{x}, \lambda, \mu) \\ &= \inf_x \ J(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x) \end{aligned}$$

Wolf dual problem

$$\mathcal{W} = \begin{cases} \max\limits_{\mathbf{x},\lambda\in\mathbf{R}^p, \mu\in\mathbf{R}^q} & \mathcal{L}(\mathbf{x}, \lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \\ \text{and} & \nabla J(x^\star) + \sum\limits_{j=1}^{p} \lambda_j \nabla h_j(x^\star) + \sum\limits_{i=1}^{q} \mu_i \nabla g_i(x^\star) = 0 \end{cases}$$

Duality – theorems (2)

Theorem (12.12, 12.13 and 12.14 Nocedal & Wright pp 346)

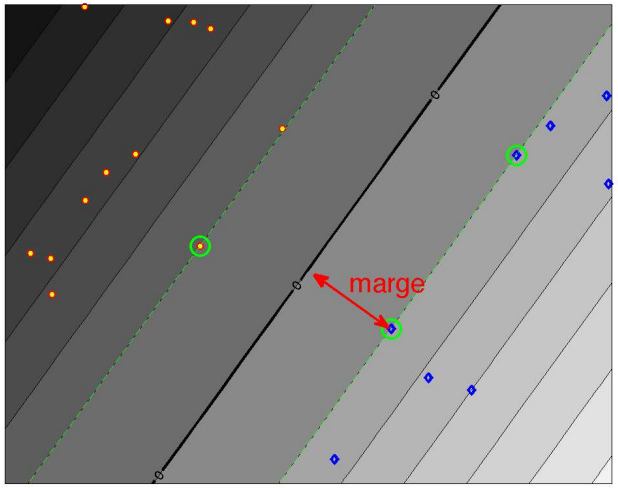
If f, g and h are convex and continuously differentiable[a]
Then the solution of the dual problem is the same as the solution of the primal

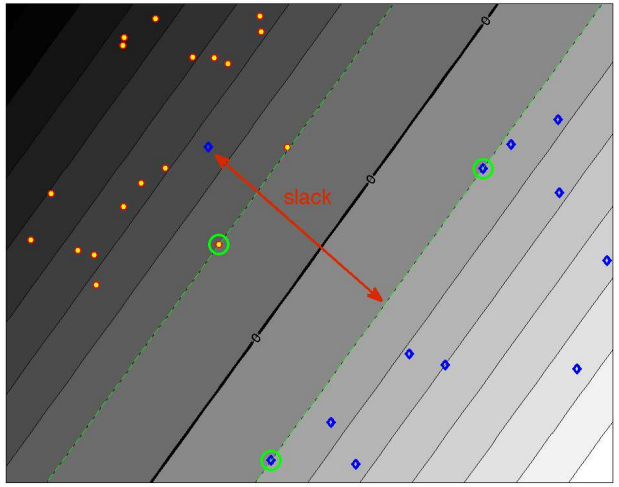
[a] under some conditions *e.g.* linear independence constraint qualification

$$
\begin{aligned}
(\lambda^\star, \mu^\star) &= \text{solution of problem } \mathcal{D} \\
\mathbf{x}^\star &= \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^\star, \mu^\star)
\end{aligned}
$$

Road map

1. Linear SVM
   - Linear classification
   - The margin
   - Linear SVM: the problem
   - Optimization in 5 slides
   - Dual formulation of the linear SVM
   - Solving the dual
   - The non separable case





Linear SVM dual formulation - The lagrangian

$$\begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \qquad i = 1, n \end{cases}$$

Looking for the lagrangian saddle point $\max\limits_{\alpha} \min\limits_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha)$ with so called lagrange multipliers $\alpha_i \geq 0$

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\big)$$

α_i represents the influence of constraint thus the influence of the training example (x_i, y_i)

