Self-Organizing Maps (SOM)

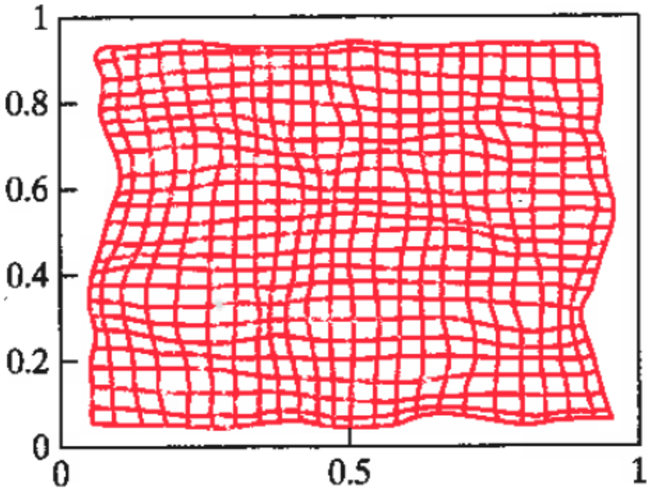
The 3 aspects of training in SOM algorithms:

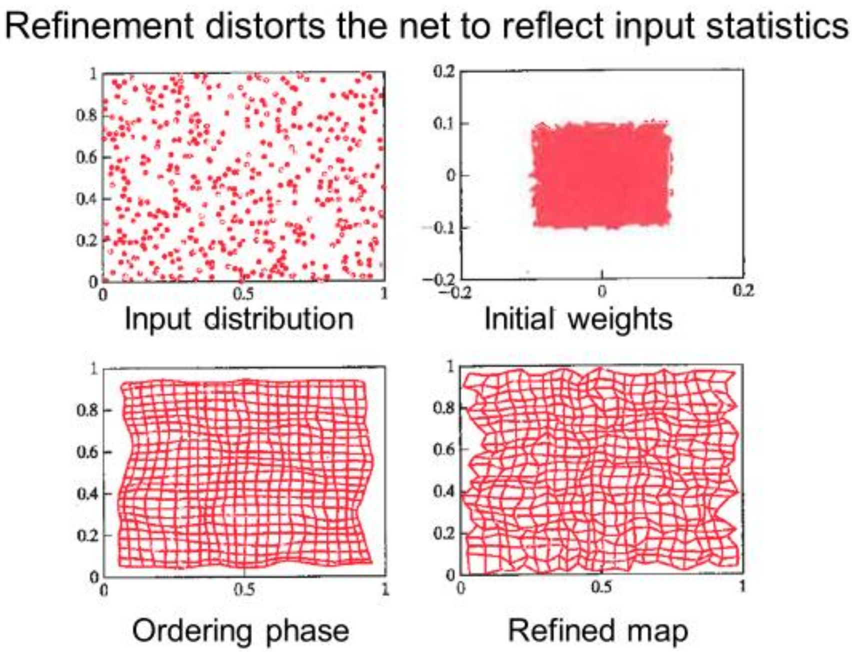
1. Output nodes compete for activation based on a discriminant function.
2. Winning node, with largest value of discriminant, becomes the center of a cooperative neighborhood.
3. Neighborhood adapts to an input pattern because cooperation increases the susceptible to large values of the discriminant function.

Elastic net over input space.

Each node is associated with a compressed representation of input space.

