Support Vector Machines Part 1

CS 760@UW-Madison

Goals for the lecture

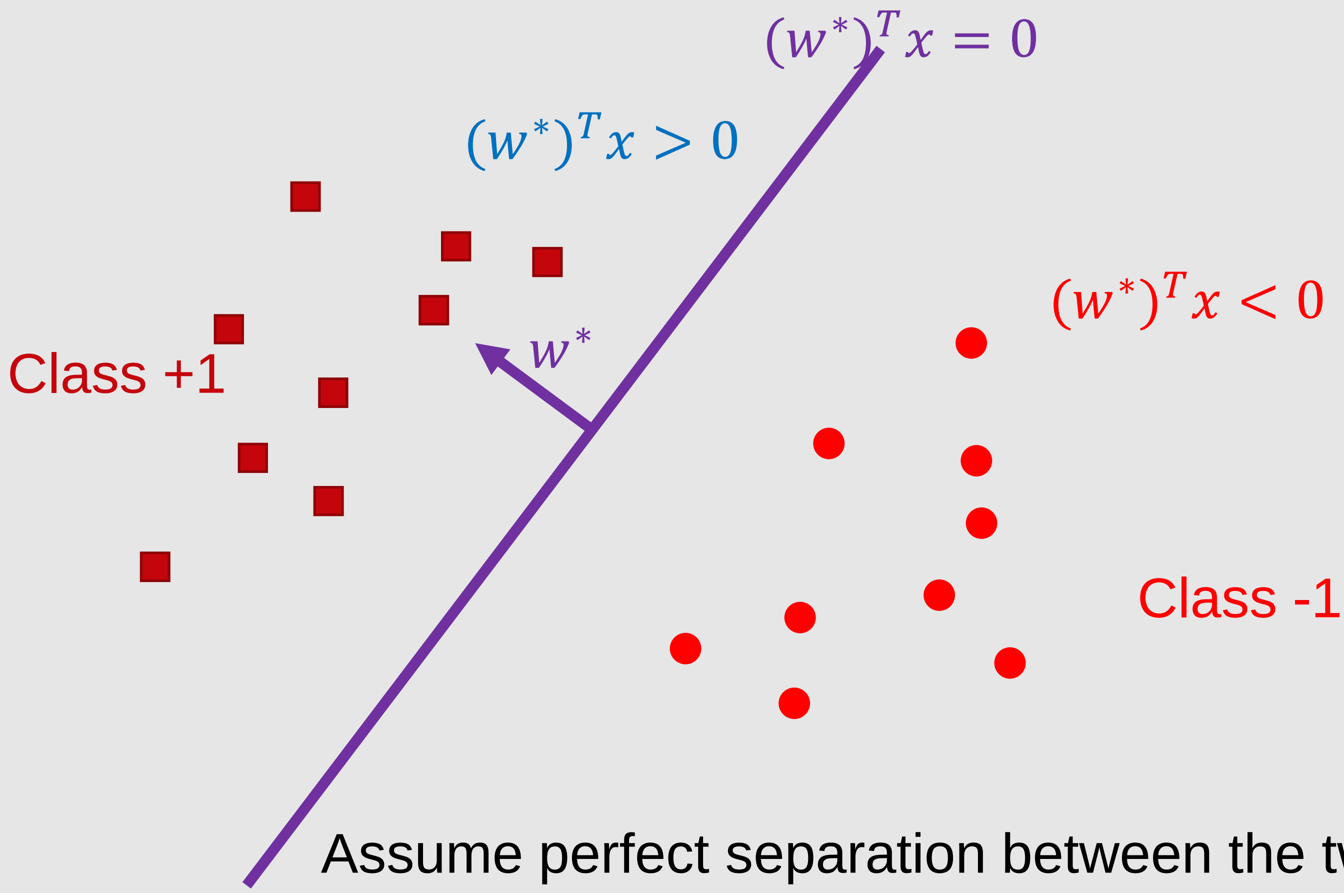
you should understand the following concepts

- the margin
- the linear support vector machine
- the primal and dual formulations of SVM learning
- support vectors

