L5 Support Vector Classification

Support Vector Machine

- Problem definition
- Geometrical picture
- Optimization problem

Optimization Problem

- Hard margin
- Convexity
- Dual problem
- Soft margin problem

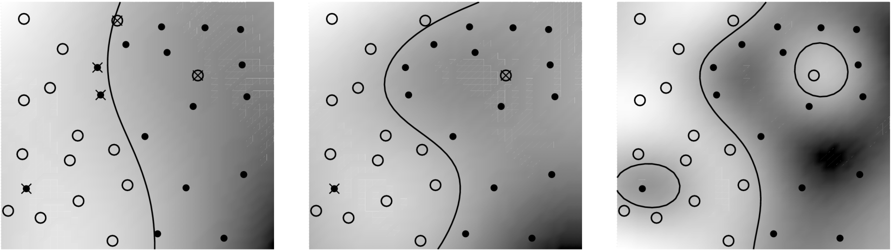
Classification

Data

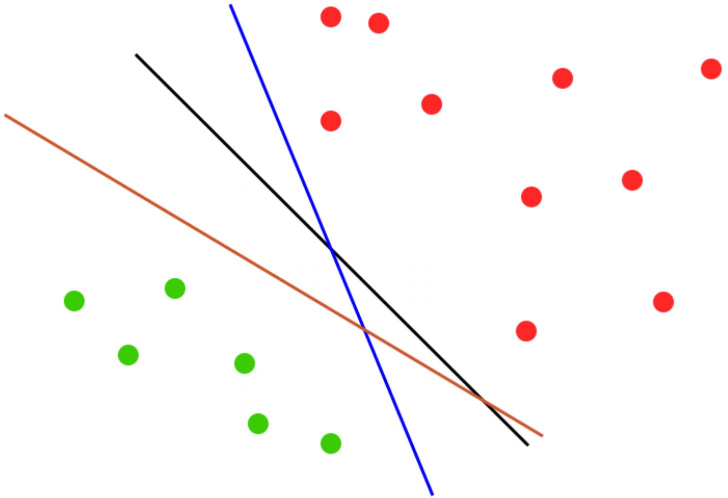
Pairs of observations (x_i, y_i) generated from some distribution $\mathrm{P}(x, y)$, e.g., (blood status, cancer), (credit transaction, fraud), (profile of jet engine, defect)

Task

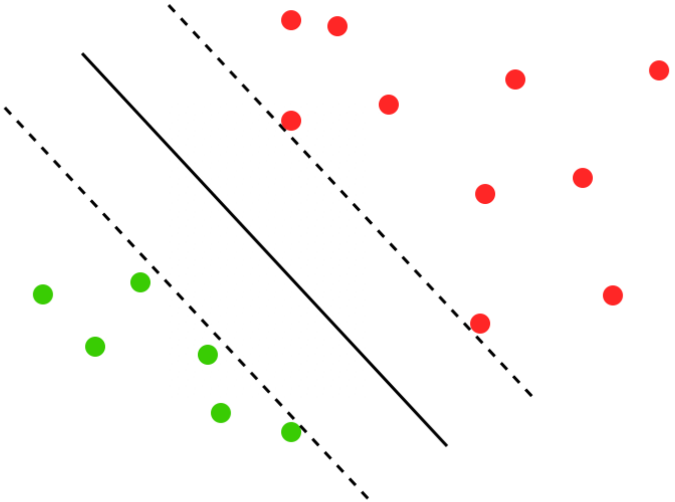
- Estimate y given x at a new location.
- Modification: find a function $f(x)$ that does the task.



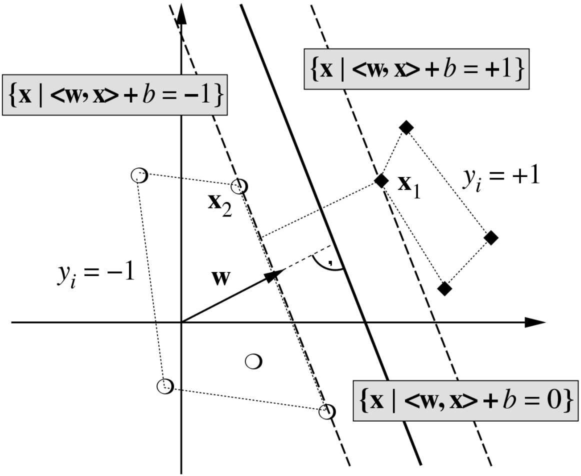
So Many Solutions



One to rule them all . . .



Optimal Separating Hyperplane



Note:

$$\langle \mathbf{w}, \mathbf{x}_1 \rangle + b = +1$$
$$\langle \mathbf{w}, \mathbf{x}_2 \rangle + b = -1$$

$$\Rightarrow \quad \langle \mathbf{w}, (\mathbf{x}_1 - \mathbf{x}_2) \rangle = 2$$

$$\Rightarrow \left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, (\mathbf{x}_1 - \mathbf{x}_2) \right\rangle = \frac{2}{\|\mathbf{w}\|}$$

Optimization Problem

Margin to Norm

- Separation of sets is given by $\frac{2}{\|w\|}$ so maximize that.
- Equivalently minimize $\frac{1}{2}\|w\|$.
- Equivalently minimize $\frac{1}{2}\|w\|^2$.

Constraints

- Separation with margin, i.e.

$$\begin{aligned} \langle w, x_i \rangle + b &\geq 1 && \text{if } y_i = 1 \\ \langle w, x_i \rangle + b &\leq -1 && \text{if } y_i = -1 \end{aligned}$$

- Equivalent constraint

$$y_i(\langle w, x_i \rangle + b) \geq 1$$

Optimization Problem

Mathematical Programming Setting

Combining the above requirements we obtain

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|w\|^2 \\ &\text{subject to} && y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \text{ for all } 1 \leq i \leq m \end{aligned}$$

Properties

- Problem is convex
- Hence it has unique minimum
- Efficient algorithms for solving it exist

Lagrange Function

Objective Function $\frac{1}{2}\|w\|^2$.

