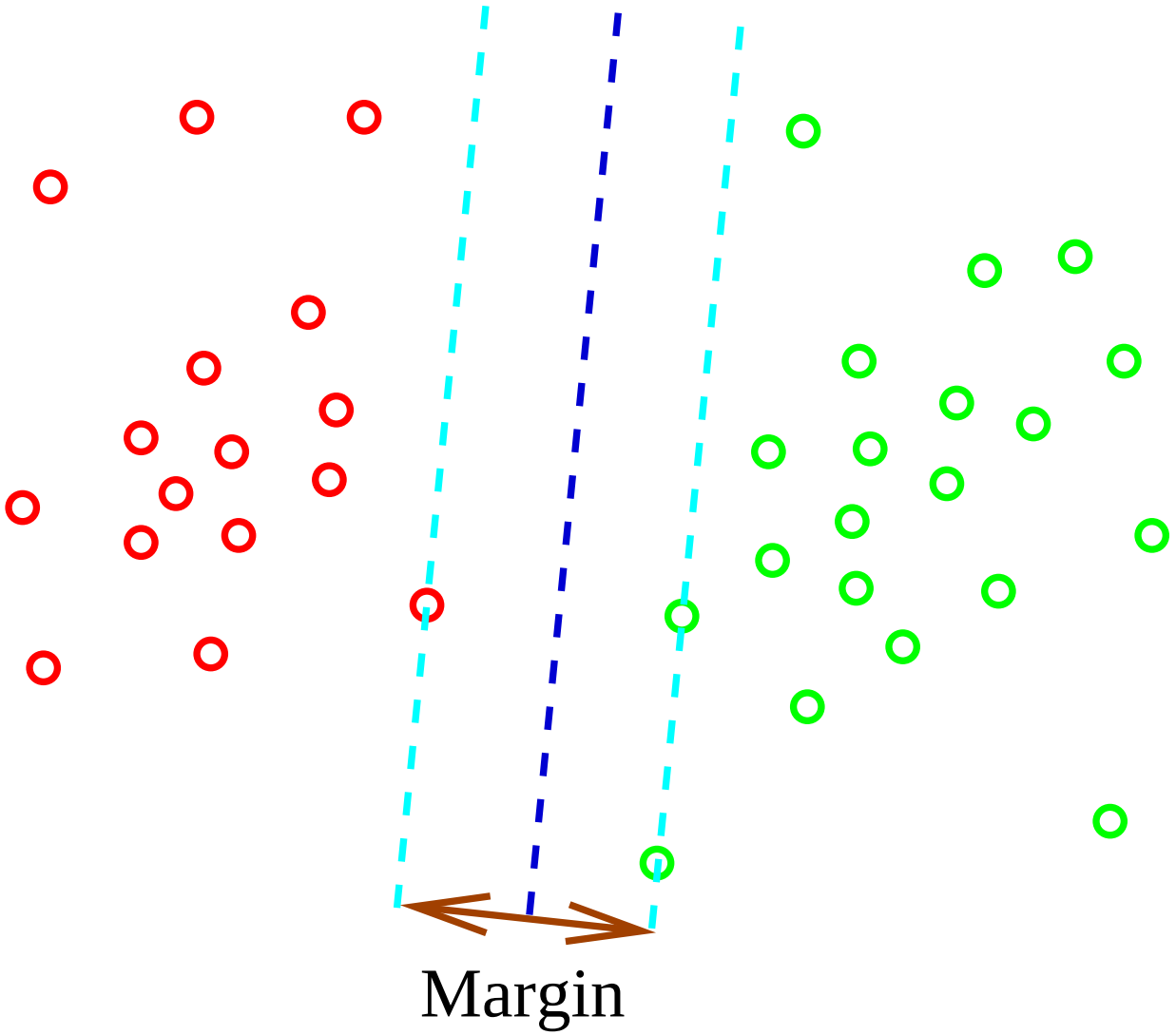
Statistical Machine Learning

UoC Stats 37700, Winter quarter

Lecture 6: Support Vector Machines.

The Support Vector Machine

- The SVM is “yet another” linear classifier.
- The main underlying geometric idea is to find a separation of the data that achieves a “large margin”.
- The intuitive idea is that if training points of the two classes are separated with a larger margin, the resulting classifier is less prone to generalization error.
- We will see later what learning theory has to say about the SVM.


Margin

SVM: primal formulation

- Assume that the training points are linearly separable.
- For a hyperplane H given by (w, b), remember that the distance to H is given by

$$d(x, H) = \frac{|w \cdot x - b|}{\|w\|} .$$

- Hence, the goal is to find

$$\underset{w,b}{\text{Arg Max}} \frac{\min_i |w \cdot X_i - b|}{\|w\|}$$

under the constraints (assuming $Y \in \{-1, 1\}$):

$$\forall 1 \leq i \leq n , (w \cdot X_i - b) Y_i > 0 .$$

Reformulation

- Once again, the normalization of the parameters (w, b) is arbitrary.
- For any feasible (i.e. data-separating) (w, b) we can choose a normalization such that

$$\min_i (w \cdot X_i - b) y_i = 1 \,.$$

- The previous optimization problem is then conveniently reformulated as

$$\underset{w,b}{\text{Arg Min}} \, \|w\|^2$$

under the constraints:

$$\forall 1 \leq i \leq n \,, (w \cdot X_i - b) Y_i \geq 1 \,.$$

- Note that with this normalization the geometric margin is exactly $\|w\|^{-1}$.

A motivation for large margin classification

Theorem

Consider n points $S_X = (X_1, \ldots, X_n)$ *and consider the set* $\mathcal{F}$ *of linear separators* $(w, 0)$ *passing through the origin (i.e.* $b = 0$*) and such that the distance of* S_X *to the corresponding hyperplane (margin) is lower bounded by* Λ*.*
If $\mathcal{F}$ *shatters* S_X*, then* $n \leq \frac{R^2}{\Lambda^2}$*, where* $R = \max_i \|X_i\|$ *.*

- Observe that this result is *dimension independent*!
- Note that this result does not exactly fit into the VC theory (why?).
- It can however result in a rigorous result in the “transductive” case (why?).
- It provides at least a first justification – at this point still non totally rigorous – that for large margin classification methods, the “margin” is the relevant criterion of complexity rather than the dimension. Hence we should not be afraid of applying them to very-high dimensional data.