- Optional: variants of SVM
- Optional: Lagrange Multiplier

Motivation

Linear classification



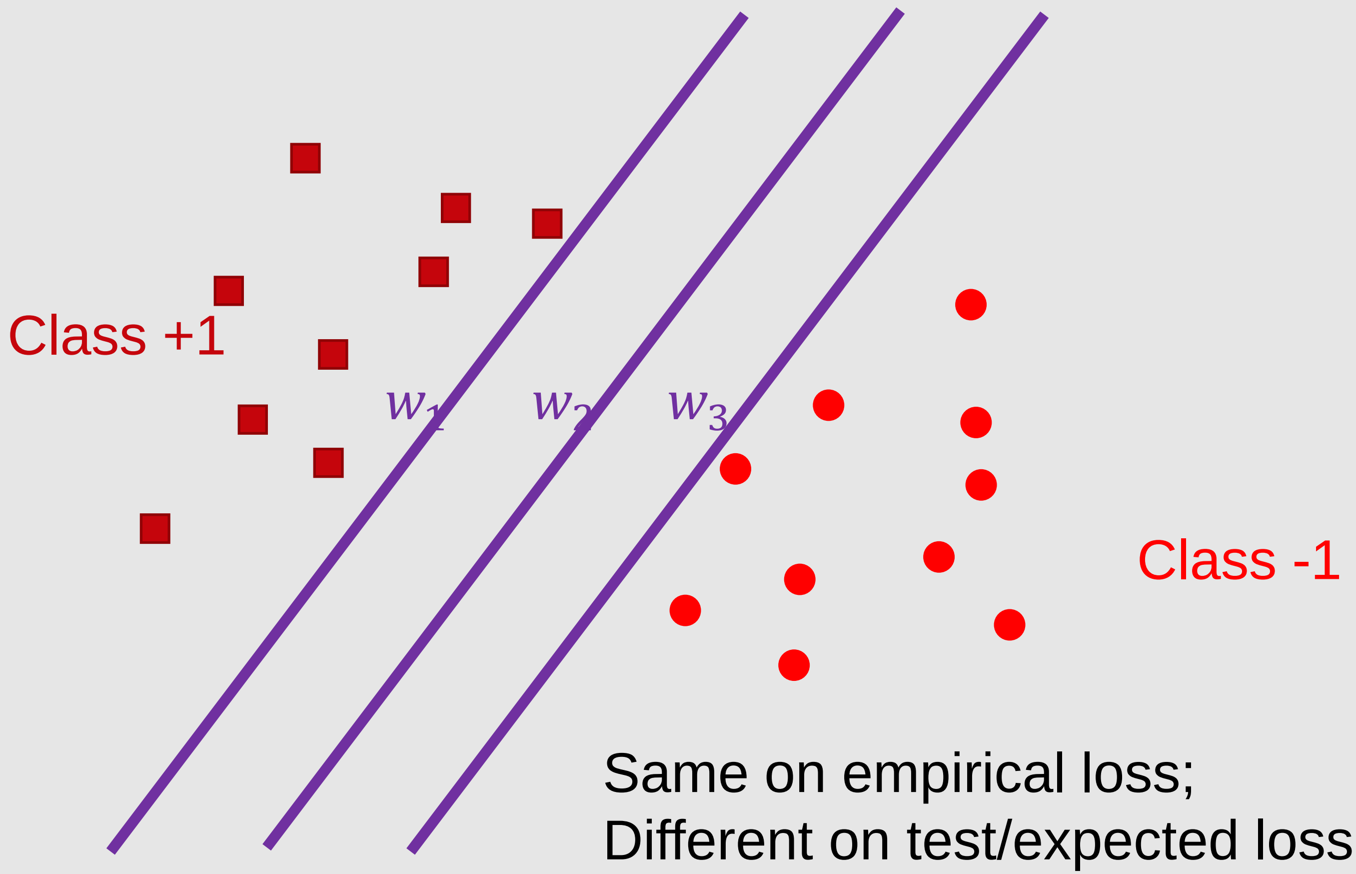
Assume perfect separation between the two classes

Attempt

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Hypothesis $y = \text{sign}(f_w(x)) = \text{sign}(w^T x)$
  - $y = +1$ if $w^T x > 0$
  - $y = -1$ if $w^T x < 0$

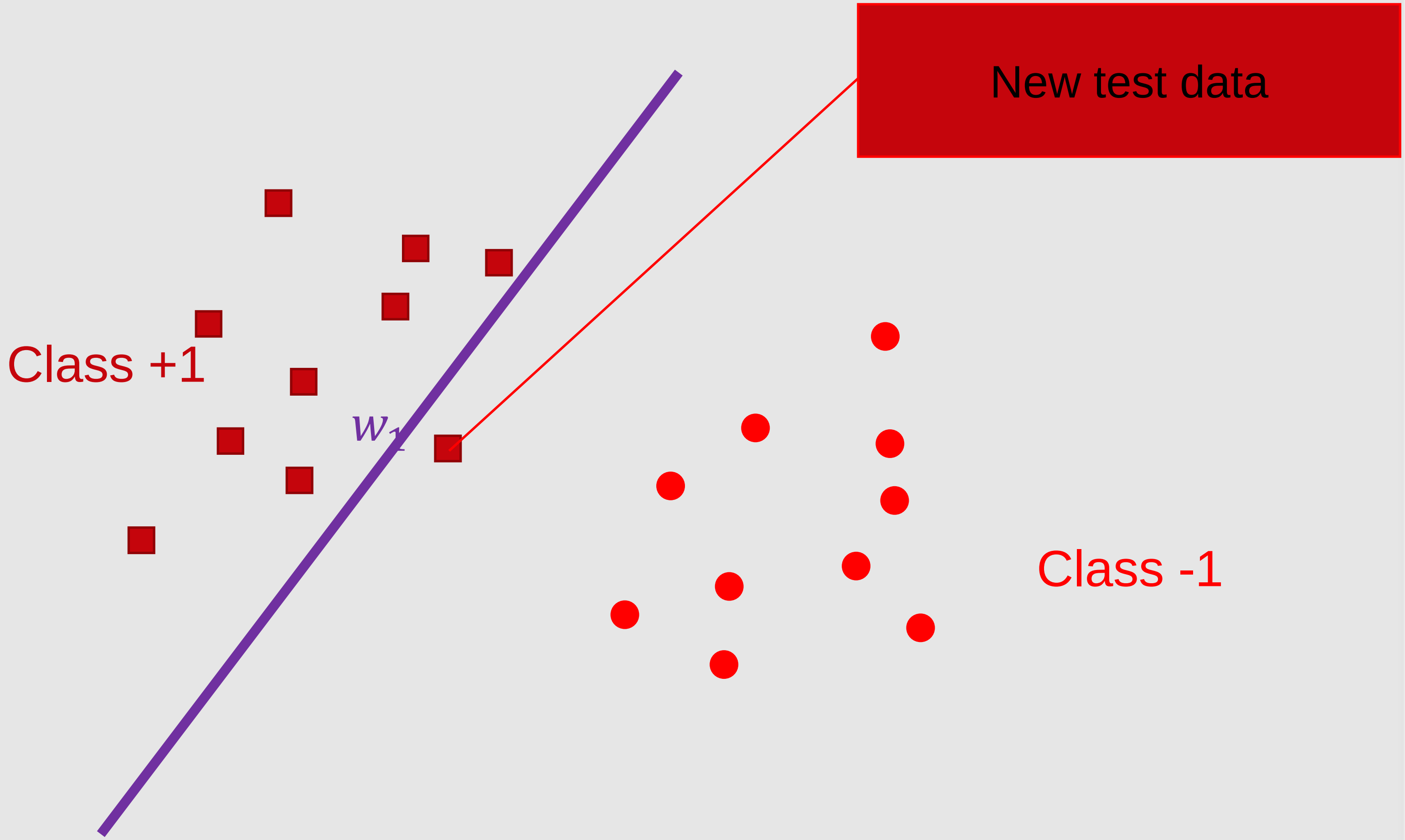
- Let's assume that we can optimize to find w

Multiple optimal solutions?