Phase of SOM convergence illustrated: Ordering phase




Refinement distorts the net to reflect input statistics

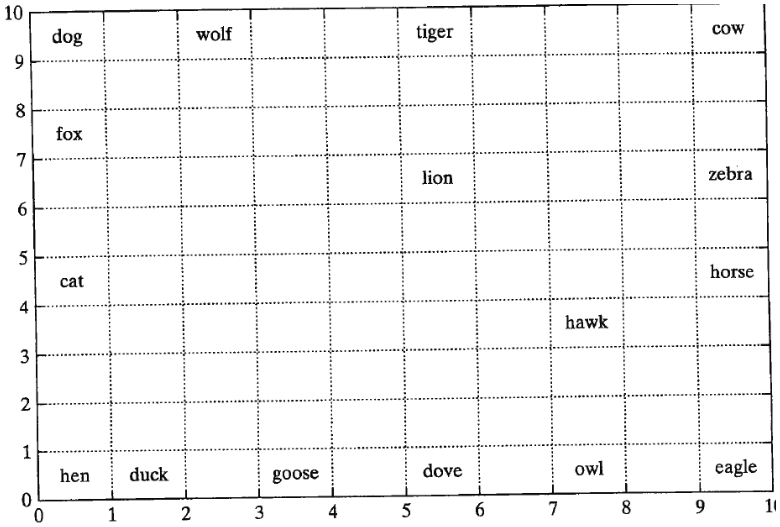
Input distribution

Initial weights

Ordering phase

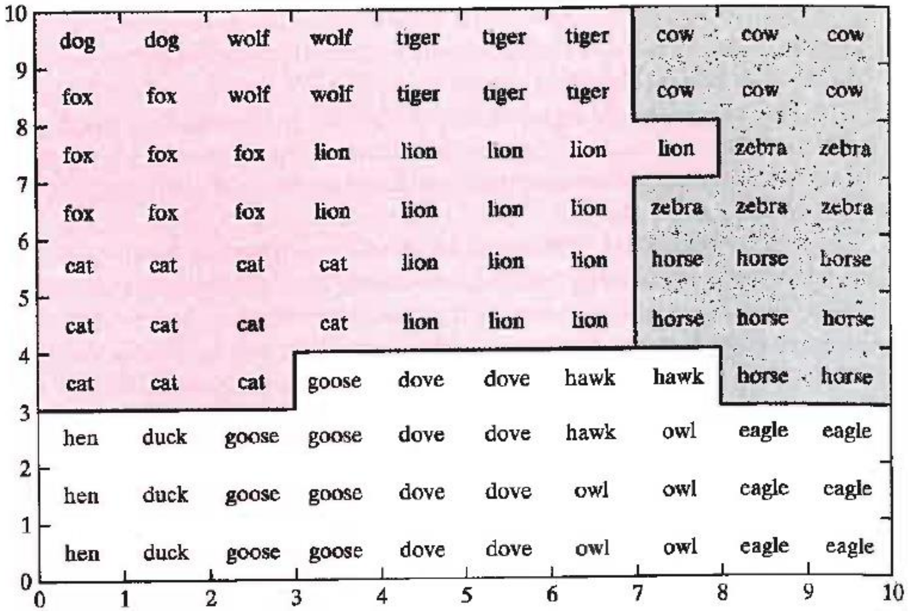
Refined map

Contextual map: Labels output nodes based on responses of the lattice to "test patterns"; labeled instances, not used in training. Labeled node has strongest response to the indicated animal type. Topological ordering of data.



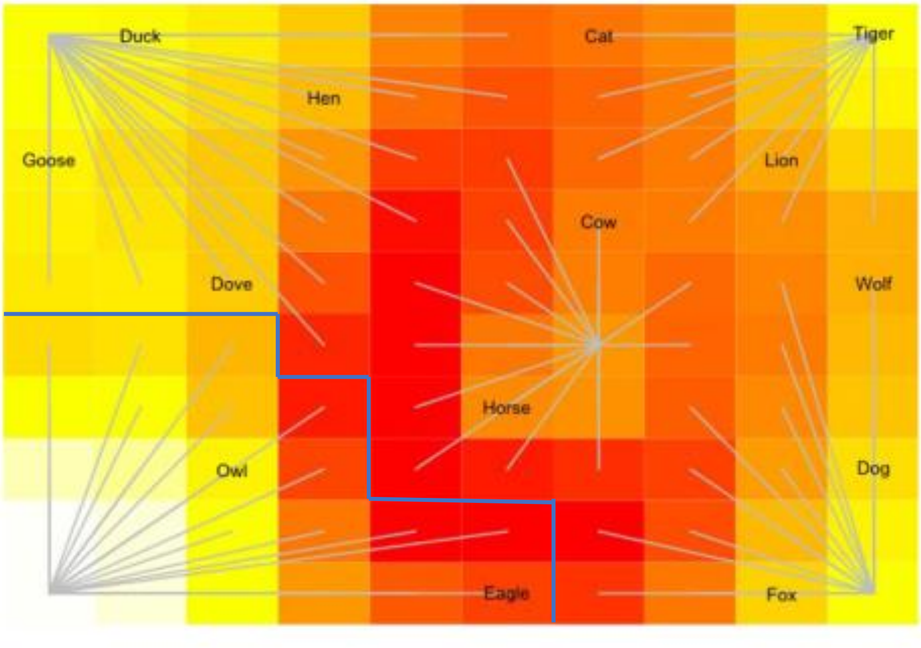
dog		wolf			tiger				cow
fox									
					lion				zebra
cat									horse
							hawk		
hen	duck		goose		dove		owl		eagle