The Karush-Kuhn-Tucker theorem

Theorem

Consider the optimization problem: $\min_{x\in\Omega} f(x)$,
under the constraints:

$$\forall 1 \leq i \leq k\,, g_i(x) \leq 0\,; \qquad \forall 1 \leq j \leq k'\,, h_i(x) = 0\,,$$

where $f, (g_i), (h_i)$ *are convex functions and* Ω *is a convex set. Define the Lagrangian*

$$L(x, \alpha, \beta) = f(x) + \sum_i \alpha_i g_i(x) + \sum_j \beta_j h_j(x)\,;$$

and pose

$$\theta(\alpha, \beta) = \inf_x L(x, \alpha, \beta)\,.$$

Then the solution of the initial problem is the same as the solution of: $\sup_{\alpha,\beta} \theta(\alpha, \beta)$ *under the constraints* $\forall 1 \leq i \leq k \ : \alpha_i \geq 0$.

KKT conditions

Theorem (Continued)

Furthermore, necessary and sufficient conditions for the existence of (x^*, α^*, β^*) *realizing the above problems are:*

$$\frac{\partial L(x^*, \alpha^*, \beta^*)}{\partial x} = 0\,;$$

$$\frac{\partial L(x^*, \alpha^*, \beta^*)}{\partial \beta} = 0\,;$$

$$\frac{\partial L(x^*, \alpha^*, \beta^*)}{\partial \alpha} \geq 0 \qquad (\textit{ i.e. : } g_i(x^*) \leq 0;)$$

$$\alpha_i^* \geq 0$$

$$\alpha_i^* g_i(x^*) = 0\,.$$

Note: the last contraint tells us that the Lagrange coefficients α_i^* are non-zero only for the “active” constraints at the solution.

The Dual problem for the SVM

- The KKT theorem applied to the SVM problem gives

$$\theta(\lambda) = -\frac{1}{2}\sum_{i,j} \lambda_i \lambda_j Y_i Y_j (X_i \cdot X_j) + \sum_i \lambda_i$$

to maximize under the constraints:

$$\forall 1 \leq i \leq n \;\; \lambda_i \geq 0 \text{ and } \sum_i \lambda_i Y_i = 0\,.$$

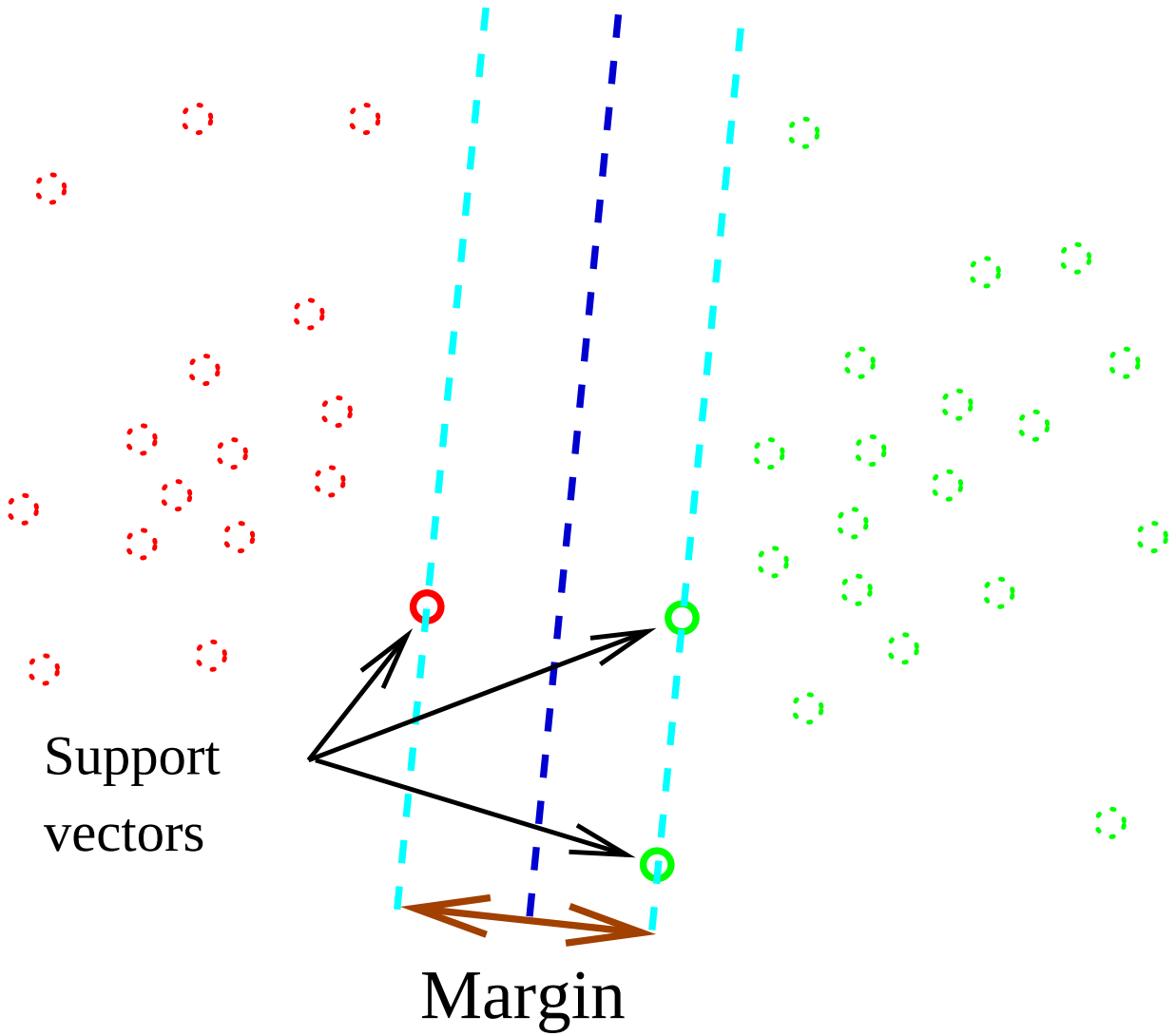
- Furthermore, the first optimality condition gives