Constraints $c_i(w, b) := 1 - y_i(\langle w, x_i \rangle + b) \leq 0$

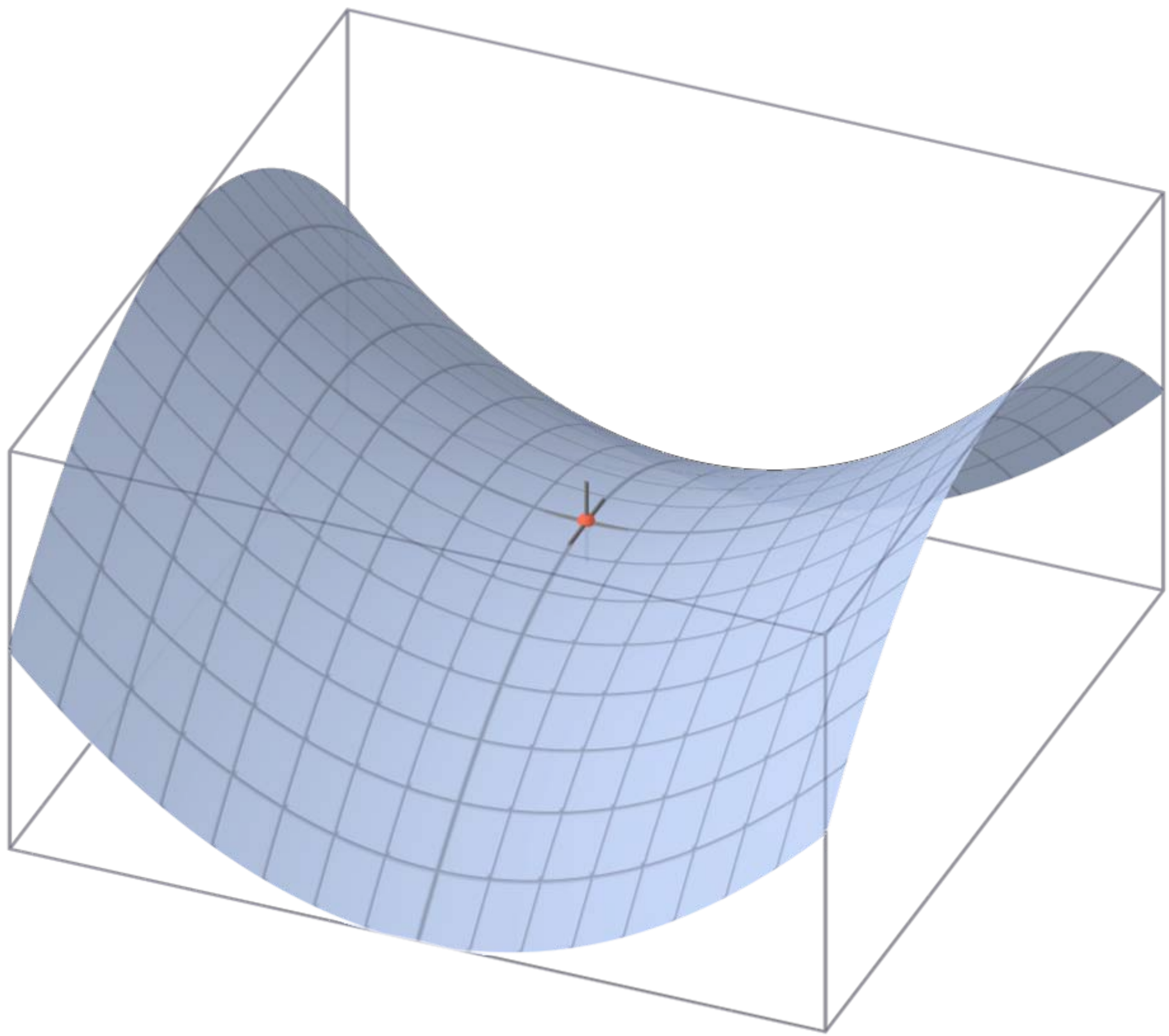
Lagrange Function

$$\begin{aligned} L(w, b, \alpha) &= \text{PrimalObjective} + \sum_i \alpha_i c_i \\ &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\langle w, x_i \rangle + b)) \end{aligned}$$

Saddle Point Condition

Derivatives of L with respect to w and b must vanish.

Saddle Point of $z = x^2 - y^2$



Support Vector Machines

Optimization Problem

$$\text{minimize } \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle - \sum_{i=1}^{m}\alpha_i$$

$$\text{subject to } \sum_{i=1}^{m}\alpha_i y_i = 0 \text{ and } \alpha_i \geq 0$$

Support Vector Expansion

$$w = \sum_i \alpha_i y_i x_i \text{ and hence } f(x) = \sum_{i=1}^{m}\alpha_i y_i \langle x_i, x\rangle + b$$

Kuhn Tucker Conditions

$$\alpha_i(1 - y_i(\langle x_i, x\rangle + b)) = 0$$

Proof (optional)

Lagrange Function

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\langle w, x_i \rangle + b))$$

Saddlepoint condition

$$\partial_w L(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y_i x_i \quad = 0 \iff \quad w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

$$\partial_b L(w, b, \alpha) = - \sum_{i=1}^{m} \alpha_i y_i x_i \quad = 0 \iff \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

To obtain the dual optimization problem we have to substitute the values of w and b into L. Note that the dual variables α_i have the constraint $\alpha_i \geq 0$.

Proof (optional)

Dual Optimization Problem

After substituting in terms for b, w the Lagrange function becomes

$$-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^{m}\alpha_i$$

$$\text{subject to } \sum_{i=1}^{m}\alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \text{ for all } 1 \leq i \leq m$$

Practical Modification

Need to **maximize** dual objective function. Rewrite as

$$\text{minimize } \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle - \sum_{i=1}^{m}\alpha_i$$

subject to the above constraints.

Support Vector Expansion

Solution in $$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

- w is given by a linear combination of training patterns x_i. **Independent of the dimensionality of x.**
- w depends on the Lagrange multipliers α_i.

Kuhn-Tucker-Conditions

- At optimal solution $\mathrm{Constraint} \cdot \mathrm{Lagrange\ Multiplier} = 0$
- In our context this means

$$\alpha_i (1 - y_i(\langle w, x_i \rangle + b)) = 0.$$

Equivalently we have

$$\alpha_i \neq 0 \Longrightarrow y_i \left(\langle w, x_i \rangle + b \right) = 1$$

Only points at the decision boundary can contribute to the solution.

Mini Summary

Linear Classification

- Many solutions
- Optimal separating hyperplane
- Optimization problem

Support Vector Machines

- Quadratic problem
- Lagrange function
- Dual problem

Interpretation

- Dual variables and SVs
- SV expansion
- Hard margin and infinite weights

Kernels

Nonlinearity via Feature Maps

Replace x_i by $\Phi(x_i)$ in the optimization problem.

Equivalent optimization problem

$$\text{minimize } \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^{m} \alpha_i$$

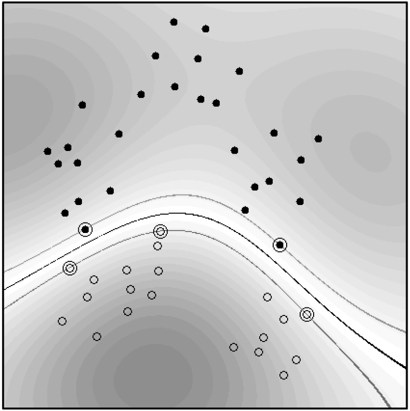
$$\text{subject to } \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0$$

Decision Function

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i) \text{ implies}$$

$$f(x) = \langle w, \Phi(x)\rangle + b = \sum_{i=1}^{m} \alpha_i y_i k(x_i, x) + b.$$

Examples and Problems



Advantage

Works well when the data is noise free.

Problem

Already a single wrong observation can ruin everything — we require $y_i f(x_i) \geq 1$ for all i.