Semantic map: Lattice sites label by animal type inducing the strongest response at that node. Similar types of data are clustered.



dog	dog	wolf	wolf	tiger	tiger	tiger	cow	cow	cow
fox	fox	wolf	wolf	tiger	tiger	tiger	cow	cow	cow
fox	fox	fox	lion	lion	lion	lion	lion	zebra	zebra
fox	fox	fox	lion	lion	lion	lion	zebra	zebra	zebra
cat	cat	cat	cat	lion	lion	lion	horse	horse	horse
cat	cat	cat	cat	lion	lion	lion	horse	horse	horse
cat	cat	cat	goose	dove	dove	hawk	hawk	horse	horse
hen	duck	goose	goose	dove	dove	hawk	owl	eagle	eagle
hen	duck	goose	goose	dove	dove	owl	owl	eagle	eagle
hen	duck	goose	goose	dove	dove	owl	owl	eagle	eagle

U-mat with stars: Use the stars to draw a boundary on the cluster that contains Owl and Eagle.



Support Vector Machines (SVM)

SVM is an example of constrained optimization.
Simplex method can be used when both objective function and constraints are linear.

Why can't the simplex method be used on SVMs?

> In SVM, we find the optimal hyperplane (aka decision boundary) by minimizing a quadratic objective function subject to linear constraints.

How do we know that the solution to the SVM optimization problem is unique?

> In the SVM optimization problem, the quadratic objective function is convex and bounded from below.

Active set in the quadratic programming problem

The active set determines which constraints will influence the final result of optimization.

What is the active set in the SVM quadratic programming problem?

> In the SVM quadratic programming problem, support vectors define the active set: attribute vectors with constraints that have non-zero Lagrange multipliers.

Solving constrained optimization by Lagrange multipliers

NOT EXPECTED TO BE ON QUIZ

find the stationary point of $f(x_1, x_2) = 1 - x_1^2 – x_2^2$

subject to the constraint $g(x_1, x_2) = x_1 + x_2 = 1$

Form the Lagrangian

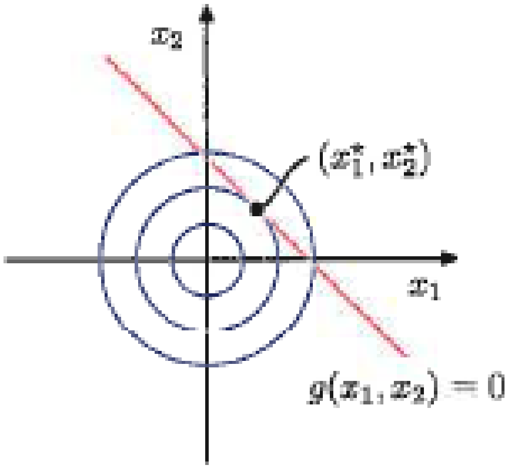
$L(\mathbf{x}, \lambda) = f(x_1, x_2) + \lambda(g(x_1, x_2) - c)$

$L(\mathbf{x}, \lambda) = 1-x_1^2-x_2^2 +\lambda(x_1+x_2-1)$

Set the partial derivatives of L with respect x_1, x_2, and λ equal to zero

$L(\mathbf{x}, \lambda) = 1-x_1^2-x_2^2 +\lambda(x_1+x_2-1)$

$-2x_1 + \lambda = 0$
$-2x_2 + \lambda = 0$
$x_1 + x_2 -1 = 0$
Solve for x_1 and x_2