$$w^* = \sum_i \lambda_i Y_i X_i\,.$$

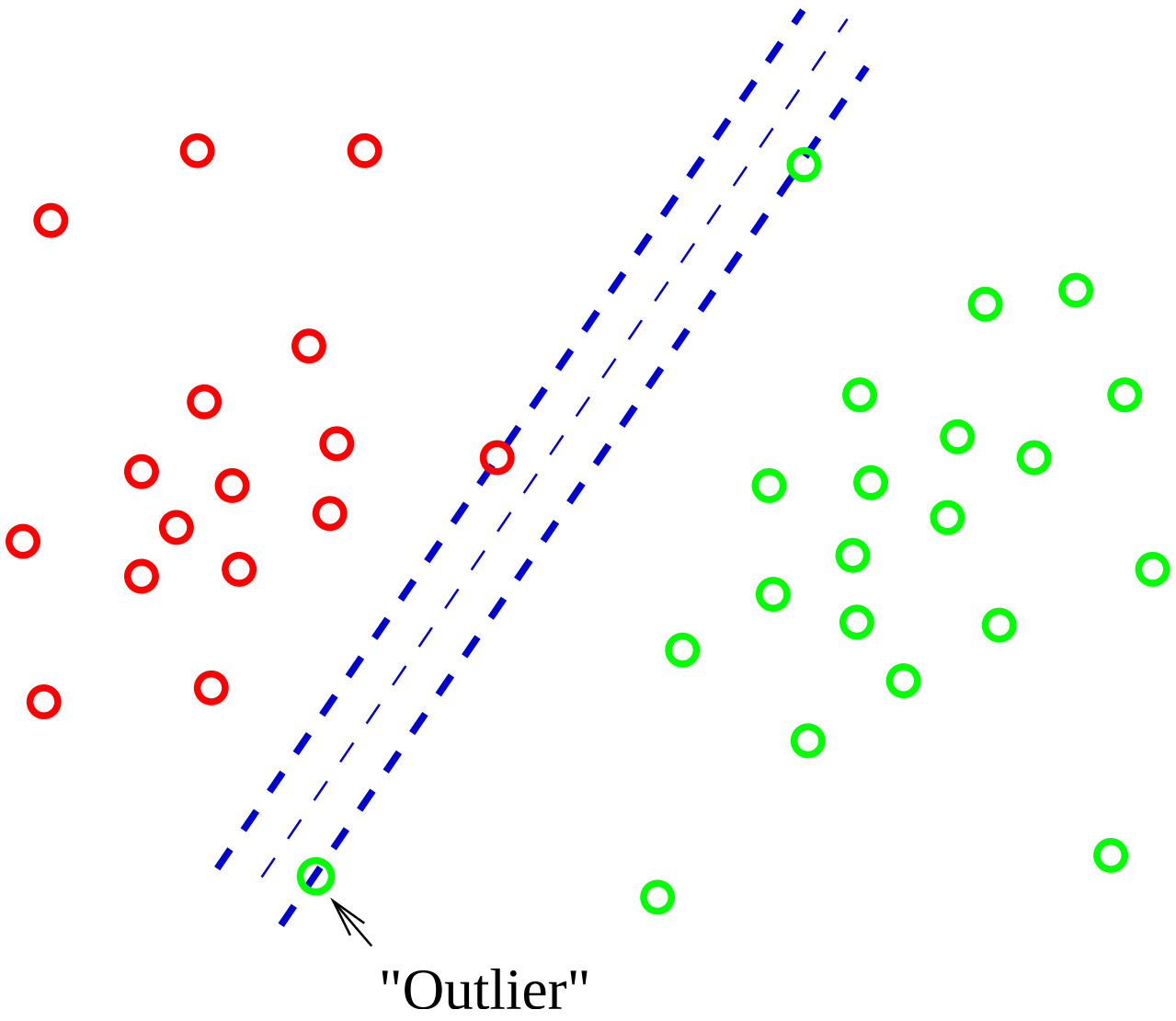
Some important observations:

- The optimizing w^* is a linear combination of the training points (positive examples have nonnegative weights and vice-versa: compare perceptron).
- The dual optimization problem is entirely determined by the knowledge of the dot products $X_i \cdot X_j$.
- (This is relevant if the dimension of the ambient space is larger than the number of points)
- From the KT conditions, only the examples that are correspond to an active constraints, i.e. are exactly “on the margin”, have non-zero cofficients in the above examples: support vectors .
- Any active constraint allows to compute the optimal offset b^* .