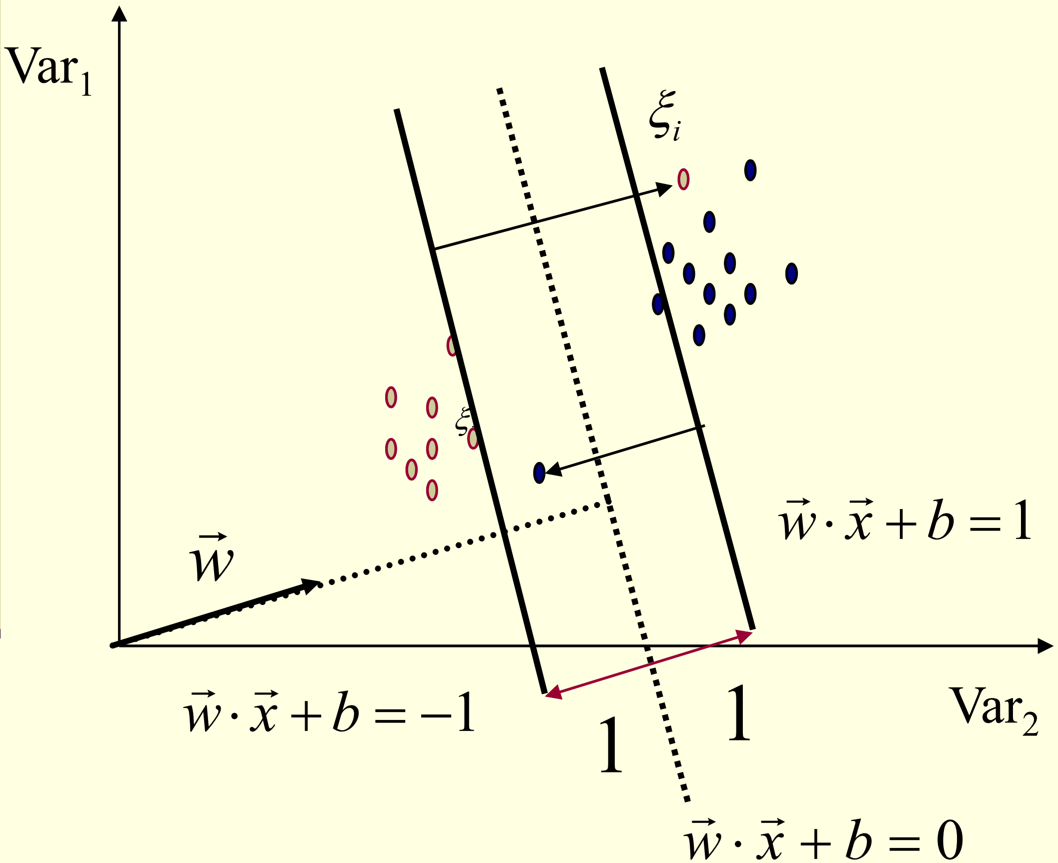
Idea

Limit the influence of individual observations by making the constraints less stringent (introduce slacks).

Support Vector Machines

- Three main ideas:
  1. Define what an optimal hyperplane is (in way that can be identified in a computationally efficient way): maximize margin
  2. Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications
  3. Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space

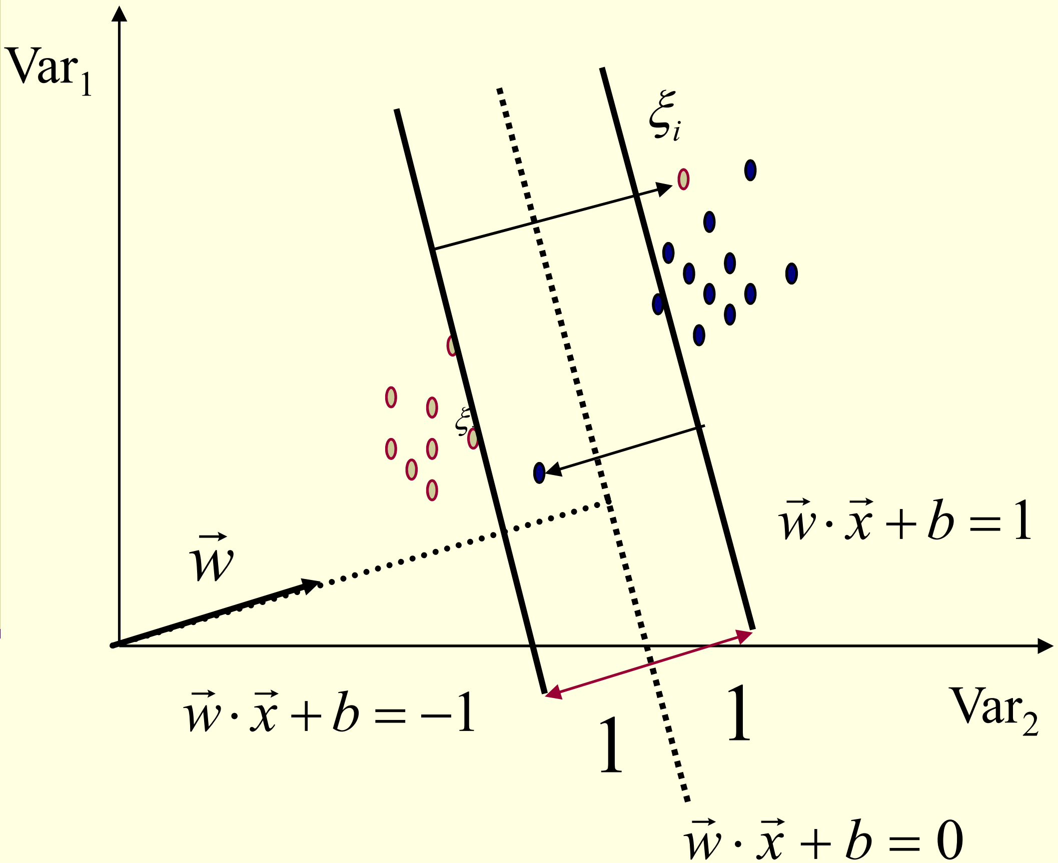
Non-Linearly Separable Data



Introduce slack variables ξ_i

Allow some instances to fall within the margin, but penalize them

Formulating the Optimization Problem



Constraint becomes :

$$y_i(w\cdot x_i+b)\geq 1-\xi_i,\ \forall x_i$$
$$\xi_i\geq 0$$

Objective function penalizes for misclassified instances and those within the margin