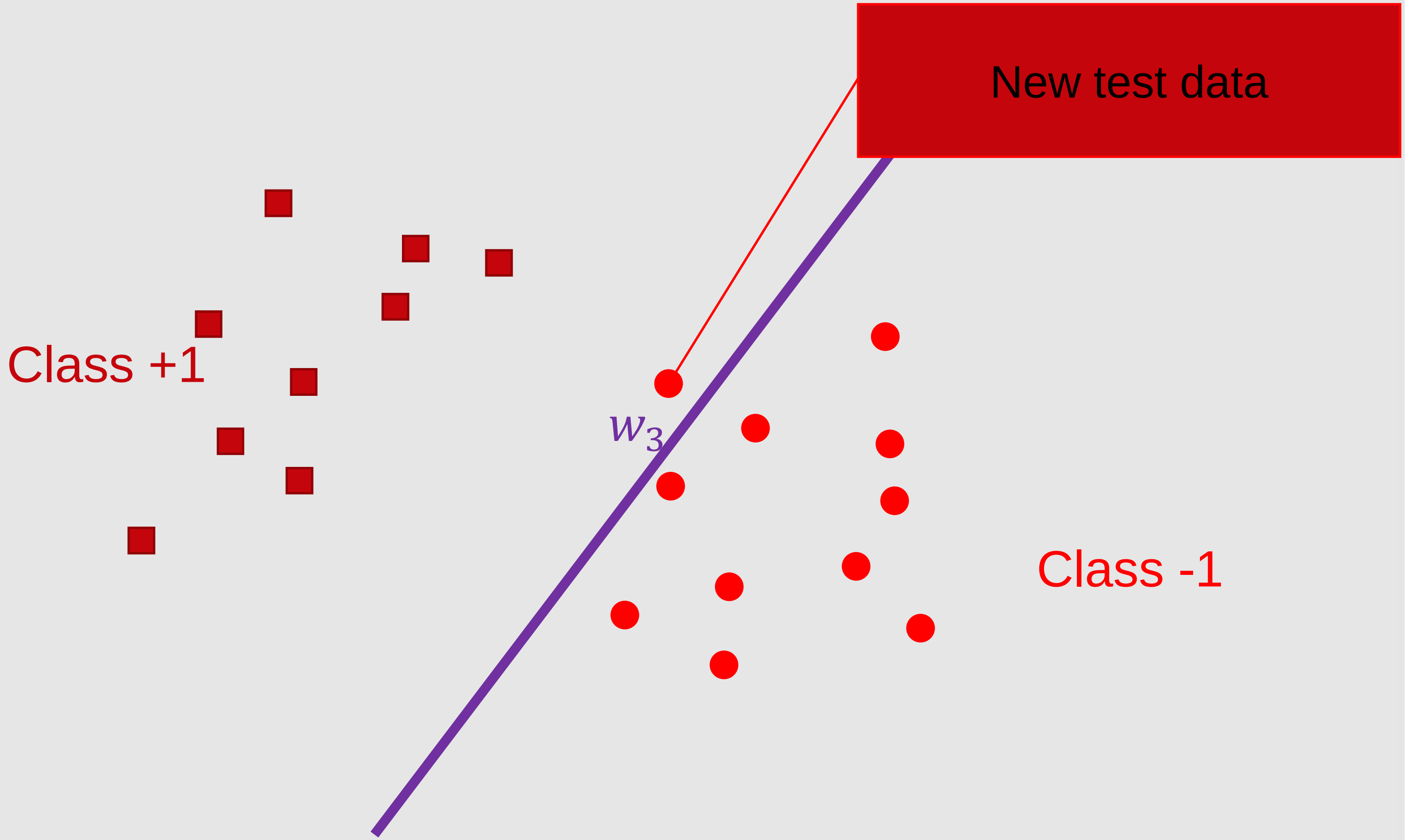


Same on empirical loss;
Different on test/expected loss

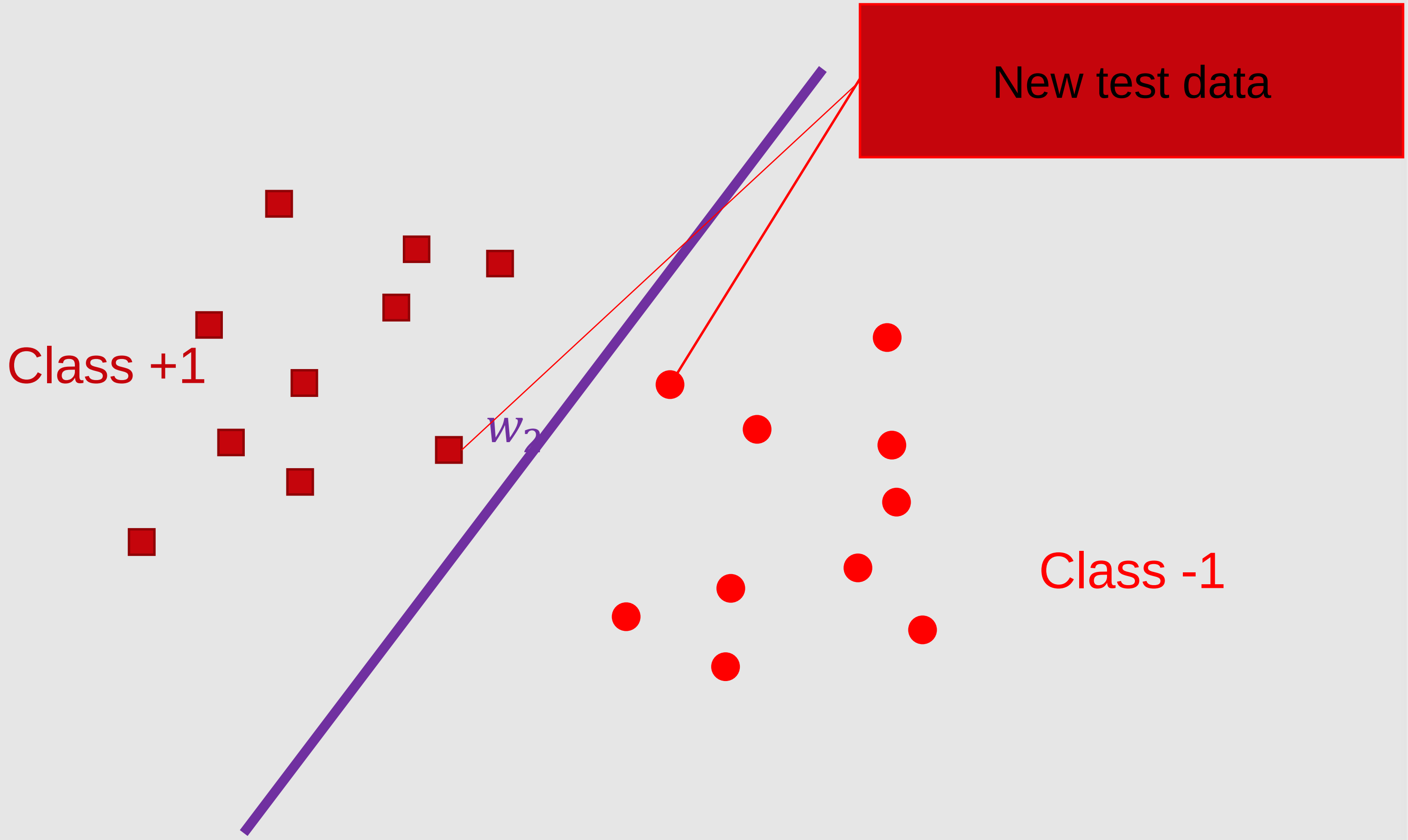
What about w_1?