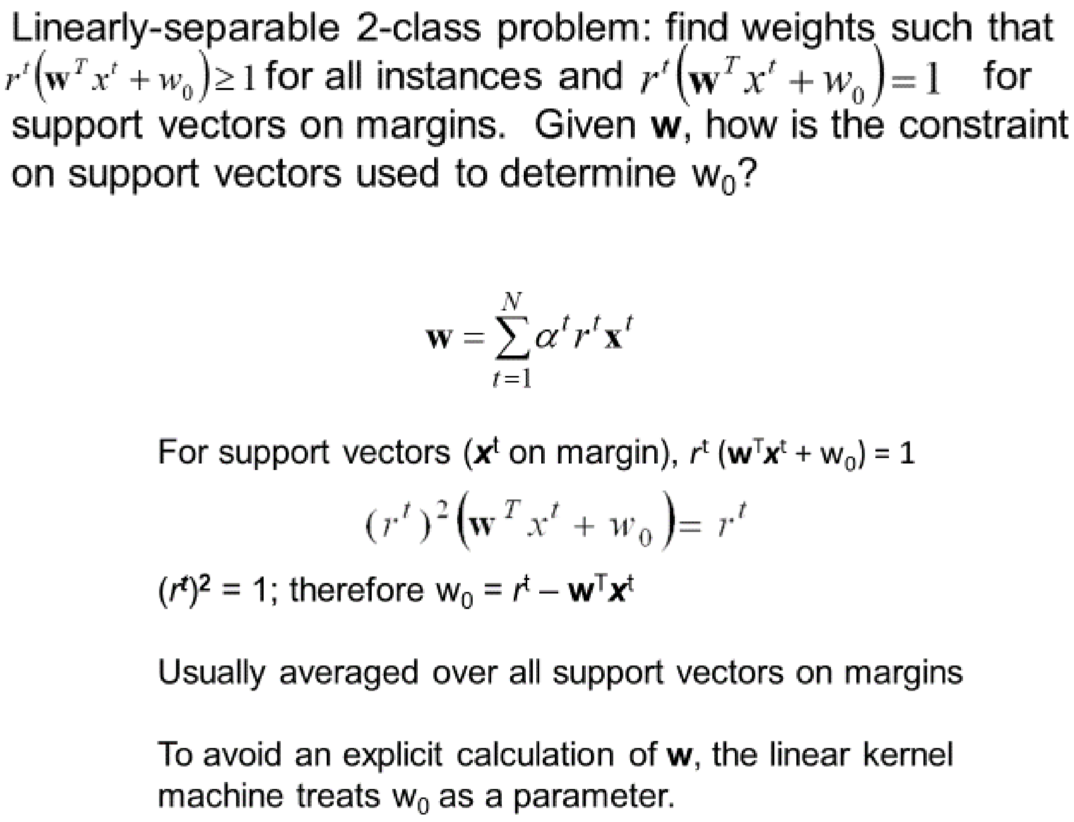
Solution is
$x_1^* = x_2^* = ½$

In this case, not necessary to find λ
λ sometimes called "undetermined multiplier"

In the SVM constrained optimization, we maximize a “dual”; a Lagrangian with primal variables replaced by Lagrange multipliers.

What calculus allows us how to make these replacements?

> In constrained optimization with a Lagrangian, we set the partial derivatives of L with respect to primal variables and Lagrange multipliers to zero to find the stationary point.
>
> This gives equations that allow primal variables to be written as a function of Lagrange multipliers.



Linearly-separable 2-class problem: find weights such that $r^t\left(\mathbf{w}^T x^t + w_0\right) \geq 1$ for all instances and $r^t\left(\mathbf{w}^T x^t + w_0\right) = 1$ for support vectors on margins. Given **w**, how is the constraint on support vectors used to determine w_0?

$$\mathbf{w} = \sum_{t=1}^{N} \alpha^t r^t \mathbf{x}^t$$

For support vectors ($\mathbf{x}^t$ on margin), $r^t\,(\mathbf{w}^T\mathbf{x}^t + w_0) = 1$

$$(r^t)^2\left(\mathbf{w}^T x^t + w_0\right) = r^t$$

$(r^t)^2 = 1$; therefore $w_0 = r^t - \mathbf{w}^T\mathbf{x}^t$

Usually averaged over all support vectors on margins

To avoid an explicit calculation of **w**, the linear kernel machine treats w_0 as a parameter.

In binary classification of linearly separable data,
SVM minimized $||\mathbf{w}||^2$.

Why does this result in maximum margins?

Distance of x^t from the decision boundary is $|g(x^t)|/||w||$.