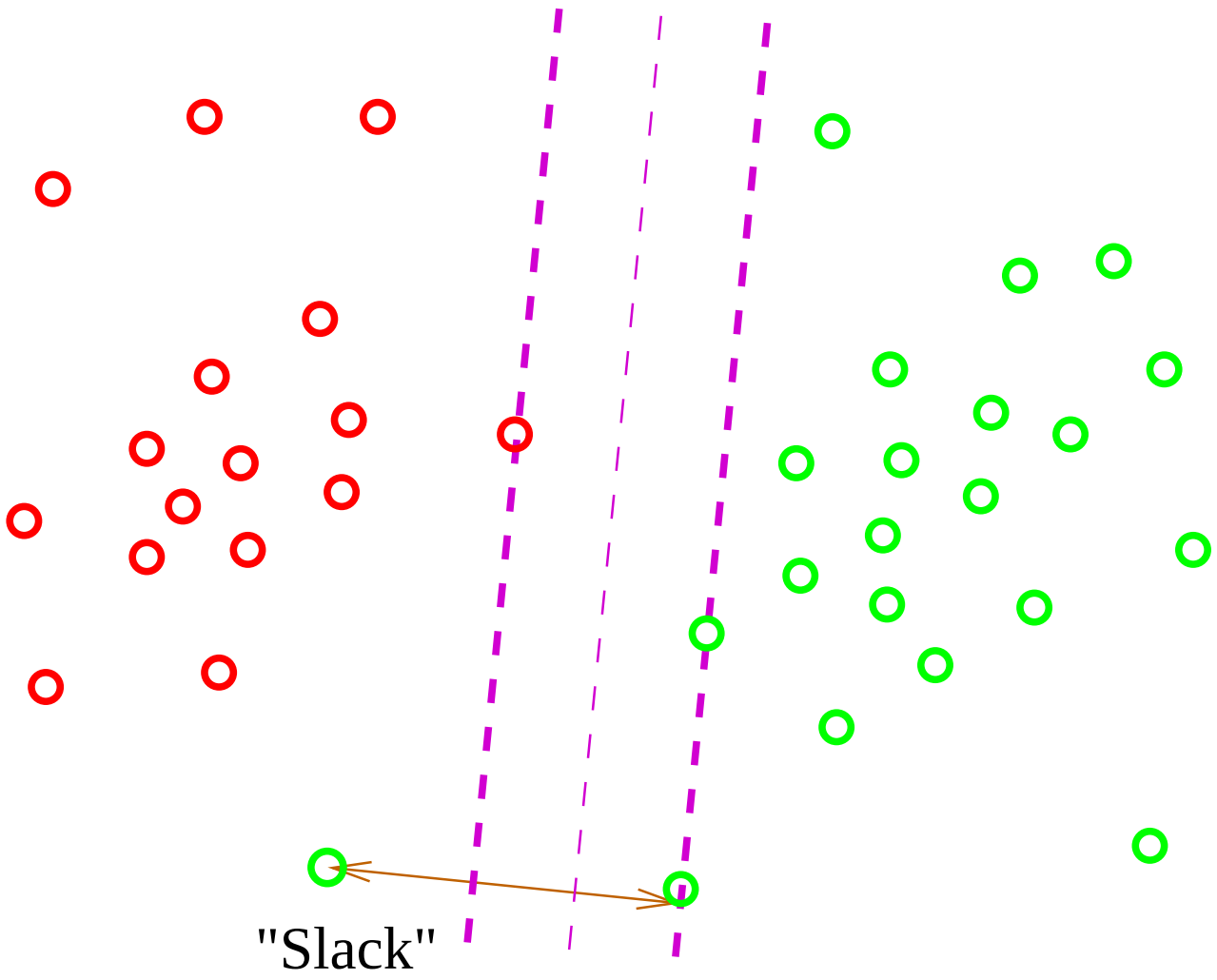
Support vectors



Problems with the hard margin SVM



Problems with the hard margin SVM



The soft margin SVM

- Relax the initial constraints to

$$\forall 1 \leq i \leq n \; : Y_i \left(w \cdot X_i + b \right) \geq 1 - \xi_i \qquad \xi_i \geq 0 \, .$$

  ξ_i are called the "slack variables" .

- Include the slack variables as a penalization term with a factor $C \geq 0$ in the objective function, hence becoming

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \, ,$$

  under the above constraints.

- The multiplier C can be made class-dependent for example to compensate for unbalancedness of the classes' prior distribution.

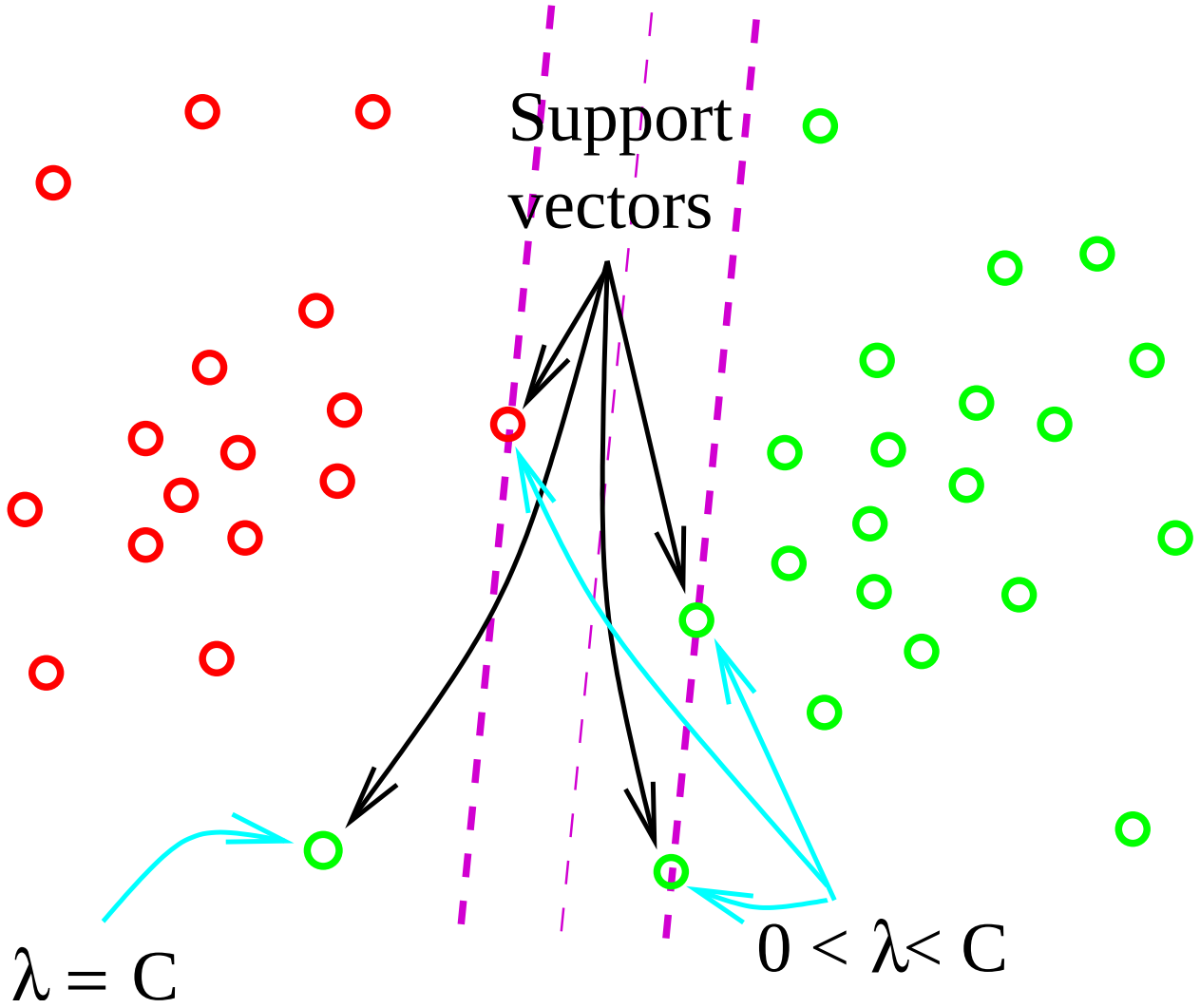
The dual problem for the soft SVM

- The KKT theorem applied to the soft margin SVM problem gives

$$\theta(\lambda) = -\frac{1}{2}\sum_{i,j}\lambda_i\lambda_j Y_i Y_j (X_i \cdot X_j) + \sum_i \lambda_i$$

  (same as hard margin SVM!) to maximize under the constraints: $\forall a \leq i \leq n \qquad 0 \leq \lambda_i \leq C$ and $\sum_i \lambda_i Y_i = 0$.
- Again, the optimal w^* is a combination of training examples, and we have a notion of support vectors.