$$\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

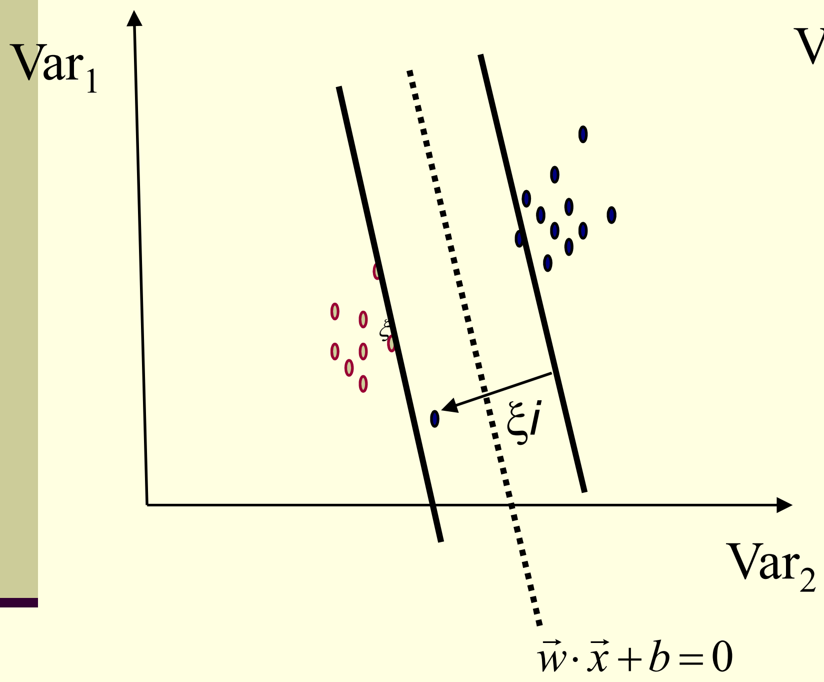
C trades-off margin width and misclassifications

Linear, Soft-Margin SVMs

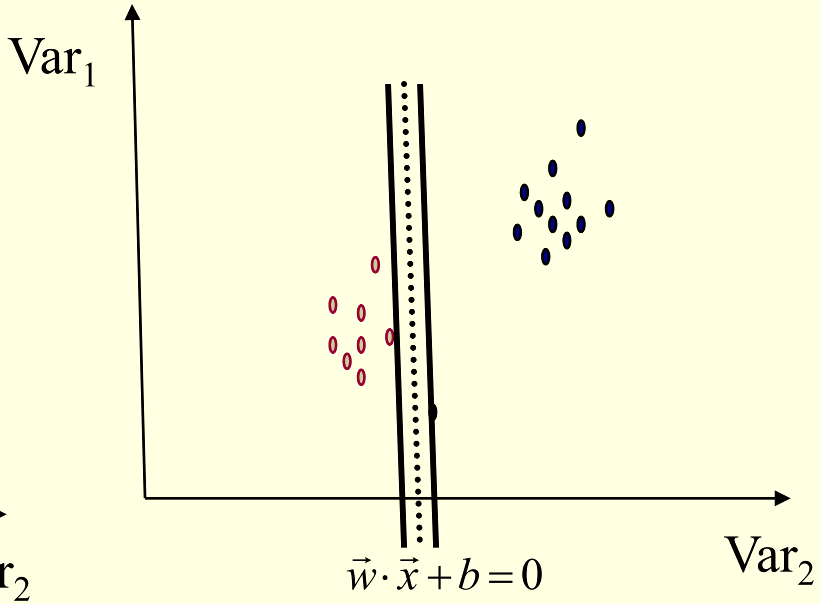
$$\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \qquad y_i(w\cdot x_i + b) \geq 1 - \xi_i,\ \forall x_i$$
$$\xi_i \geq 0$$

- Algorithm tries to maintain ξ_i to zero while maximizing margin
- Notice: algorithm does not minimize the *number* of misclassifications (NP-complete problem) but the sum of distances from the margin hyperplanes
- Other formulations use ξ_i^2 instead
- As $C \to \infty$, we get closer to the hard-margin solution

Robustness of Soft vs Hard Margin SVMs



Soft Margin SVN



Hard Margin SVN

Soft vs Hard Margin SVM

- Soft-Margin always have a solution
- Soft-Margin is more robust to outliers
  - Smoother surfaces (in the non-linear case)
- Hard-Margin does not require to guess the cost parameter (requires no parameters at all)

Optimization Problem (Soft Margin)

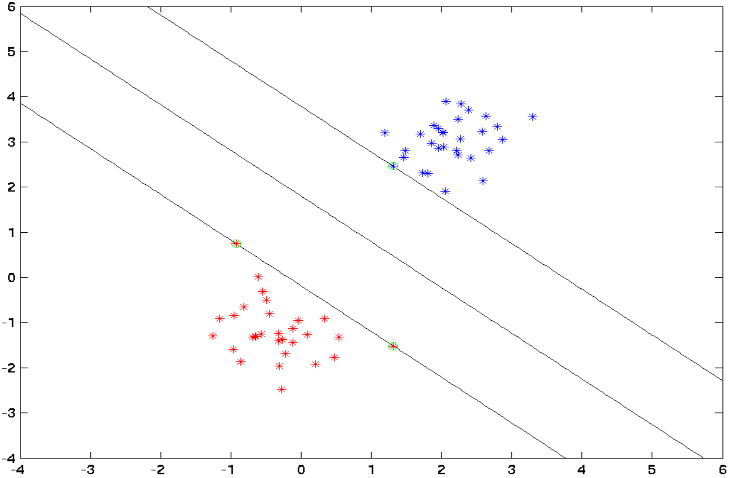
Recall: Hard Margin Problem

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|w\|^2 \\ &\text{subject to} && y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \end{aligned}$$

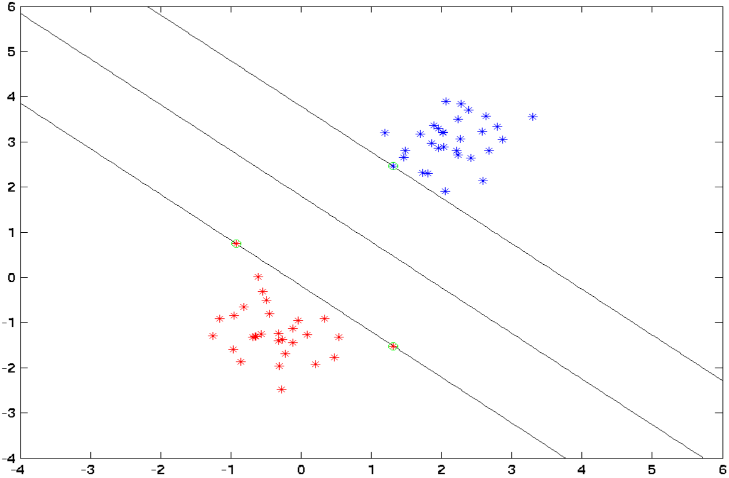
Softening the Constraints

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{subject to} && y_i(\langle w, x_i \rangle + b) - 1 + \xi_i \geq 0 \text{ and } \xi_i \geq 0 \end{aligned}$$