Optimality conditions

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\big)$$

Computing the gradients: $\begin{cases} \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \alpha) & = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = \sum_{i=1}^{n} \alpha_i \; y_i \end{cases}$

we have the following optimality conditions

$$\begin{cases} \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \alpha) & = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i \; y_i = 0 \end{cases}$$

Linear SVM dual formulation

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\big)$$

Optimality: $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad \sum_{i=1}^{n} \alpha_i \; y_i = 0$

$$\begin{aligned}
\mathcal{L}(\alpha) &= \tfrac{1}{2} \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i}_{\mathbf{w}^\top \mathbf{w}} - \textstyle\sum_{i=1}^{n} \alpha_i y_i \underbrace{\sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j^\top}_{\mathbf{w}^\top} \mathbf{x}_i - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0} + \textstyle\sum_{i=1}^{n} \alpha_i \\
&= -\frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i + \sum_{i=1}^{n} \alpha_i
\end{aligned}$$

Dual linear SVM is also a quadratic program

$$\text{problem } \mathcal{D} \quad \begin{cases} \min\limits_{\alpha \in \mathbf{R}^n} & \frac{1}{2}\alpha^\top G \alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

with G a symmetric matrix $n \times n$ such that $G_{ij} = y_i y_j \mathbf{x}_j^\top \mathbf{x}_i$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d+1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d+1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \, y_i(\mathbf{x}^\top \mathbf{x}_i) + b$$

Cold case: the least square problem

Linear model

$$y_i = \sum_{j=1}^{d} \mathbf{w}_j x_{ij} + \varepsilon_i \quad , \qquad i = 1, n$$

n data and d variables; $d < n$

$$\min_{\mathbf{w}} = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} x_{ij} \mathbf{w}_j - y_i \right)^2 = \|X\mathbf{w} - \mathbf{y}\|^2$$

Solution: $\widetilde{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{y}$

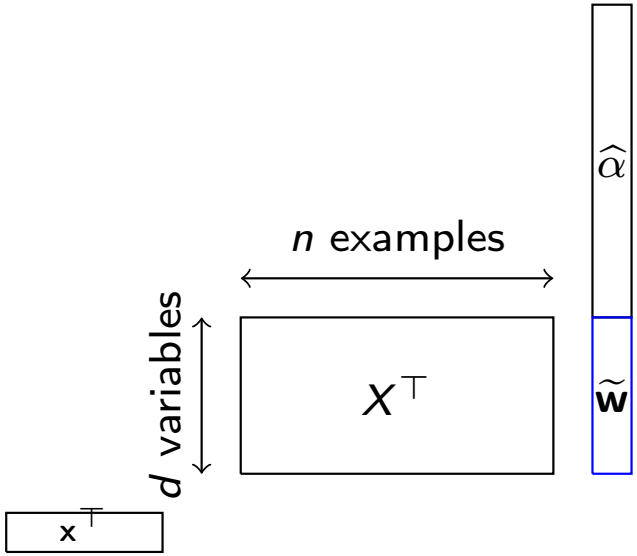
$$f(\mathbf{x}) = \mathbf{x}^\top \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{\mathbf{w}}}$$

What is the influence of each data point (matrix X lines) ?

data point influence (contribution)

for any new data point $\mathbf{x}$

$$
\begin{aligned}
f(\mathbf{x}) &= \mathbf{x}^\top (X^\top X)(X^\top X)^{-1} \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{\mathbf{w}}} \\
&= \mathbf{x}^\top X^\top \underbrace{X(X^\top X)^{-1}(X^\top X)^{-1} X^\top \mathbf{y}}_{\widehat{\alpha}}
\end{aligned}
$$