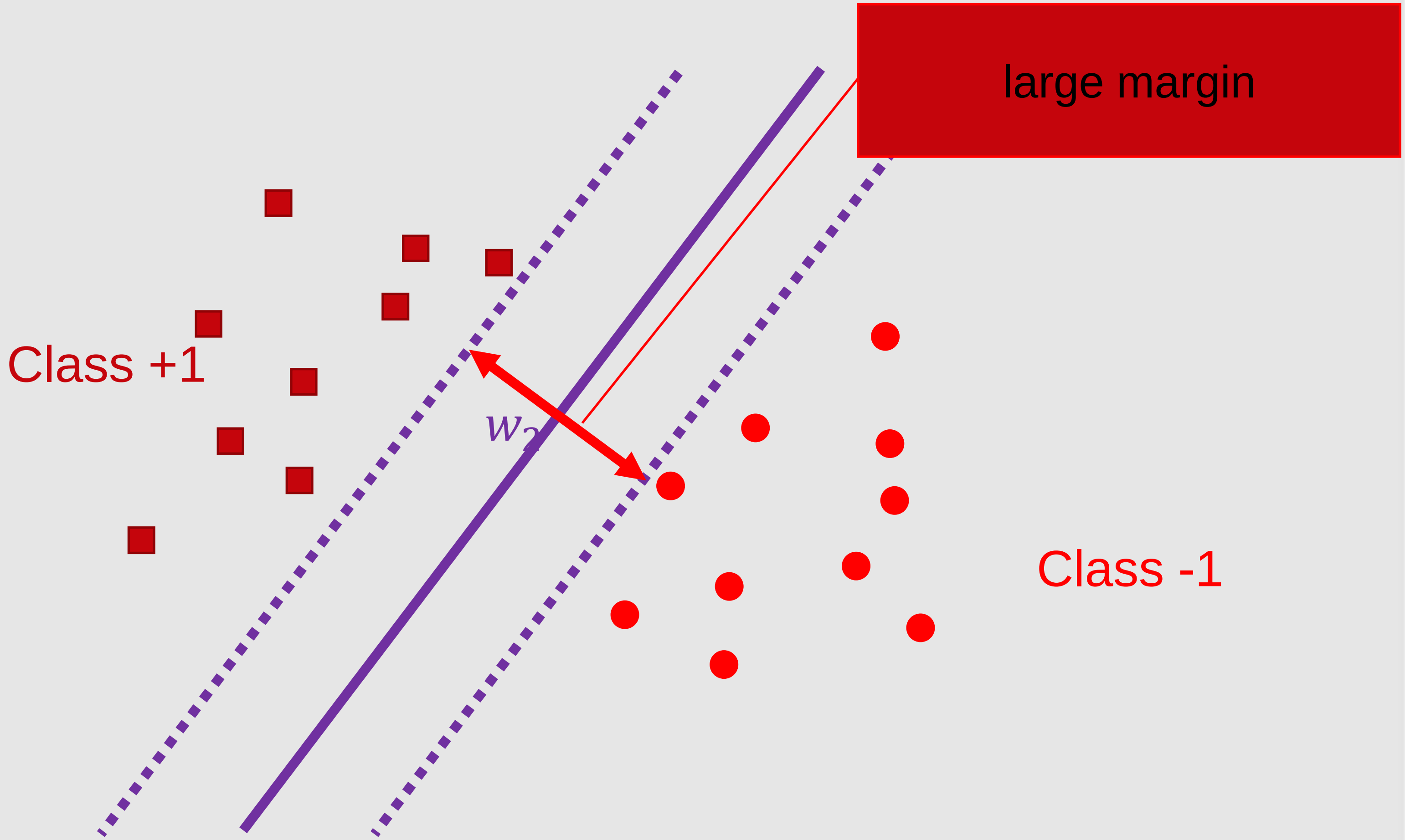
What about w_3?



