



CS 6375

Support Vector Machines

Rishabh Iyer

University of Texas at Dallas

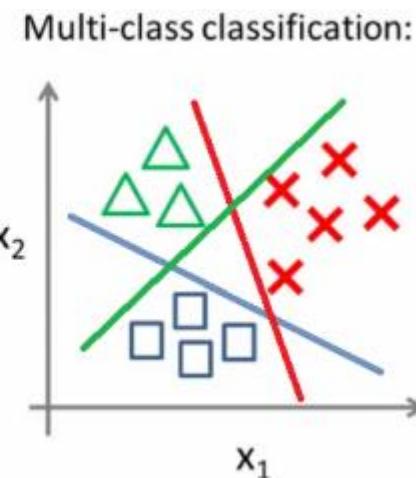
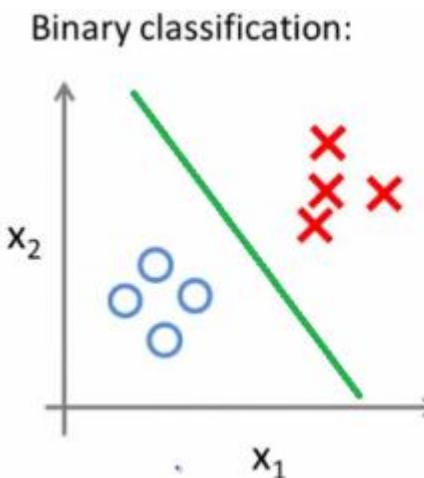
Announcements (09/12/2022)

- Assignment 1 will be out by EOW
- Either next Mon or Wed, we will have a discussion on project ideas
- Midterm exam will be in the 2nd or 3rd week of October.

Recap: Classification

Classification vs Regression

- Input: pairs of points $(x^{(1)}, y^{(1)}), \dots, (x^{(M)}, y^{(M)})$ with $x^{(m)} \in \mathbb{R}^n$
- $y^{(m)} \in [0, k - 1]$
- If $k = 2$, we get Binary classification



Recap: Hypothesis Space

- **Hypothesis space:** set of allowable functions $f: \underline{X} \rightarrow Y$
- Goal: find the “best” element of the hypothesis space
 - How do we measure the quality of f ?