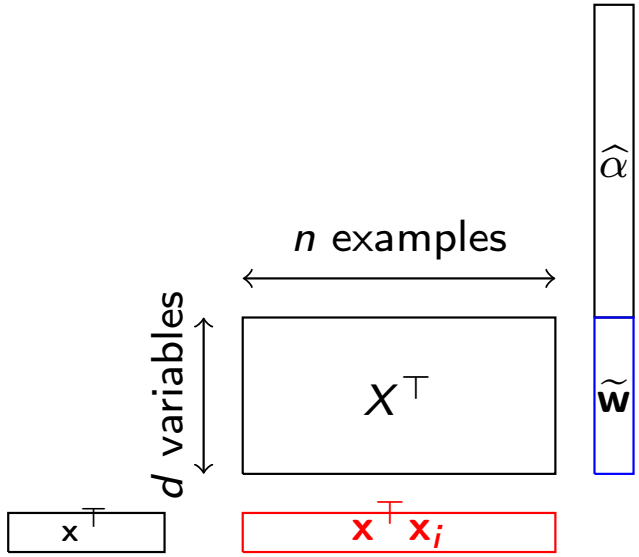


$$
f(\mathbf{x}) = \sum_{j=1}^{d} \widetilde{\mathbf{w}}_j x_j
$$

data point influence (contribution)

for any new data point $\mathbf{x}$

$$
\begin{aligned}
f(\mathbf{x}) &= \mathbf{x}^\top \, (X^\top X)(X^\top X)^{-1} \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{\mathbf{w}}} \\
&= \mathbf{x}^\top \, X^\top \, \underbrace{X(X^\top X)^{-1}(X^\top X)^{-1} X^\top \mathbf{y}}_{\widehat{\alpha}}
\end{aligned}
$$



$$
f(\mathbf{x}) \;=\; \sum_{j=1}^{d} \widetilde{\mathbf{w}}_j x_j = \sum_{i=1}^{n} \widehat{\alpha}_i \, (\mathbf{x}^\top \mathbf{x}_i)
$$

from variables to examples

$$
\underbrace{\widehat{\alpha} = X(X^\top X)^{-1}\widetilde{\mathbf{w}}}_{n \text{ examples}} \qquad \text{et} \qquad \underbrace{\widetilde{\mathbf{w}} = X^\top \widehat{\alpha}}_{d \text{ variables}}
$$

what if $d \geq n$!

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d+1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

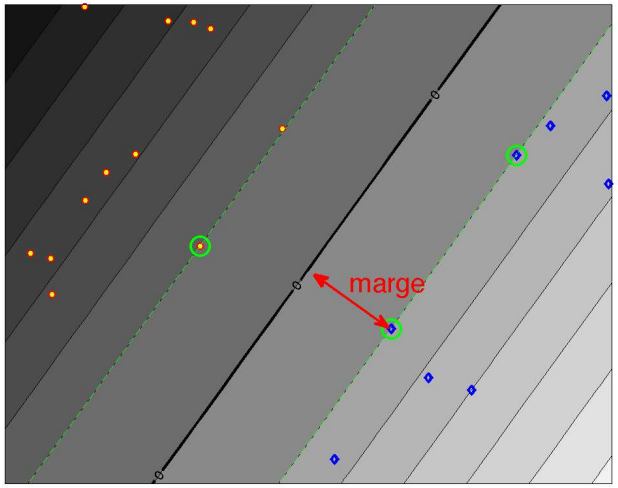
$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad i = 1, n \end{cases}$$

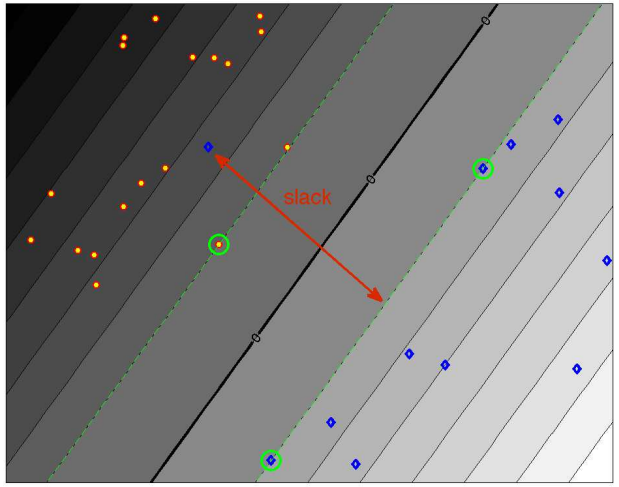
- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \, y_i(\mathbf{x}^\top \mathbf{x}_i) + b$$