Most confident: w_2



Intuition: margin



Margin

Margin

We are going to prove the following math expression for margin using a geometric argument

- Lemma 1: x has distance $\frac{|f_w(x)|}{||w||}$ to the hyperplane $f_w(x) = w^T x = 0$
- Lemma 2: x has distance $\frac{|f_{w,b}(x)|}{||w||}$ to the hyperplane $f_{w,b}(x) = w^T x + b = 0$

Need two geometric facts:

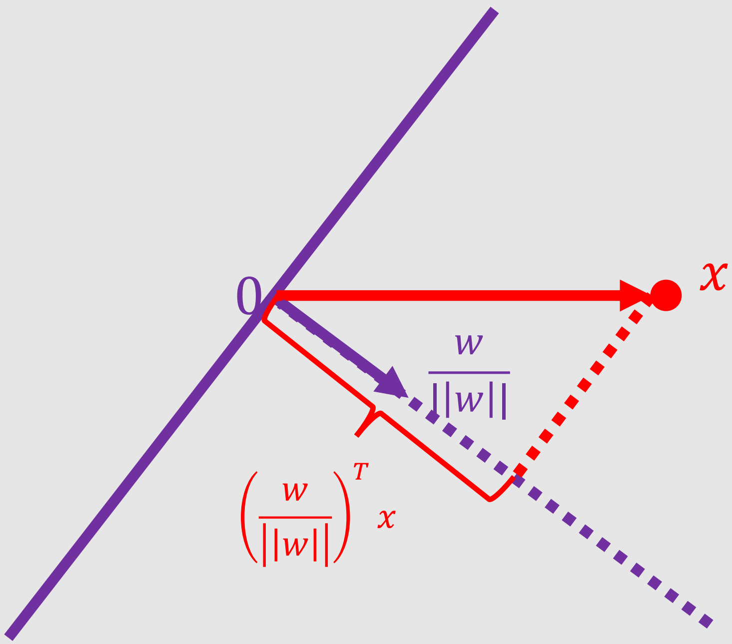
- w is orthogonal to the hyperplane $f_{w,b}(x) = w^T x + b = 0$
- Let v be a direction (i.e., unit vector). Then the length of the projection of x on v is $v^T x$

Margin

- Lemma 1: x has distance $\frac{|f_w(x)|}{||w||}$ to the hyperplane $f_w(x) = w^T x = 0$

Proof:

- w is orthogonal to the hyperplane
- The unit direction is $\frac{w}{||w||}$
- The projection of x is $\left(\frac{w}{||w||}\right)^T x = \frac{f_w(x)}{||w||}$



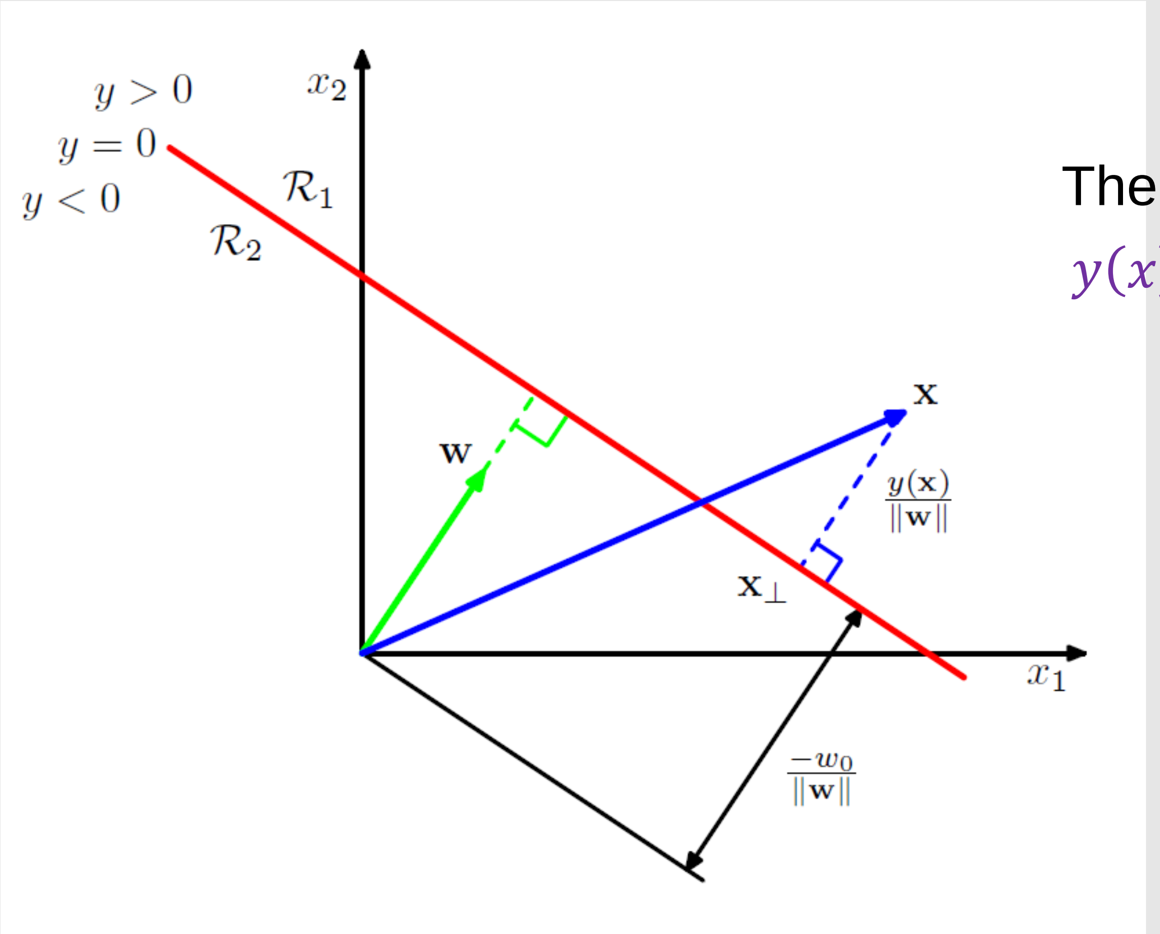
Margin: with bias

- Lemma 2: x has distance $\frac{|f_{w,b}(x)|}{||w||}$ to the hyperplane $f_{w,b}(x) = w^T x + b = 0$