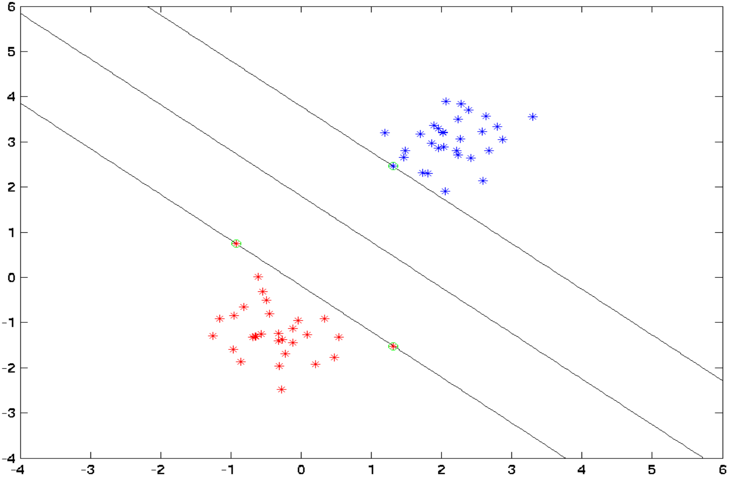
Linear SVM $C = 1$



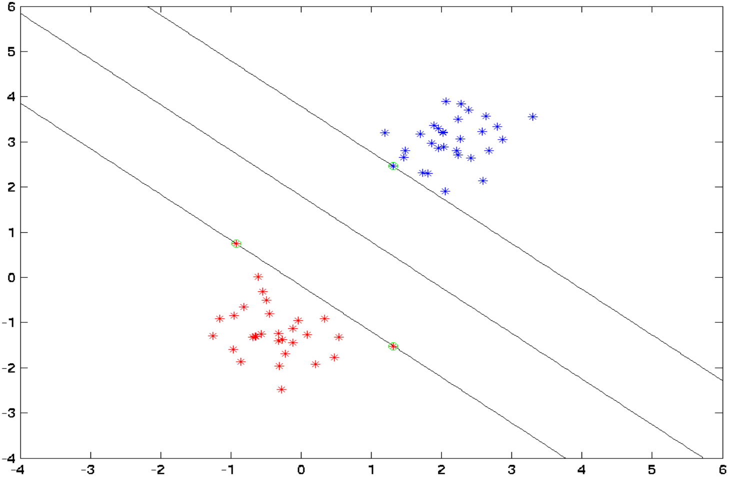
Linear SVM $C = 2$



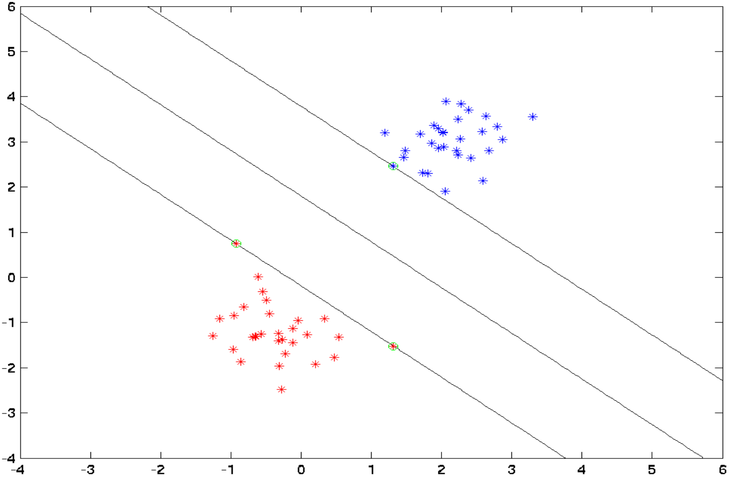
Linear SVM $C = 5$



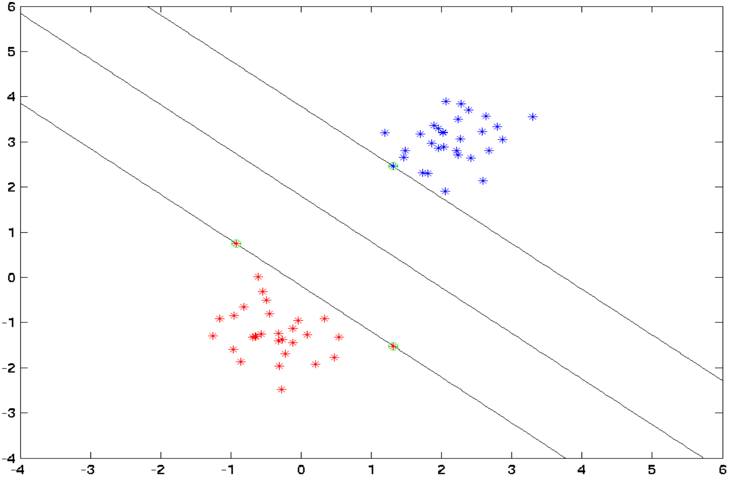
Linear SVM $C = 10$



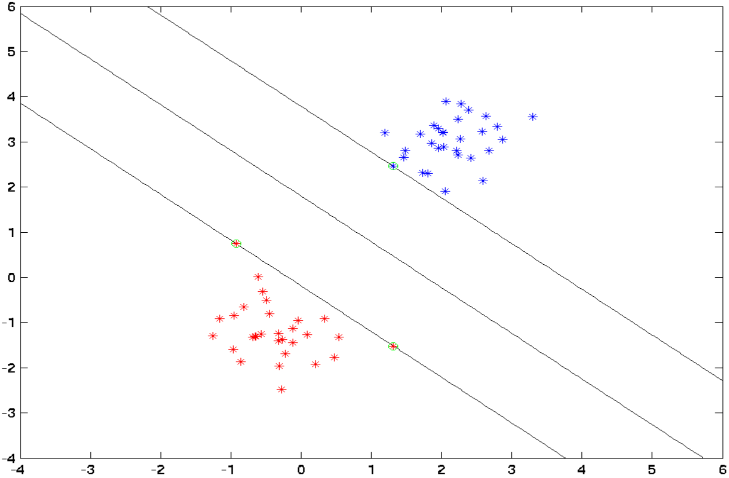
Linear SVM $C = 20$



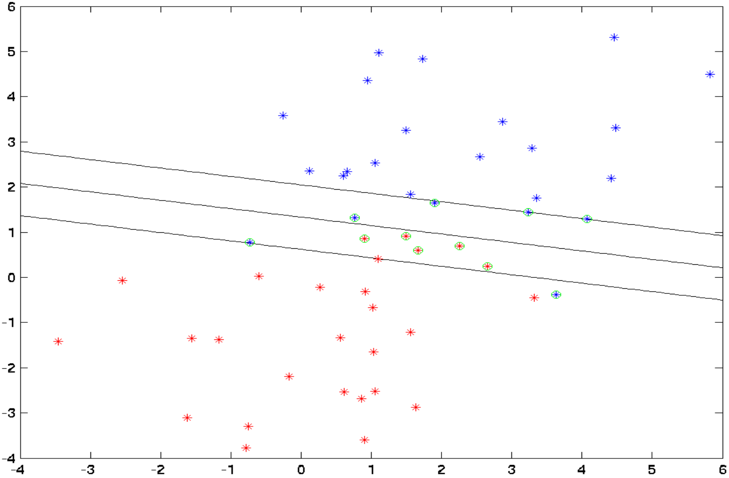
Linear SVM $C = 50$



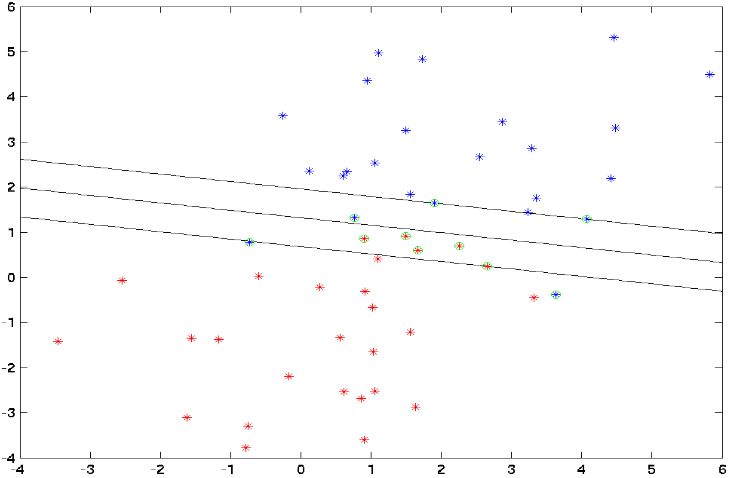
Linear SVM $C = 100$



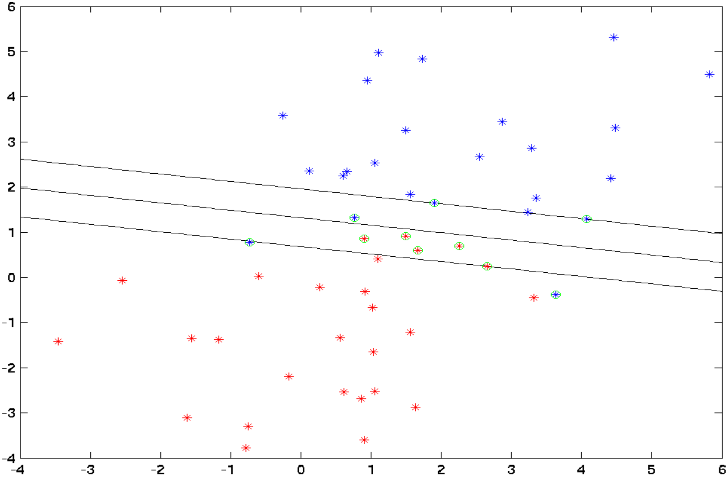