Support vectors for the soft SVM



The soft SVM is a regularized ERM procedure

- Going back to the primal problem, we can refomulate it without constraints as

$$\min_{w,b} \sum_i \left(1 - Y_i \left(w \cdot X_i + b\right)\right)_+ + \frac{1}{C} \|w\|^2 \ .$$

- Hence, the soft SVM can be equally seen as a ridge regression (of sorts) with an ad-hoc loss function

$$\ell(f, X, Y) = \left(1 - Y_i f(X_i)\right)_+$$

  where f belongs to the set of linear classifiers.
- Having a class-dependent coefficient C will correspond to a reweighting of the examples depending on the class.
- This "hinge loss" is a convex upper bound for the usual 0-1 misclassification loss function.
- Using other losses or regularizers is also possible. . .

The representer theorem

- Consider an optimization problem of the form

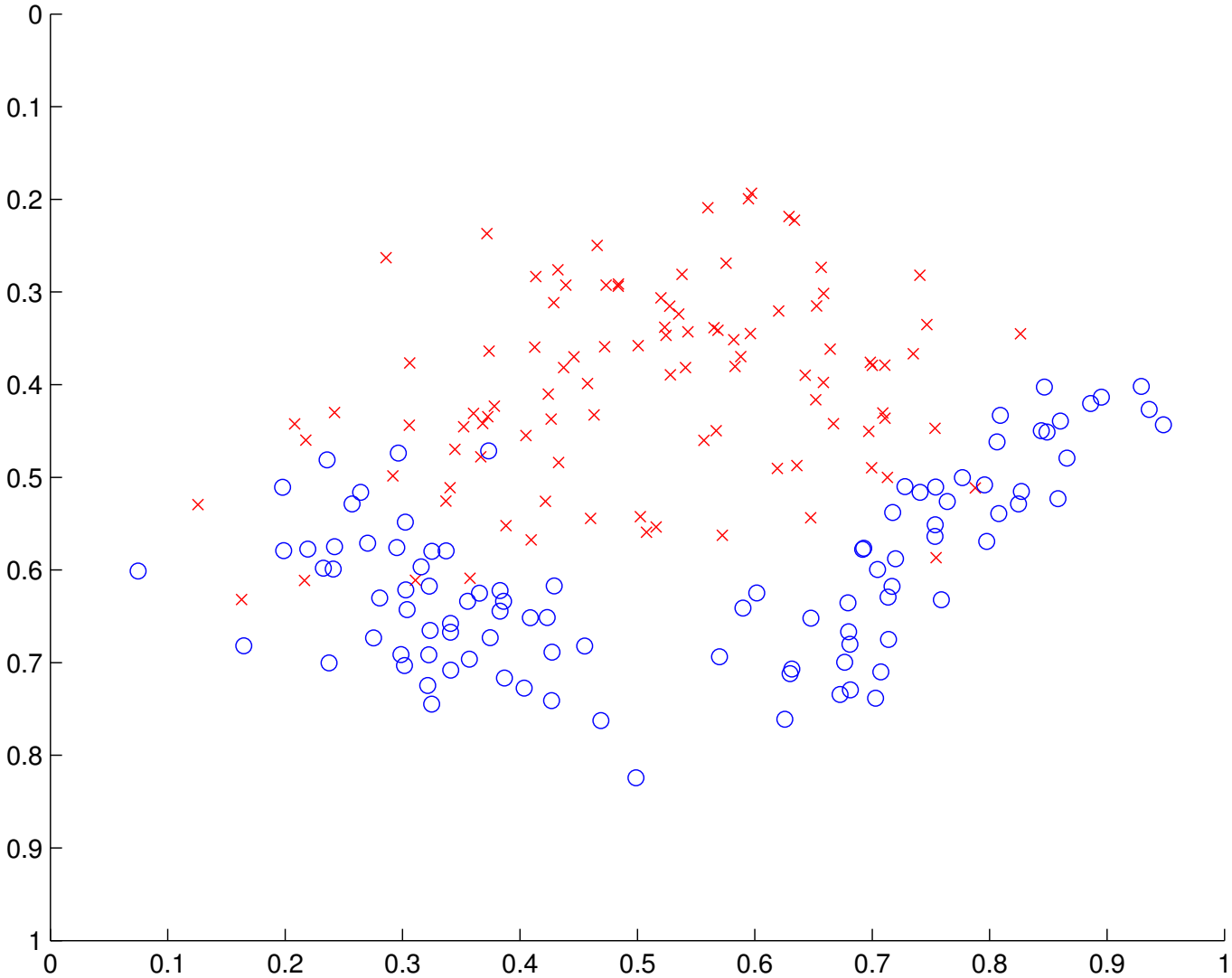
$$\underset{w,b}{\mathrm{Arg\,Min}}\, \Psi((X_i \cdot w)_{1\leq i\leq n}, b, \|w\|)\,,$$

  where Ψ is a function nondecreasing in its last variable.
- Then the solution w^* is a linear combination of the X_i's,