Proof:

- Let $x = x_\perp + r\frac{w}{||w||}$, then $|r|$ is the distance
- Multiply both sides by w^T and add b
- Left hand side: $w^T x + b = \ f_{w,b}(x)$
- Right hand side: $w^T x_\perp + r\frac{w^T w}{||w||} + b = 0 + r||w||$

Margin: with bias



The notation here is:

$y(x) = w^T x + w_0$

Figure from *Pattern Recognition and Machine Learning*, Bishop

Support Vector Machine (SVM)

SVM: objective

- Absolute margin over all training data points:

$$\gamma = \min_i \frac{|f_{w,b}(x_i)|}{||w||}$$

- Since only want correct $f_{w,b}$, and recall $y_i \in \{+1, -1\}$, we define the margin to be

$$\gamma = \min_i \frac{y_i f_{w,b}(x_i)}{||w||}$$

- If $f_{w,b}$ incorrect on some x_i, the margin is negative

SVM: objective

- Maximize margin over all training data points:

$$\max_{w,b} \gamma = \max_{w,b} \min_{i} \frac{y_i f_{w,b}(x_i)}{||w||} = \max_{w,b} \min_{i} \frac{y_i (w^T x_i + b)}{||w||}$$

- A bit complicated …

SVM: simplified objective

- Observation: when (w, b) scaled by a factor c, the margin unchanged

$$\frac{y_i(cw^T x_i + cb)}{||cw||} = \frac{y_i(w^T x_i + b)}{||w||}$$

- Let's consider a fixed scale such that

$$y_{i^*}(w^T x_{i^*} + b) = 1$$

where x_{i^*} is the point closest to the hyperplane

SVM: simplified objective

- Let's consider a fixed scale such that

$$y_{i^*}(w^T x_{i^*} + b) = 1$$

where x_{i^*} is the point closet to the hyperplane

- Now we have for all data

$$y_i(w^T x_i + b) \geq 1$$

and at least for one i the equality holds

- Then the margin over all training points is $\frac{1}{||w||}$

SVM: simplified objective

- Optimization simplified to

$$\min_{w,b} \frac{1}{2} \left| |w| \right|^2$$

$$y_i(w^T x_i + b) \geq 1, \forall i$$

- How to find the optimum $\hat{w}^*$?
- Solved by Lagrange multiplier method

SVM: optimization

SVM: optimization

- Optimization (Quadratic Programming):

$$\min_{w,b} \frac{1}{2}\left||w|\right|^2$$

$$y_i(w^T x_i + b) \geq 1, \forall i$$

- Generalized Lagrangian:

$$\mathcal{L}(w, b, \boldsymbol{\alpha}) = \frac{1}{2}\left||w|\right|^2 - \sum_i \alpha_i [y_i(w^T x_i + b) - 1]$$

where $\boldsymbol{\alpha}$ is the Lagrange multiplier

SVM: optimization

- KKT conditions:

$$\frac{\partial \mathcal{L}}{\partial w} = 0, \rightarrow w = \sum_i \alpha_i y_i x_i \quad (1)$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0, \rightarrow 0 = \sum_i \alpha_i y_i \quad (2)$$

- Plug into $\mathcal{L}$:

$$\mathcal{L}(w, b, \boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (3)$$

combined with $0 = \sum_i \alpha_i y_i \,, \alpha_i \geq 0$

SVM: optimization

Only depend on inner products

- Reduces to dual problem:

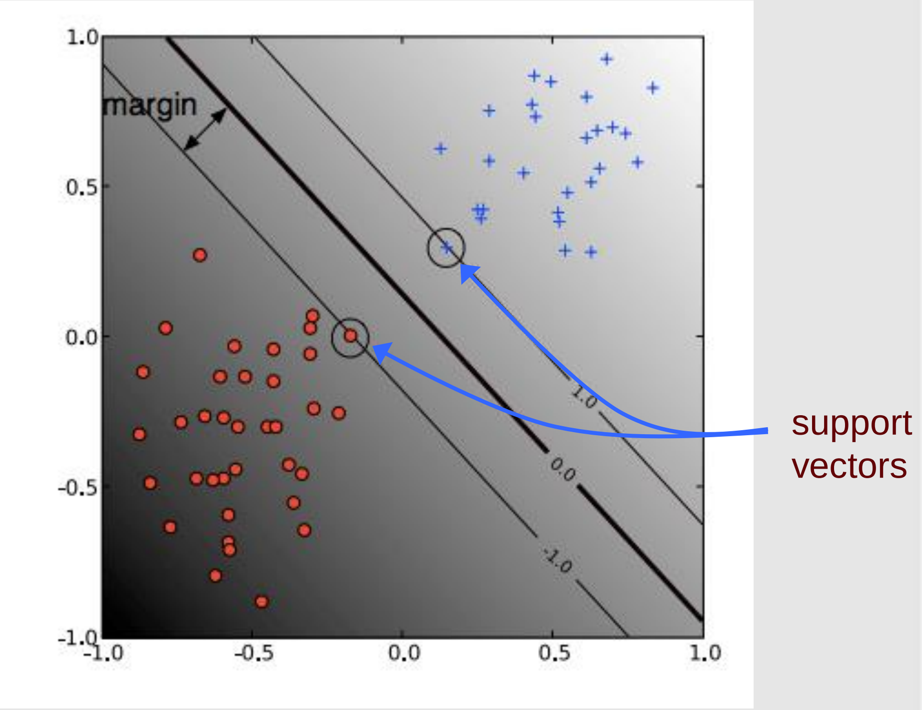
$$\mathcal{L}(w, b, \boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\sum_i \alpha_i y_i = 0, \alpha_i \geq 0$$

- Since $w = \sum_i \alpha_i y_i x_i$, we have $w^T x + b = \sum_i \alpha_i y_i x_i^T x + b$

Support Vectors

- final solution is a sparse linear combination of the training instances
- those instances with $\alpha_i > 0$ are called *support vectors*
  - they lie on the margin boundary
- solution NOT changed if delete the instances with $\alpha_i = 0$