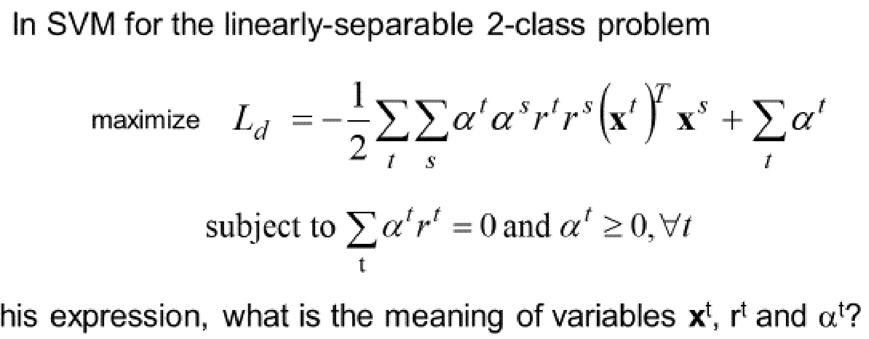
With bipolar labels, $|g(x^t)| = r^t(w^T x^t + w_0)$

Minimizing $||\mathbf{w}||^2$ subject to $r_t(w^T x^t + w_0) \geq 1$ maximizes the distance of all data points from the decision boundary

In constrained optimization, it is often possible to convert the primal problem (i.e. the original form of the optimization problem) to a dual form.

In general, solution of the dual problem provides only an upper bound on the solution of the primal problem (called duality gap). Why does maximizing the dual in SVM give an exact solution?

In SVM we always have a convex optimization problem with linear constraints.
For this type of constrained optimization, the duality gap is zero.



In SVM for the linearly-separable 2-class problem

$$\text{maximize} \quad L_d = -\frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s \left(\mathbf{x}^t\right)^T \mathbf{x}^s + \sum_t \alpha^t$$

$$\text{subject to } \sum_{\text{t}} \alpha^t r^t = 0 \text{ and } \alpha^t \geq 0, \forall t$$

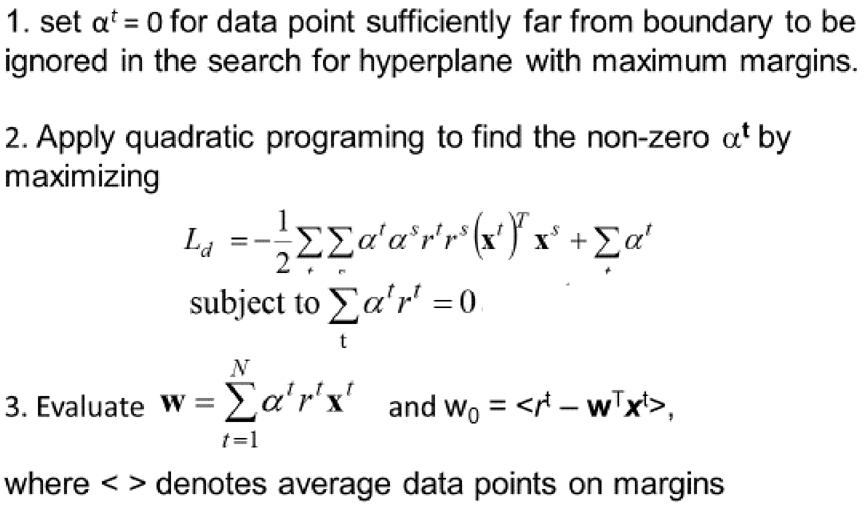
In this expression, what is the meaning of variables $\mathbf{x}^t$, r^t and α^t?

$\boldsymbol{x}^t$: Attribute vector

r^t : Label on that attribute vector

α^t : Lagrange multiplier that is the result of the constraint on that attribute vector

In SVM for the linearly-separable 2-class problem, what are the 3 steps to find the decision boundary with maximum margins?



1. set $\alpha^t = 0$ for data point sufficiently far from boundary to be ignored in the search for hyperplane with maximum margins.