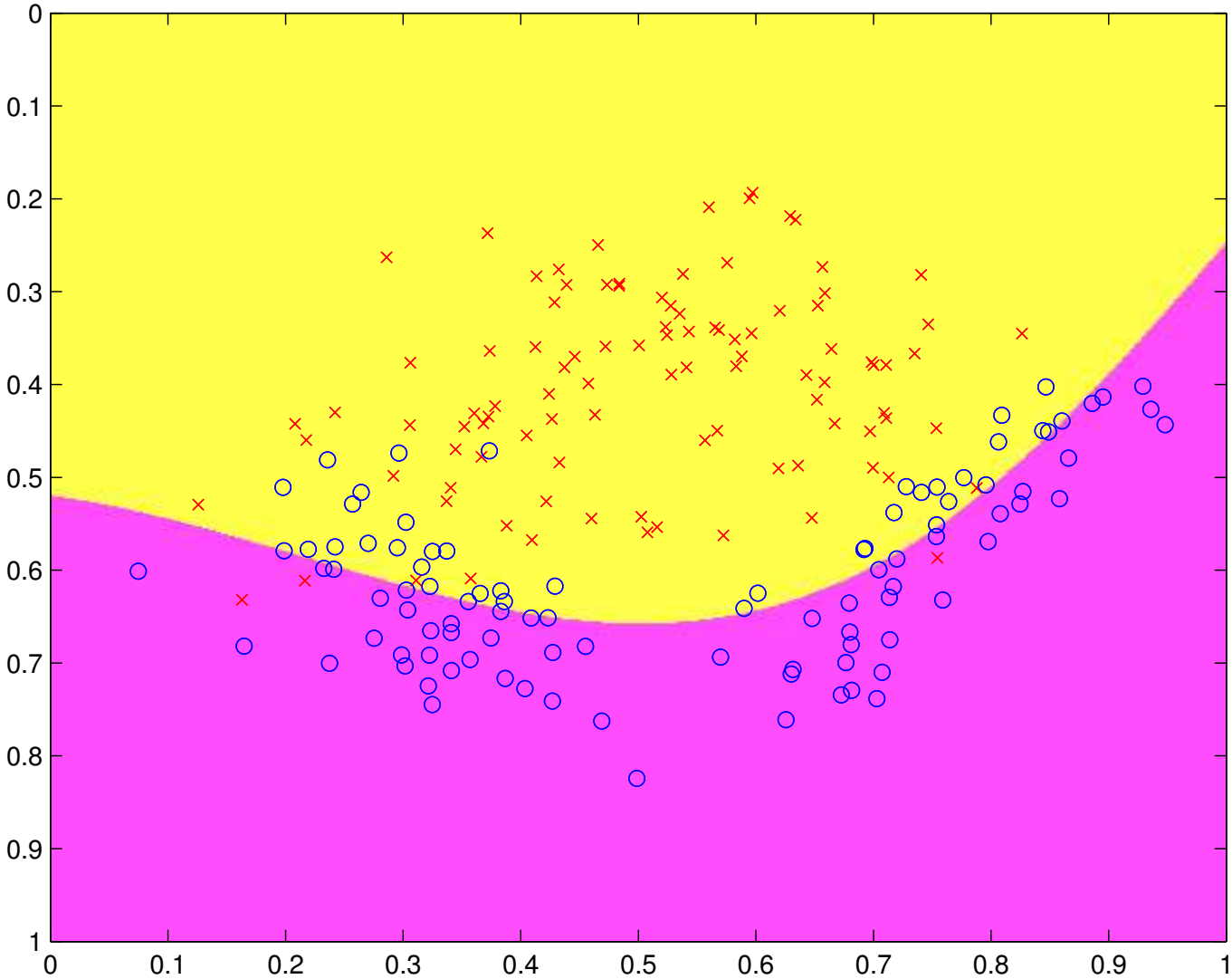
$$w^* = \sum_i a_i X_i\,.$$

- Again, this is relevant in the case where the ambient space is of dimension larger than the number of examples (and posssibly infinite dimension).