Linear SVM $C = 1$



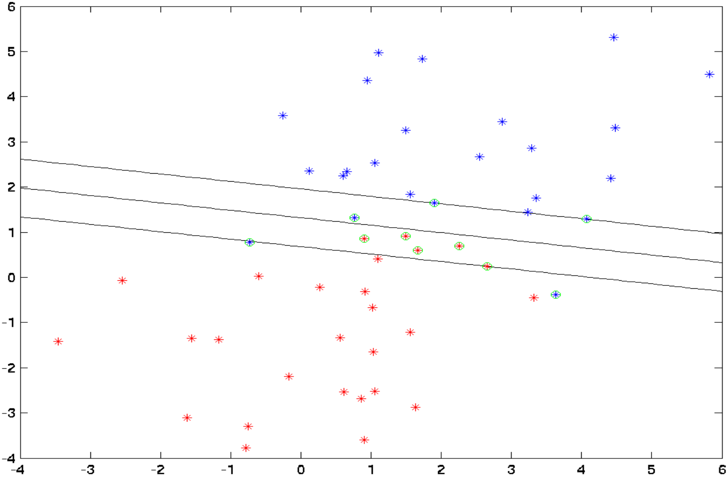
Linear SVM $C = 2$



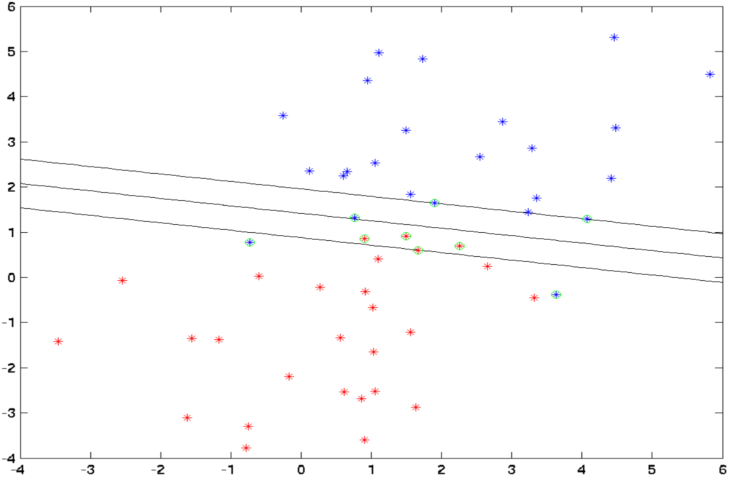
Linear SVM $C = 5$



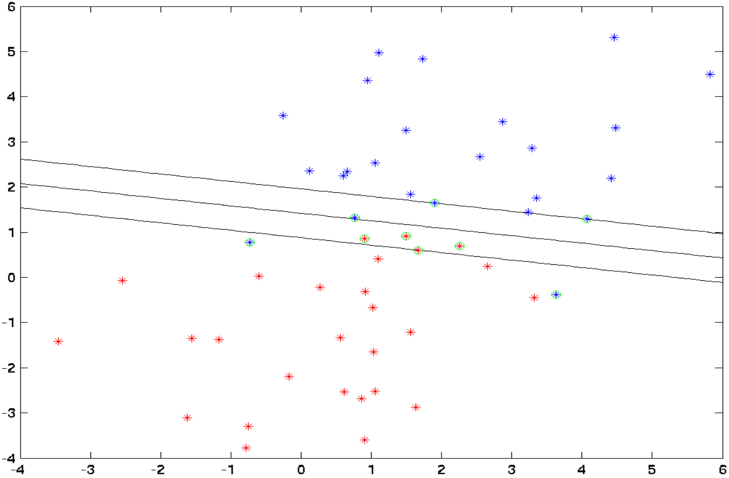
Linear SVM $C = 10$



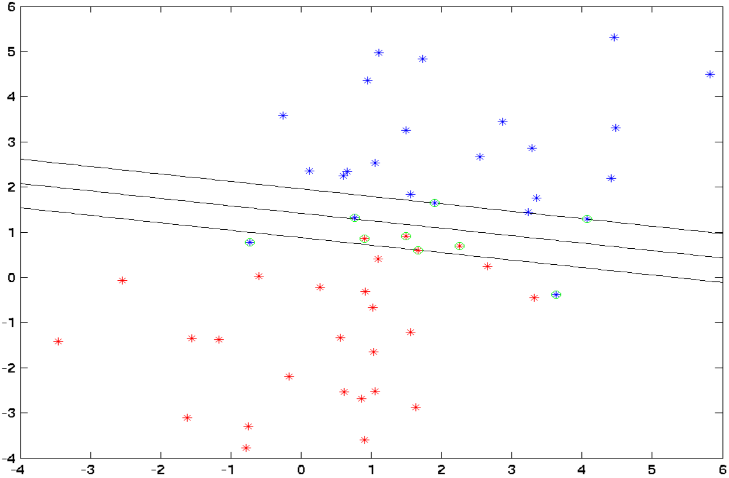
Linear SVM $C = 20$



Linear SVM $C = 50$



Linear SVM $C = 100$



Insights

Changing C

- For clean data C doesn't matter much.
- For noisy data, large C leads to narrow margin (SVM tries to do a good job at separating, even though it isn't possible)