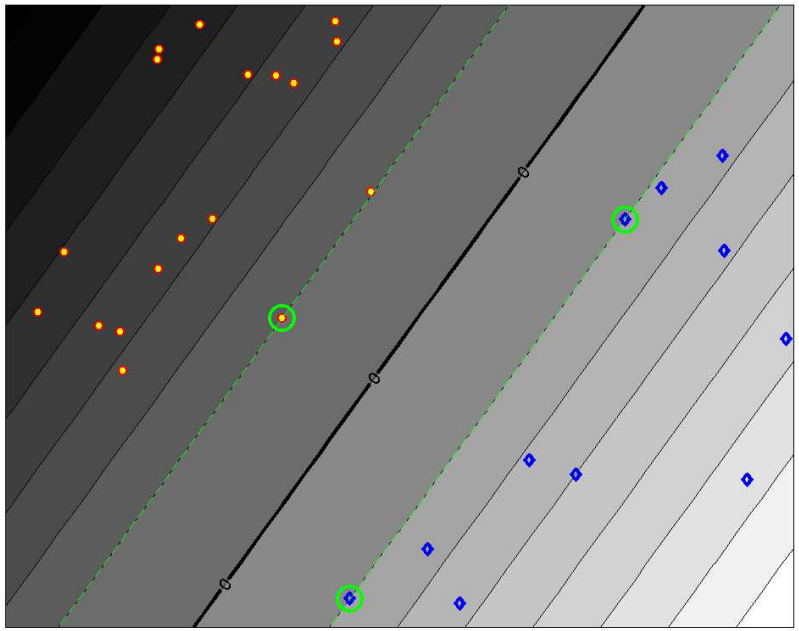
Road map

1. Linear SVM
   - Linear classification
   - The margin
   - Linear SVM: the problem
   - Optimization in 5 slides
   - Dual formulation of the linear SVM
   - Solving the dual
   - The non separable case





Solving the dual (1)

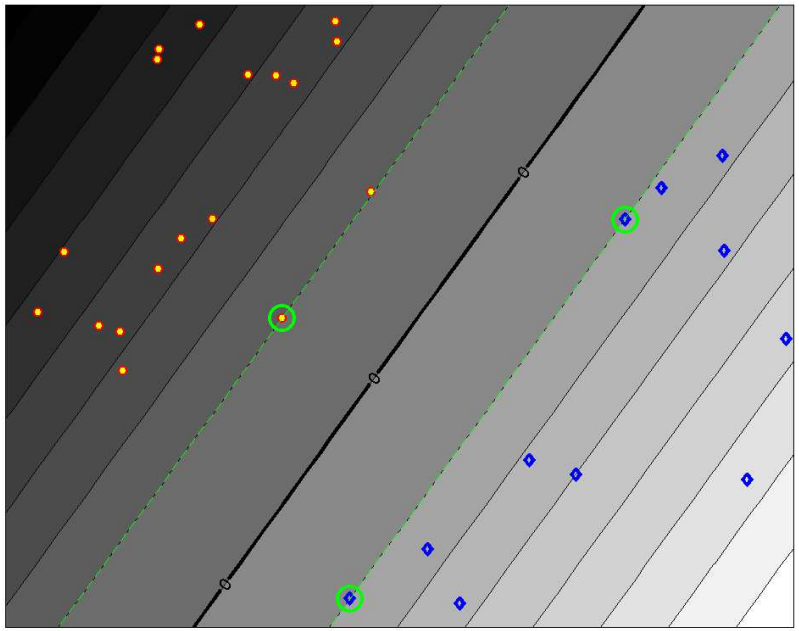


Data point influence

- $\alpha_i = 0$ this point is useless
- $\alpha_i \neq 0$ this point is said to be support