Optional: Lagrange Multiplier

Lagrangian

- Consider optimization problem:

$$\min_{w} \; f(w)$$

$$h_i(w) = 0, \forall 1 \leq i \leq l$$

- Lagrangian:

$$\mathcal{L}(w, \boldsymbol{\beta}) = \; f(w) + \sum_i \beta_i h_i(w)$$

where β_i's are called Lagrange multipliers

Lagrangian

- Consider optimization problem:

$$\min_{w} \; f(w)$$

$$h_i(w) = 0, \forall 1 \leq i \leq l$$

- Solved by setting derivatives of Lagrangian to 0

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0$$

Generalized Lagrangian

- Consider optimization problem:

$$\min_{w} \ f(w)$$

$$g_i(w) \leq 0, \forall 1 \leq i \leq k$$

$$h_j(w) = 0, \forall 1 \leq j \leq l$$

- Generalized Lagrangian:

$$\mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \ f(w) + \sum_{i} \alpha_i g_i(w) + \sum_{j} \beta_j h_j(w)$$

where α_i, β_j's are called Lagrange multipliers

Generalized Lagrangian

- Consider the quantity:

$$\theta_P(w) := \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- Why?

$$\theta_P(w) = \begin{cases} f(w), & \text{if } w \text{ satisfies all the constraints} \\ +\infty, & \text{if } w \text{ does not satisfy the constraints} \end{cases}$$

- So minimizing $f(w)$ is the same as minimizing $\theta_P(w)$

$$\min_w f(w) = \min_w \theta_P(w) = \min_w \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

Lagrange duality

- The primal problem

$$p^* \coloneqq \min_w f(w) = \min_w \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- The dual problem

$$d^* \coloneqq \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \min_w \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- Always true:

$$d^* \leq p^*$$

Lagrange duality

- The primal problem

$$p^* := \min_w f(w) = \min_w \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- The dual problem

$$d^* := \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \min_w \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- Interesting case: when do we have

$$d^* = p^*?$$

Lagrange duality

- Theorem: under proper conditions, there exists $(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ such that

$$d^* = \mathcal{L}(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = p^*$$

Moreover, $(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ satisfy Karush-Kuhn-Tucker (KKT) conditions:

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0, \qquad \alpha_i g_i(w) = 0$$

$$g_i(w) \le 0, \;\; h_j(w) = 0, \qquad \alpha_i \ge 0$$

Lagrange duality

- Theorem: under proper conditions, there exists $(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ such that

$$d^* = \mathcal{L}(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = p^*$$

dual complementarity

Moreover, $(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ satisfy Karush-Kuhn-Tucker (KKT) conditions:

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0, \qquad \alpha_i g_i(w) = 0$$

$$g_i(w) \leq 0, \; h_j(w) = 0, \qquad \alpha_i \geq 0$$

Lagrange duality

- Theorem: under proper conditions, there exists $(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ such that

$$d^* = \mathcal{L}(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = p^*$$

primal constraints

satisfy Karush-Kuhn-Tu

dual constraints

conditions:

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0, \qquad \alpha_i g_i(w) = 0$$

$$g_i(w) \leq 0, \ \ h_j(w) = 0, \qquad \alpha_i \geq 0$$

Lagrange duality

- What are the proper conditions?
- A set of conditions (Slater conditions):
  - f, g_i convex, h_j affine, and exists w satisfying all $g_i(w) < 0$
- There exist other sets of conditions
  - Check textbooks, e.g., Convex Optimization by Boyd and Vandenberghe

Optional: Variants of SVM

Hard-margin SVM

- Optimization (Quadratic Programming):

$$\min_{w,b} \frac{1}{2}\left|\left|w\right|\right|^2$$

$$y_i(w^T x_i + b) \geq 1, \forall i$$

Soft-margin SVM [Cortes & Vapnik, *Machine Learning* 1995]

- if the training instances are not linearly separable, the previous formulation will fail
- we can adjust our approach by using *slack variables* (denoted by ζ_i) to tolerate errors

$$\min_{w,b,\zeta_i} \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_i \zeta_i$$