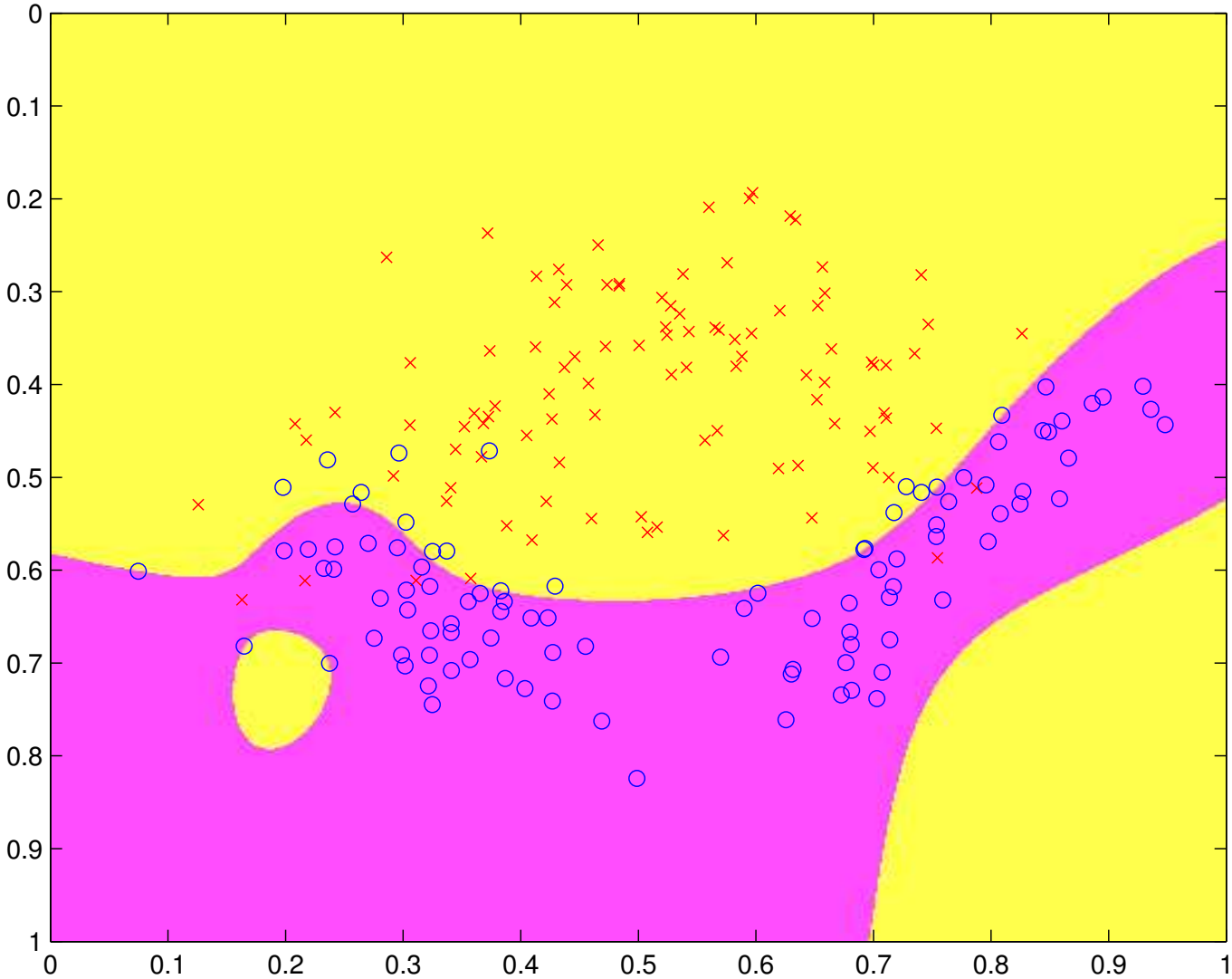
Role of the the regularizing constant



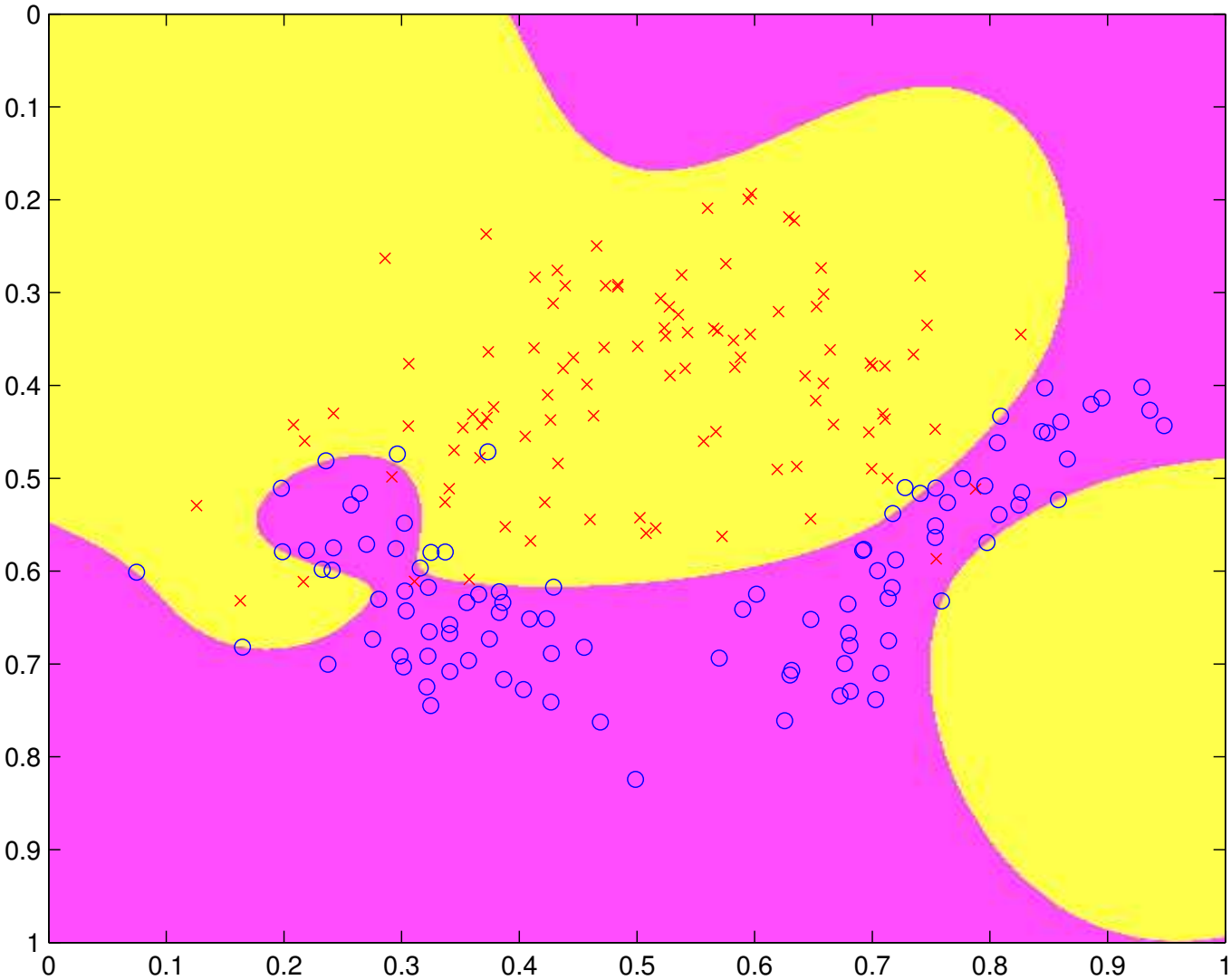
Role of the the regularizing constant



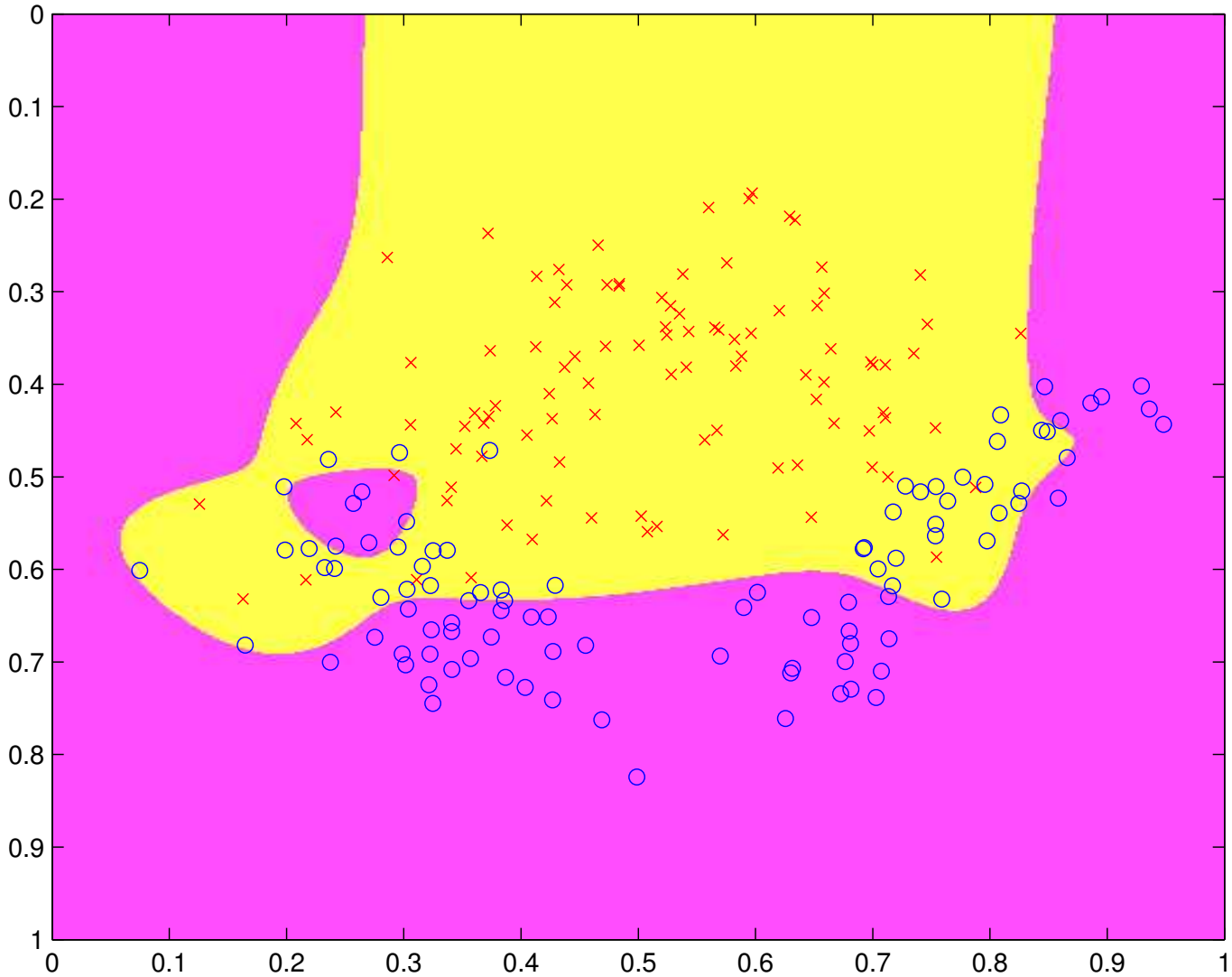
Role of the the regularizing constant



Role of the the regularizing constant



Role of the the regularizing constant



Support vectors and leave-one-out error

- Consider a non-support vector training point (X_i, Y_i), i.e. such that the corresponding coefficient λ_i in the expansion of w^* is zero.
- Consider the training sample S^{-i} with this point removed; then the SVM solution of S^{-i} coincide with the solution on S.
- Since a non-support vector is correctly classified, we deduce the inequality on the leave-one out error

$$LOO(S, SVM) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\widehat{f}^{-i}(X_i) \neq Y_i\} \leq \frac{\#SV.}{n}$$

- This is a weak justification that if the number of SV is small, we expect a better generalization error.

Implementation issues

- The optimization can be solved using various iterative procedures.
- The convergence is generally monitored either by looking at the KKT constraint conditions for the solution:

$$Y_i f(X_i) \begin{cases} \geq 1 & \text{if } \lambda_i = 0 \ ; \\ = 1 & \text{if} 0 < \lambda_i < C \ ; \\ \leq 1 & \text{if } \lambda_i = C \ , \end{cases}$$

  and stop when they are satisfied up to some error ϵ .
- Alternatively, control the "duality gap":

$$f(w, \xi) - \theta(\lambda) = C \sum_i \xi_i + \sum_i \lambda_i - 2\theta(\lambda) \, ,$$

$$\text{where} \quad \xi_i = (1 - Y_i (\sum_j Y_j \alpha_j X_i \cdot X_j + b))_+ \, ,$$

  and b is picked such that $Y_i f(X_i) = 1$ for some arbitrary i_0 s.t. $0 < \lambda_{i_0} < C$.

Naive implementation : gradient ascent

- Consider the case where we “replace” the free offset b by adding a constant coordinate to all samples (why is that actually not equivalent?)
- A very simple solution is to perform naive gradient ascent on the dual problem:

$$\frac{\partial \theta}{\partial \lambda_i} = 1 - Y_i \sum_j \lambda_j Y_j (X_i \cdot X_j) \, .$$

- One update step is then

$$\lambda_i \leftarrow \left[\lambda_i + \eta \frac{\partial \theta}{\partial \lambda_i} \right]_{[0,C]} \, .$$