2. Apply quadratic programing to find the non-zero $\alpha^{\mathbf{t}}$ by maximizing

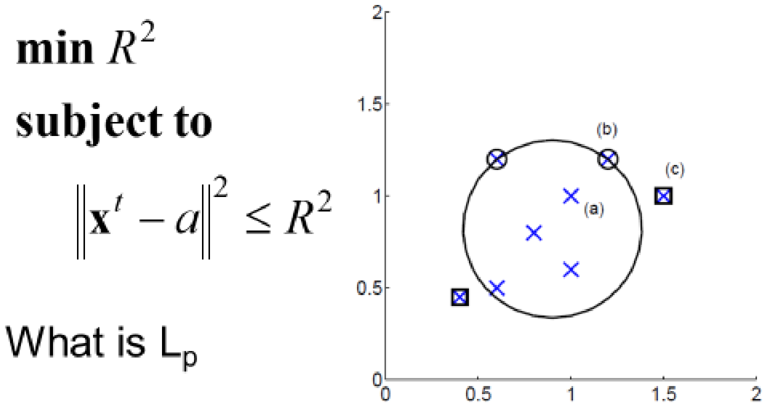
$$L_d = -\frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s \left(\mathbf{x}^t\right)^T \mathbf{x}^s + \sum_t \alpha^t$$

$$\text{subject to } \sum_{\text{t}} \alpha^t r^t = 0$$

3. Evaluate $\mathbf{w} = \sum_{t=1}^{N} \alpha^t r^t \mathbf{x}^t$ and $w_0 = \langle r^t - \mathbf{w}^T \boldsymbol{x}^t \rangle$,

where < > denotes average data points on margins

One-Class SVM with no slack variables



min R^2

subject to

$$\left\|\mathbf{x}^t - a\right\|^2 \leq R^2$$

What is L_p

$$\mathbf{L_p} = R^2 - \sum_t \alpha^t (R^2 - \| \mathbf{x^t} - \mathbf{a} \|^2)$$

What is the derivative of L_p with respect to R?

$$\frac{\delta L_p}{\delta R} = 2R - 2R \sum \alpha^t$$

What constraint on Lagrange multipliers results from setting this derivative equal to zero?

$$\sum \alpha^t = 1$$

C-SVM: binary classification of non-linearly separable data

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t\left[r^t\left(\mathbf{w}^T x^t + w_0\right) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$

What are the primal variables?

$\boldsymbol{w}$, w_0, and ξ^t

What are the dual variables?

α^t and m^t are Lagrange multipliers.

What role does constant C play?

C is the regularization parameter for the penalty on soft error.

How does increasing C likely effect the width of margins?

Increasing C penalizes soft error more likely decreasing the width of margins with fewer points in the margins.

n-SVM: alternative to C-SVM

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{\mathrm{N}}\sum_t \xi^t - \sum_t \alpha^t\left[r^t\left(\mathbf{w}^T x^t + w_0\right) - \rho + \xi^t\right] - \sum_t \mu^t \xi^t - \delta\rho$$

What are the primal variables?

$\boldsymbol{w}$, w_0, ρ, and ξ^t

What are the dual variables?