Noisy data

- Clean data has few support vectors
- Noisy data leads to data in the margins
- More support vectors for noisy data

Dual Optimization Problem

Optimization Problem

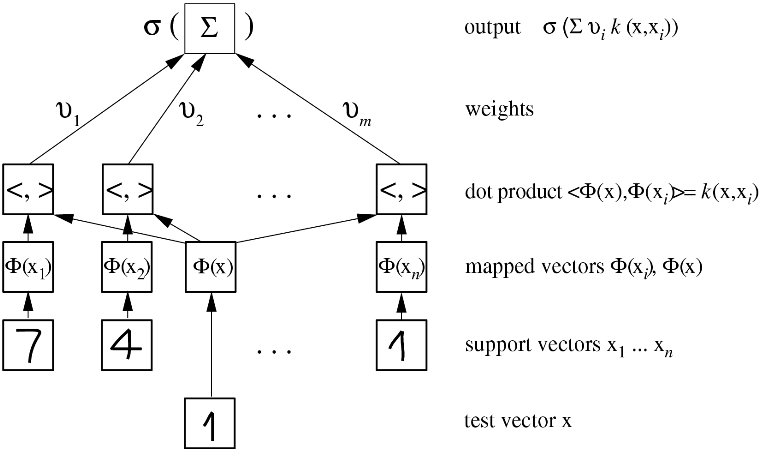
$$\text{minimize } \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^{m} \alpha_i$$

$$\text{subject to } \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } C \geq \alpha_i \geq 0 \text{ for all } 1 \leq i \leq m$$

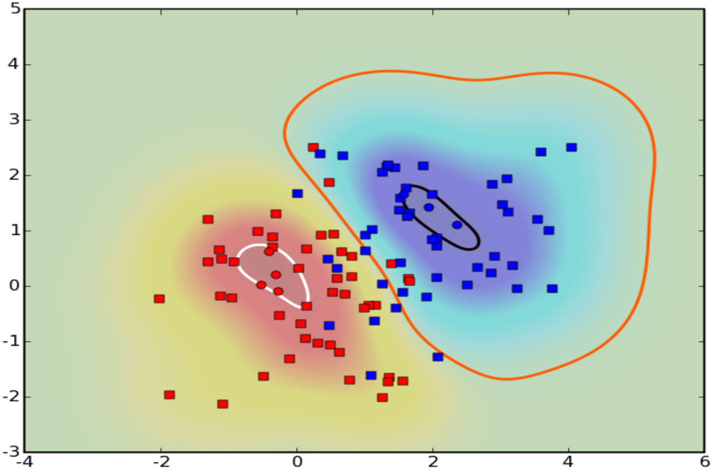
Interpretation

- Almost same optimization problem as before
- Constraint on weight of each α_i (bounds influence of pattern).
- Efficient solvers exist (more about that tomorrow).