$$y_i(w^T x_i + b) \geq 1 - \zeta_i, \zeta_i \geq 0, \forall i$$

- C determines the relative importance of maximizing margin vs. minimizing slack

The effect of C in soft-margin SVM

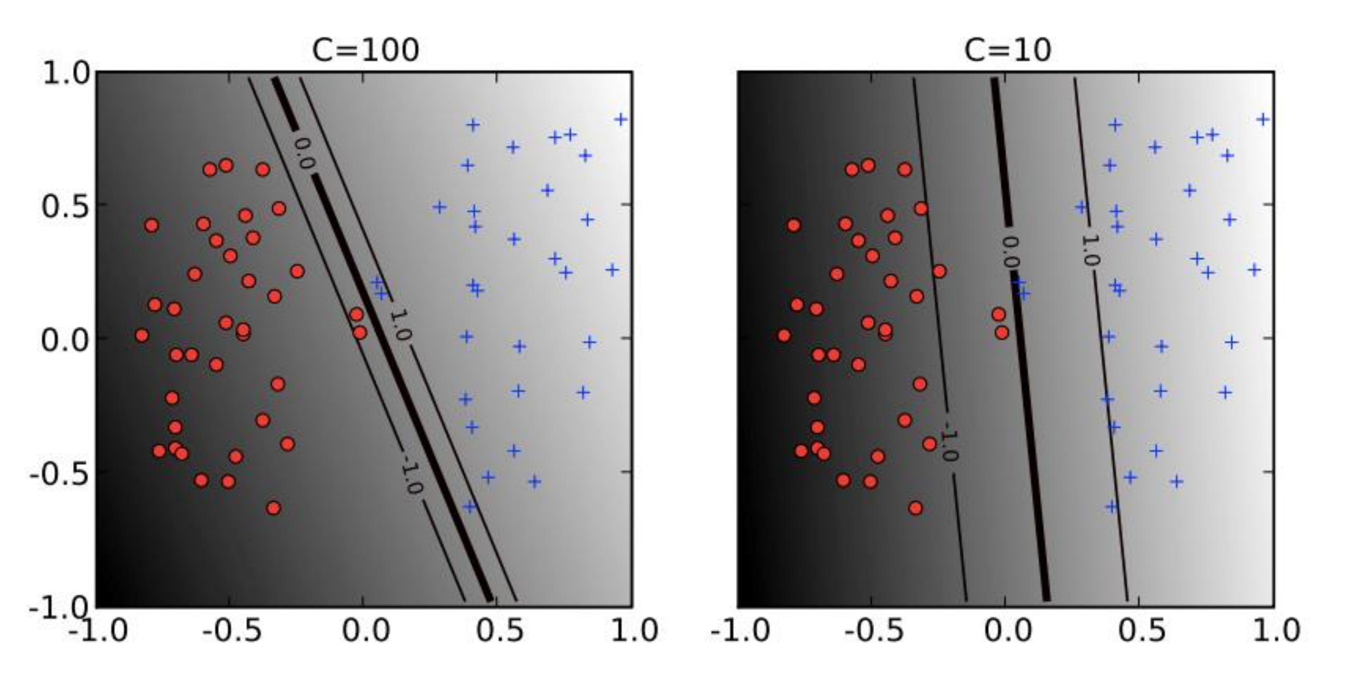
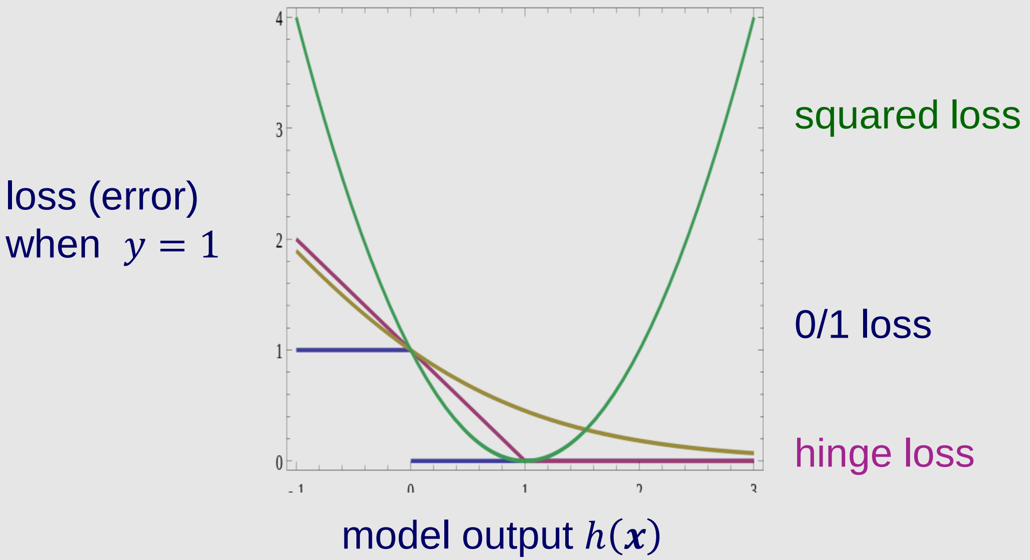


Figure from Ben-Hur & Weston,
Methods in Molecular Biology 2010

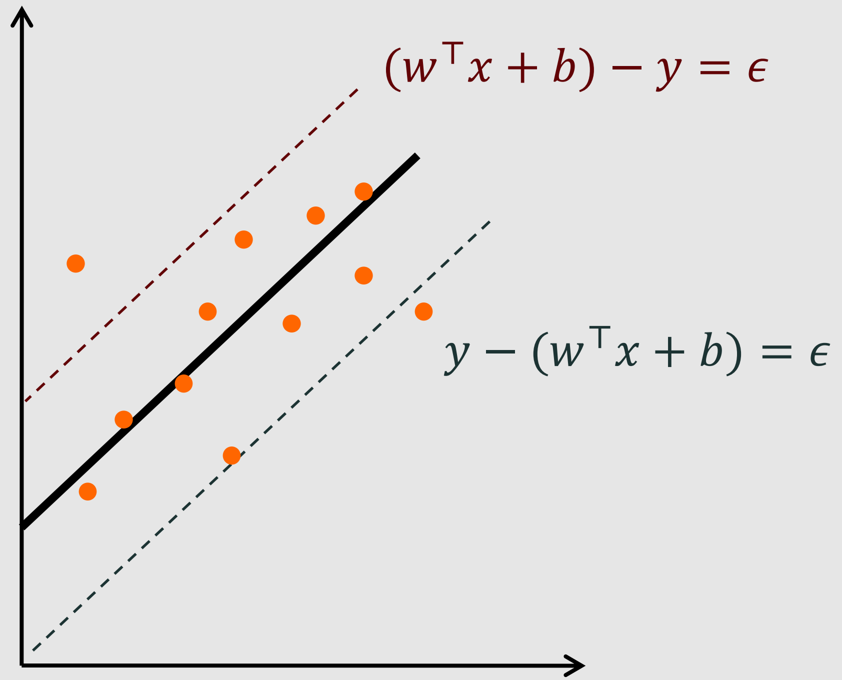
Hinge loss

- when we covered neural nets, we talked about minimizing squared loss and cross-entropy loss
- SVMs minimize *hinge loss*



Support Vector Regression

- the SVM idea can also be applied in regression tasks
- an ϵ-insensitive error function specifies that a training instance is well explained if the model's prediction is within ϵ of y_i



Support Vector Regression

- Regression using *slack variables* (denoted by ζ_i, ξ_i) to tolerate errors

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

$$(w^T x_i + b) - y_i \leq \epsilon + \zeta_i,$$
$$y_i - (w^T x_i + b) \leq \epsilon + \xi_i,$$
$$\zeta_i, \xi_i \geq 0.$$

slack variables allow predictions for some training instances to be off by more than ϵ

THANK YOU

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Elad Hazan, Tom Dietterich, and Pedro Domingos.