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \, y_i (\mathbf{x}^\top \mathbf{x}_i) + b$$

Solving the dual (1)



Data point influence

- $\alpha_i = 0$ this point is useless
- $\alpha_i \neq 0$ this point is said to be support

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{3} \alpha_i \, y_i (\mathbf{x}^\top \mathbf{x}_i) + b$$

Decison border only depends on 3 points ($d + 1$)

Solving the dual (2)

Assume we know these 3 data points

$$\begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases} \implies \begin{cases} \min\limits_{\alpha \in \mathbb{R}^3} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \end{cases}$$

$$L(\alpha, b) = \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha + b\, y^\top \alpha$$

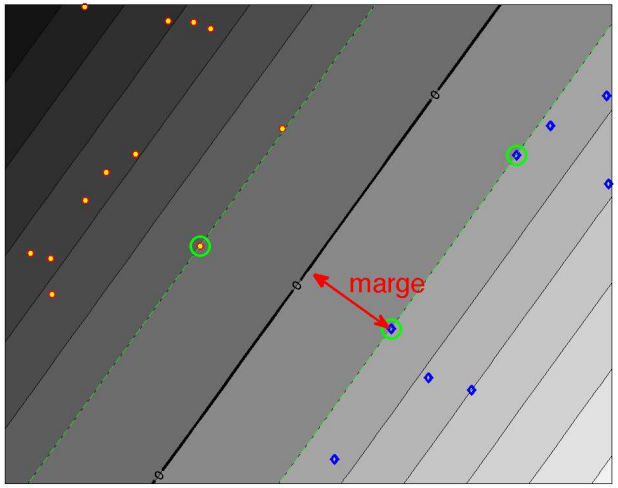
solve the following linear system

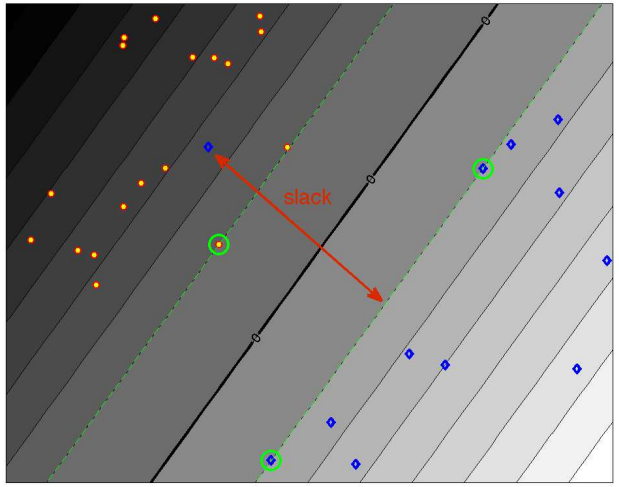
$$\begin{cases} G\alpha + \quad b\,\mathbf{y} & = \mathbf{e} \\ \mathbf{y}^\top \alpha & = 0 \end{cases}$$

```
U = chol(G);  % upper
a = U\ (U'\e);
c = U\ (U'\y);
b = (y'*a)\(y'*c)
alpha = U\ (U'\(e - b*y));
```

Road map

1. Linear SVM
   - Linear classification
   - The margin
   - Linear SVM: the problem
   - Optimization in 5 slides
   - Dual formulation of the linear SVM
   - Solving the dual
   - The non separable case