- Generally, this is actually updated one coefficient at a time (compare perceptron), and with the choice of η_i corresponding to the optimal line search, i.e. $\eta_i = (X_i \cdot X_i)^{-1}$.
- Various heuristics are used to select the order of the points, based e.g. on the violations of the KT conditions and the current estimate of the SV set.

Platt's SMO

- When considering the problem with free offset b the constraint $\sum_i \lambda_i Y_i = 0$ prevents updating only one single coefficient at a time.
- A simple idea is to update 2 coefficients simultaneously.
- Calculations of the optimum (line search) lead to the update:

$$\Delta = \frac{(f(X_1) - f(X_2) - Y_1 + Y_2)}{\|X_1 - X_2\|^2};$$
$$\lambda_2^{new} = \lambda_2 + Y_2\Delta;$$
$$\lambda_1^{new} = \lambda_1 - Y_1\Delta,$$

  (plus additional clipping constraints)
- Again, choice of the points to update follow various clever heuristics, e.g. based on a "working set".

The ν-SVM

- The choice of relaxation introducing the slack variables is somewhat arbitrary. An alternative way (maybe closer to the initial geometric view) to define an optimization problem is to consider explicitly the "margin":

$$\underset{w,b,\gamma,\xi}{\text{Arg Min}} -\gamma + C\sum_i \xi_i$$

under the constraints:

$$\|w\| = 1;\ \ \forall 1 \leq i \leq n : \xi_i \geq 0 \text{ and } Y_i(w \cdot X_i + b) \geq \gamma - \xi_i\,.$$

- The dual optimization can be shown to be

$$\underset{\lambda}{\text{Arg Max}} -\sum_{i,j} \lambda_i \lambda_j Y_i Y_j (X_i \cdot X_j)\,,$$

under constraints

$$\forall 0 \leq i \leq n\ :\ 0 \leq \lambda_i \leq C \text{ and } \sum_i \lambda_i Y_i = 0 \text{ and } \sum_i \lambda_i = 1\,.$$

Properties of the ν-SVM

- Note that the two formulations are not equivalent. In particular, for the ν-SVM we must have $C \geq n^{-1}$ otherwise the problem is infeasible .
- Putting $C = \frac{1}{\nu n}$ with $\nu \in (0, 1]$, in can be seen (through the constraints) that ν is an upper bound on the proportion of support vectors of type I and a lower bound on the total number of support vectors. This formulation is hence in a sense more “interpretable” .
- Finally, as ν varies in $(0, 1]$ the set of attained solutions coincides with the set of solutions in the standard formulation when C varies in $(0, \infty]$.