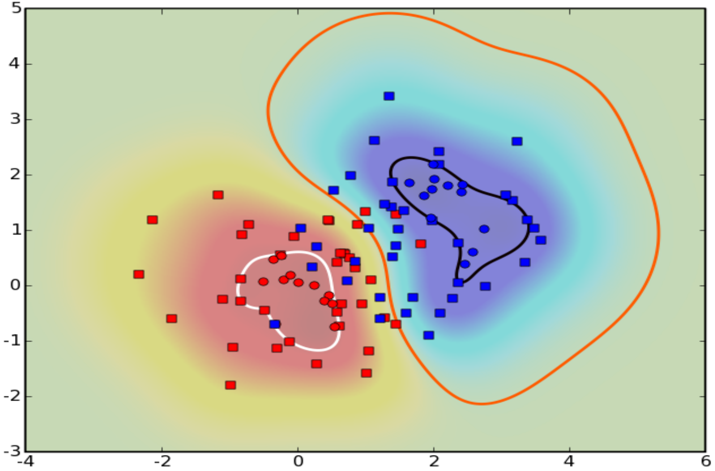
SV Classification Machine



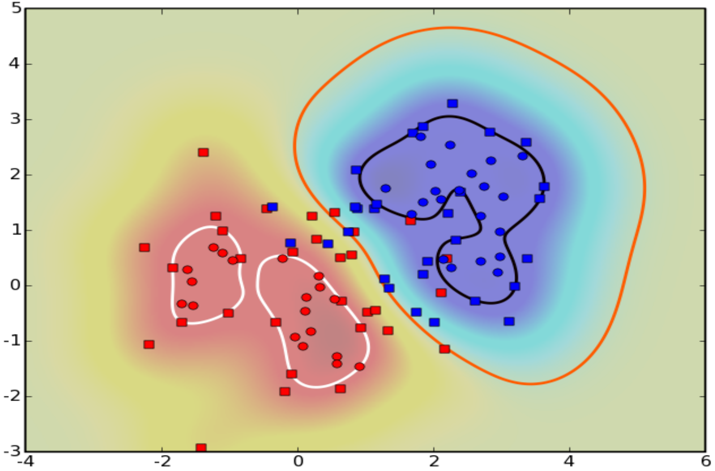
Gaussian RBF with $C = 0.1$



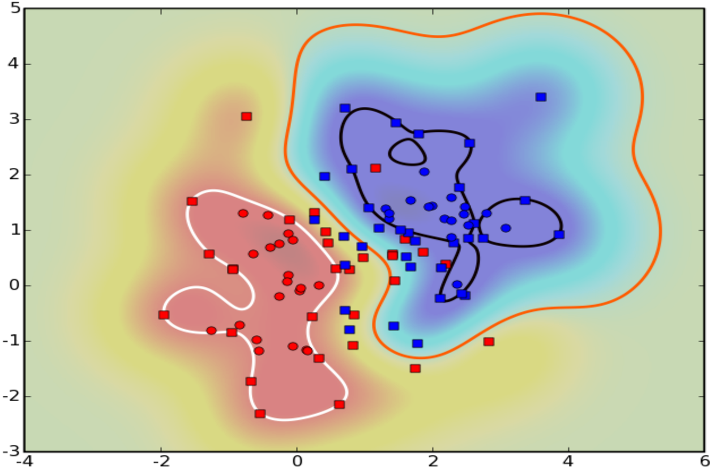
Gaussian RBF with $C = 0.2$



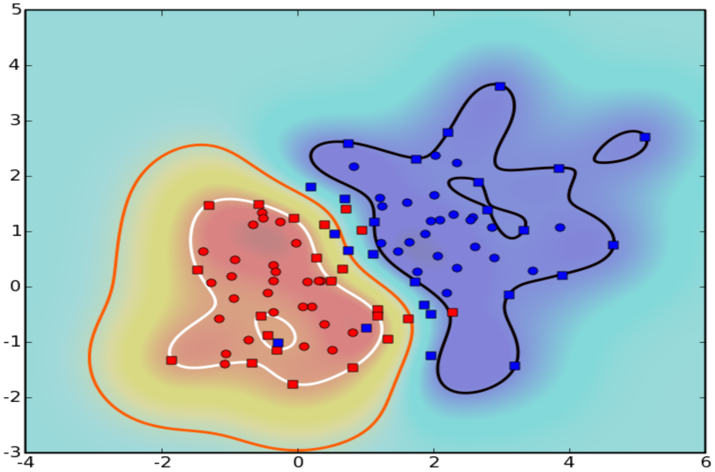
Gaussian RBF with $C = 0.4$



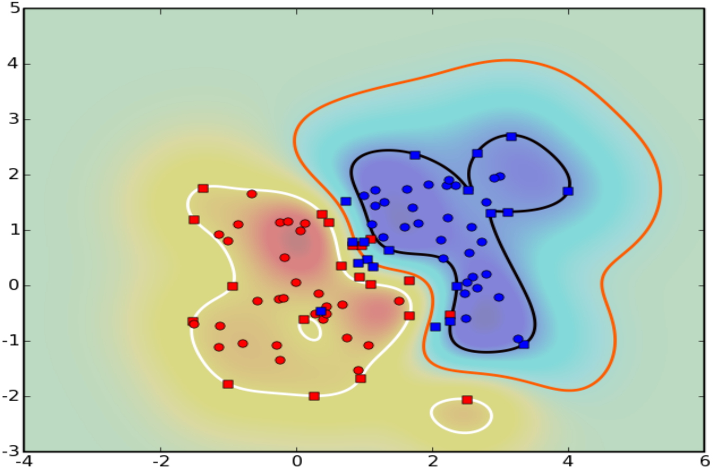
Gaussian RBF with $C = 0.8$



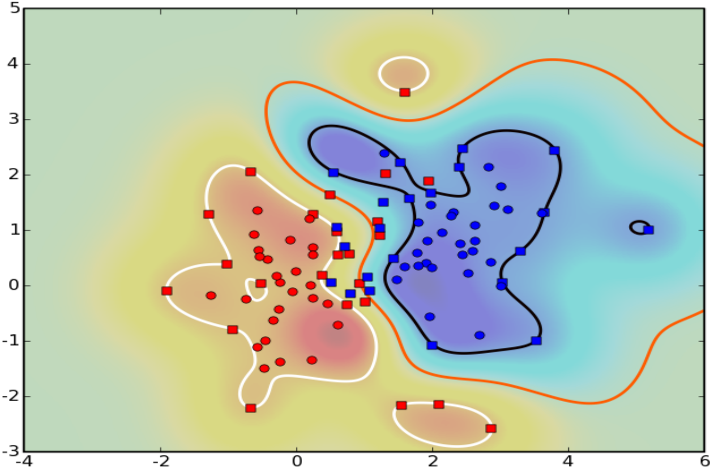
Gaussian RBF with $C = 1.6$



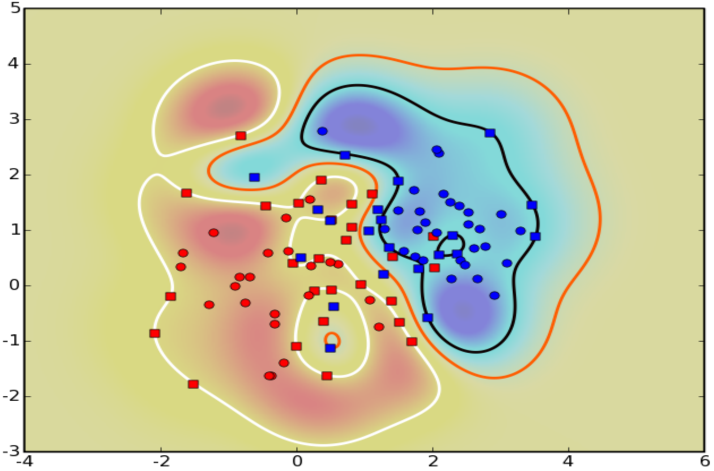
Gaussian RBF with $C = 3.2$



Gaussian RBF with $C = 6.4$



Gaussian RBF with $C = 12.8$



Summary

Support Vector Machine

- Problem definition
- Geometrical picture
- Optimization problem

Optimization Problem

- Hard margin
- Convexity
- Dual problem
- Soft margin problem

Soft Margin SVMs

C-SVM [15]: for $C > 0$, minimize

$$\tau(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i$$

subject to $y_i \cdot (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$ (margin $2/\|\mathbf{w}\|$)

ν-SVM [55]: for $0 \leq \nu < 1$, minimize

$$\tau(\mathbf{w}, \boldsymbol{\xi}, \rho) = \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{m} \sum_{i} \xi_i$$

subject to $y_i \cdot (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq \rho - \xi_i, \quad \xi_i \geq 0$ (margin $2\rho/\|\mathbf{w}\|$)

The ν-Property

SVs: $\alpha_i > 0$
"margin errors:" $\xi_i > 0$

KKT-Conditions $\Longrightarrow$

- All margin errors are SVs.
- Not all SVs need to be margin errors.
  Those which are *not* lie exactly on the edge of the margin.

Proposition:
1. *fraction of Margin Errors* $\leq \nu \leq$ *fraction of SVs.*
2. *asymptotically:* ... $= \nu =$ ...

Duals, Using Kernels

C-SVM dual: maximize

$$W(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to $0 \leq \alpha_i \leq C, \quad \sum_i \alpha_i y_i = 0.$

ν-SVM dual: maximize

$$W(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to $0 \leq \alpha_i \leq \frac{1}{m}, \quad \sum_i \alpha_i y_i = 0, \quad \sum_i \alpha_i \geq \nu$

In both cases: *decision function*:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b\right)$$

Connection between ν-SVC and C-SVC

Proposition. If ν-SV classification leads to $\rho > 0$, then C-SV classification, with C set a priori to $1/\rho$, leads to the same decision function.

Proof. Minimize the primal target, then fix ρ, and minimize only over the remaining variables: nothing will change. Hence the obtained solution $\mathbf{w}_0, b_0, \boldsymbol{\xi}_0$ minimizes the primal problem of C-SVC, for $C = 1$, subject to

$$y_i \cdot (\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq \rho - \xi_i.$$

To recover the constraint

$$y_i \cdot (\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 - \xi_i,$$

rescale to the set of variables $\mathbf{w}' = \mathbf{w}/\rho, b' = b/\rho, \boldsymbol{\xi}' = \boldsymbol{\xi}/\rho$. This leaves us, up to a constant scaling factor ρ^2, with the C-SV target with $C = 1/\rho$.