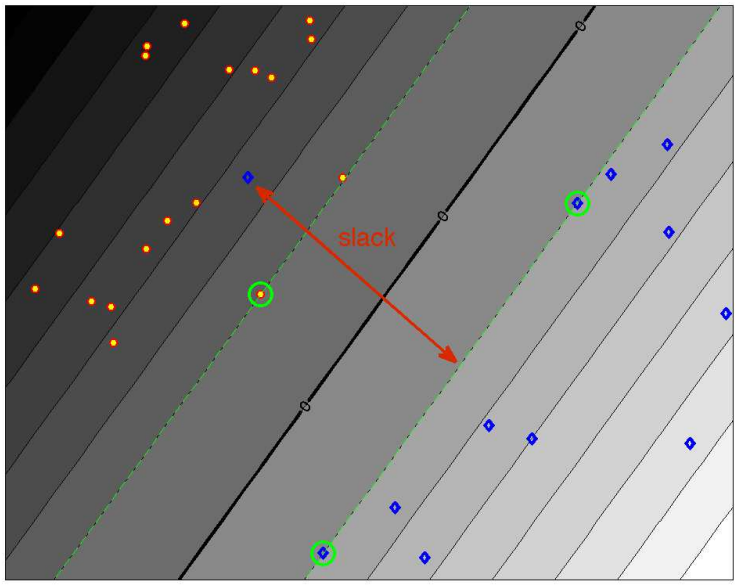




The non separable case: a bi criteria optimization problem

Modeling potential errors: introducing slack variables ξ_i

$$(x_i, y_i) \qquad \begin{cases} \text{no error:} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \Rightarrow \quad \xi_i = 0 \\ \text{error:} & \qquad\qquad\qquad\qquad\quad \xi_i = 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0 \end{cases}$$



$$\begin{cases} \min\limits_{\mathbf{w},b,\xi} & \dfrac{1}{2}\|\mathbf{w}\|^2 \\ \min\limits_{\mathbf{w},b,\xi} & \dfrac{C}{p}\displaystyle\sum_{i=1}^{n} \xi_i^p \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, n \end{cases}$$

Our hope: almost all $\xi_i = 0$

The non separable case

Modeling potential errors: introducing slack variables ξ_i

$$(x_i, y_i) \qquad \begin{cases} \text{no error:} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \Rightarrow \quad \xi_i = 0 \\ \text{error:} & \qquad\qquad\qquad\qquad\quad \xi_i = 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0 \end{cases}$$

Minimizing also the slack (the error), for a given $C > 0$

$$\begin{cases} \min\limits_{\mathbf{w}, b, \xi} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{p}\sum\limits_{i=1}^{n} \xi_i^p \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, n \\ & \xi_i \geq 0 \qquad\qquad\qquad\qquad i = 1, n \end{cases}$$

Looking for the saddle point of the lagrangian with the Lagrange multipliers $\alpha_i \geq 0$ and $\beta_i \geq 0$

$$\mathcal{L}(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{p}\sum_{i=1}^{n} \xi_i^p - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{n} \beta_i \xi_i$$

Optimality conditions ($p = 1$)

$$\mathcal{L}(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\big(y_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{n}\beta_i\xi_i$$

Computing the gradients:
$$\begin{cases} \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, b, \alpha) & = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial\mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = \sum_{i=1}^{n}\alpha_i\, y_i \\ \nabla_{\xi_i}\mathcal{L}(\mathbf{w}, b, \alpha) & = C - \alpha_i - \beta_i \end{cases}$$

- no change for $\mathbf{w}$ and b
- $\beta_i \geq 0$ and $C - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i \leq C$

The dual formulation:

$$\begin{cases} \min\limits_{\alpha\in\mathbf{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \leq C \qquad i = 1, n \end{cases}$$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi\in\mathbb{R}^n} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{with} & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, n \end{cases}$$

- $d + n + 1$ unknown
- $2n$ constraints
- classical QP
- to be used when n is too large to build G

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \leq C \qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- $2n$ box constraints
- easy to solve
- to be used when n is not too large