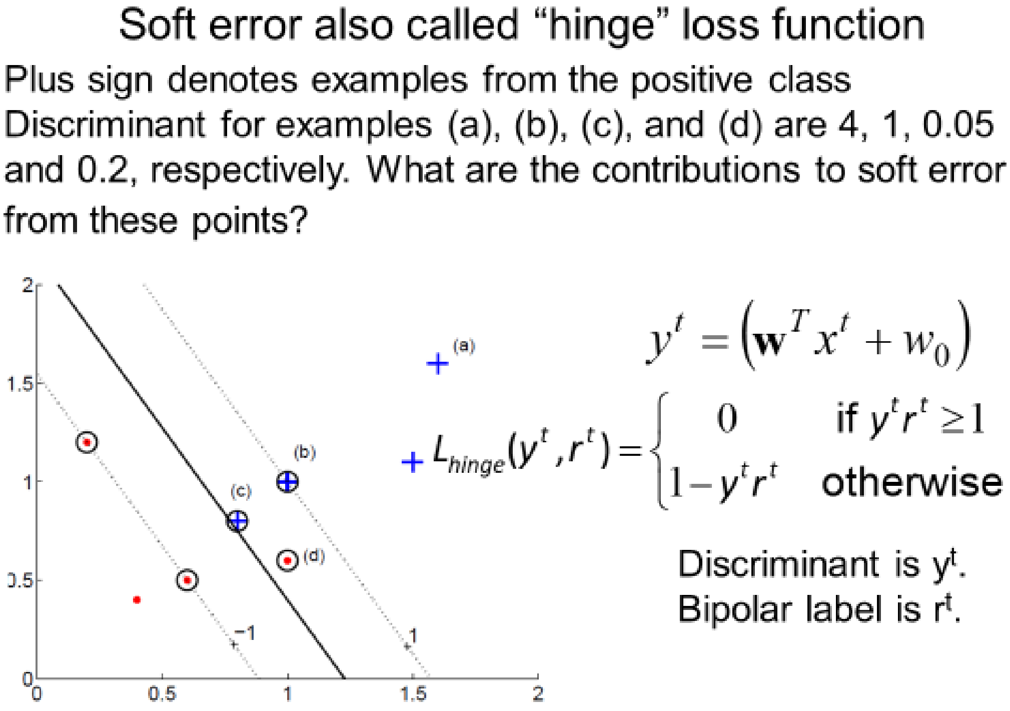
α^t, μ^t, and δ

What is the meaning of regularization parameter v?

v is the upper bound on the fraction of instances in margin.

How does increasing v likely effect the width of margins?

Increasing v likely increases width of margins with more points in margins.



Soft error also called “hinge” loss function

Plus sign denotes examples from the positive class Discriminant for examples (a), (b), (c), and (d) are 4, 1, 0.05 and 0.2, respectively. What are the contributions to soft error from these points?

$$y^t = \left(\mathbf{w}^T x^t + w_0\right)$$

$$L_{hinge}(y^t, r^t) = \begin{cases} 0 & \text{if } y^t r^t \ge 1 \\ 1 - y^t r^t & \text{otherwise} \end{cases}$$

Discriminant is y^t.
Bipolar label is r^t.

r^t is +1 for members and -1 for non-members.

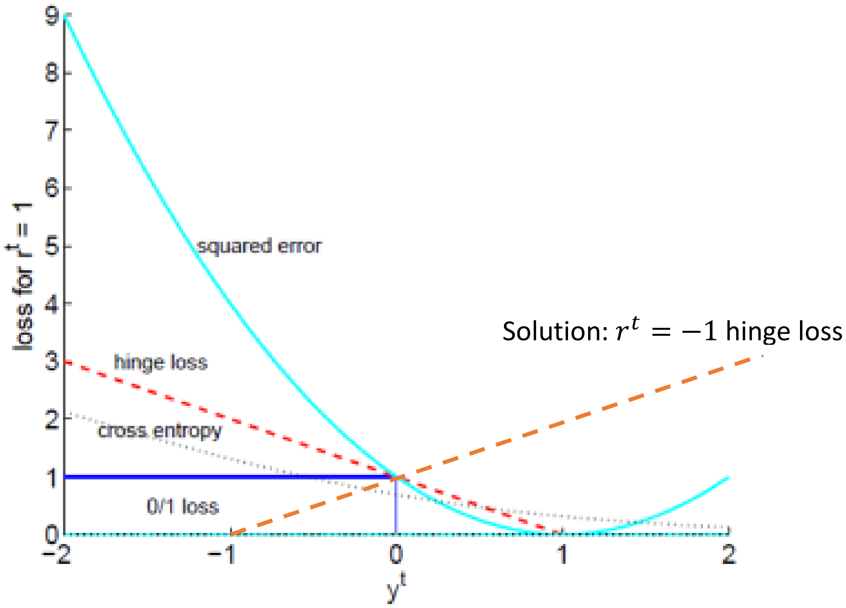
a) $4 * 1 = \; 4 \geq 1$ so $L_{hinge} = 0$

b) $1 * 1 = 1 \geq 1$ so $L_{hinge} = 0$

c) $0.05 * 1 = 0.05 < 1$ so $L_{hinge} = 1 - y^t r^t = 1 - 0.05 = 0.95$

d) $0.2 * -1 = -0.2 < 1$ so $L_{hinge} = 1 - y^t r^t = 1 - (-0.2) = 1.2$

Describe the hinge loss function for instances in the non-member class.



Solution: $r^t = -1$ hinge loss

The above chart shows the hinge loss function for instances in the member class, since $r^t = 1$