Optimality conditions ($p = 2$)

$$\mathcal{L}(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\big(y_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{n}\beta_i\xi_i$$

Computing the gradients:
$$\begin{cases} \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, b, \alpha) & = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial\mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = \sum_{i=1}^{n}\alpha_i\, y_i \\ \nabla_{\xi_i}\mathcal{L}(\mathbf{w}, b, \alpha) & = C\xi_i - \alpha_i - \beta_i \end{cases}$$

- no change for $\mathbf{w}$ and b
- $C\xi_i - \alpha_i - \beta_i = 0 \;\Rightarrow\; \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i\xi_i = -\frac{1}{2C}\sum_{i=1}^{n}\alpha_i^2$

The dual formulation:

$$\begin{cases} \min\limits_{\alpha\in\mathbf{R}^n} & \frac{1}{2}\alpha^\top(G + \frac{1}{C}I)\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \le \alpha_i \qquad\qquad i = 1, n \end{cases}$$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi\in\mathbb{R}^n} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{C}\sum\limits_{i=1}^{n}\xi_i^2 \\ \text{with} & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, n \end{cases}$$

- $d + n + 1$ unknown
- $2n$ constraints
- classical QP
- to be used when n is too large to build G

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top(G + \frac{1}{C}I)\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix is regularized
- n box constraints
- easy to solve
- to be used when n is not too large

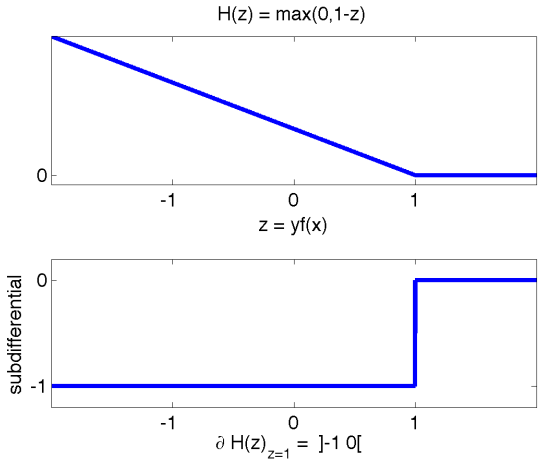
Eliminating the slack but not the possible mistakes

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi\in\mathbb{R}^n} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum\limits_{i=1}^{n}\xi_i^2 \\ \text{with} & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, n \end{cases}$$

Introducing the hinge loss

$$\xi_i = \max\big(1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1, 0\big)$$

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\max\big(1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1, 0\big)^2$$



Back to $d + 1$ variables, but this is no longer an explicit QP

SVM (minimizing the hinge loss) with variable selection

LP SVM

$$\min_{\mathbf{w},b} \quad \|\mathbf{w}\|_1 + \; C\sum_{i=1}^{n} \max\big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, 0\big)$$

Lasso SVM

$$\min_{\mathbf{w},b} \quad \|\mathbf{w}\|_1 + \; \frac{C}{2}\sum_{i=1}^{n} \max\big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, 0\big)^2$$

Elastic net SVM (doubly regularized with p=1 or p=2)

$$\min_{\mathbf{w},b} \quad \|\mathbf{w}\|_1 + \tfrac{1}{2}\|\mathbf{w}\|^2 + \; C\sum_{i=1}^{n} \max\big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, 0\big)^p$$

Danzig selector for SVM, Generalized Garrote

Conclusion: variables or data point?

- seeking for a universal learning algorithm
  - no model for $\mathbb{P}(\mathbf{x}, y)$
- the linear case: data is separable
  - the non separable case
- double objective: minimizing the error together with the regularity of the solution
  - multi objective optimisation
- dualiy : variable – example
  - use the primal when $d < n$ (in the liner case) or when matrix G is hard to compute
  - otherwise use the dual
- universality = nonlinearity
  - kernels