$$f(x, w, b) = \text{sign}(w^T x + b) \rightarrow \text{classification}$$

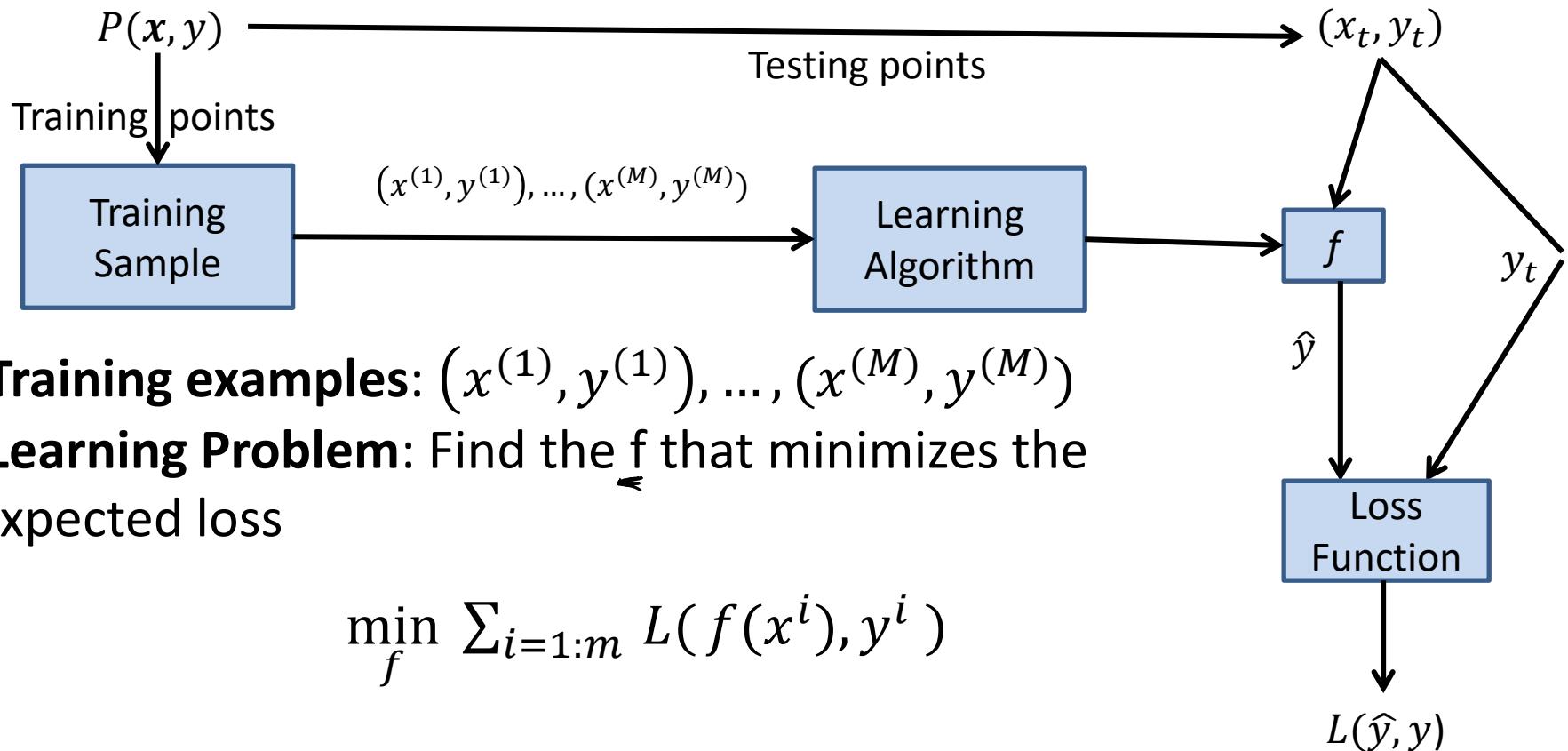
$$w^T x + b \rightarrow \text{regression}$$

Goal: $\forall i, f(x^{(i)}, w, b) \approx y^{(i)}$

Loss: $L(w, b) = [y^{(i)} - f(x^{(i)}, w, b)]^2$ (Regression)

$L(w, b)$ = $\max_3 (0, -y^{(i)} f(x^{(i)}, w, b))$ (Perception)

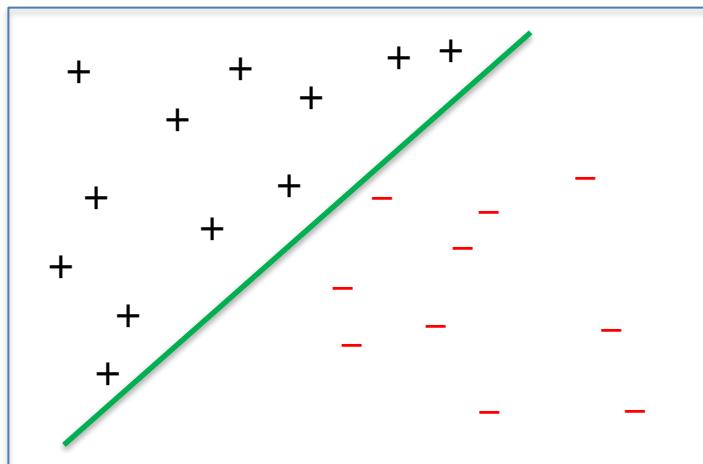
Recap: Supervised Learning Workflow



- **Testing:** Given a new point (x_t, y_t) drawn from P , the classifier is given x and predicts $\hat{y}_t = f(x_t)$
- **Evaluation:** Measure the error $Err(\hat{y}_t, y_t)$ – often same as L

Recap: Binary Classification

- Input $(x^{(1)}, y^{(1)}), \dots, (x^{(M)}, y^{(M)})$ with $x^{(m)} \in \mathbb{R}^n$ and $y^{(m)} \in \{-1, +1\}$
- We can think of the observations as points in \mathbb{R}^n with an associated sign (either +/- corresponding to 0/1)
- An example with $n = 2$



In this case, we say
that the
observations are
linearly separable

0/1 Loss Vs Perceptron Loss

- Zero/One Loss which counts the number of mis-classifications:

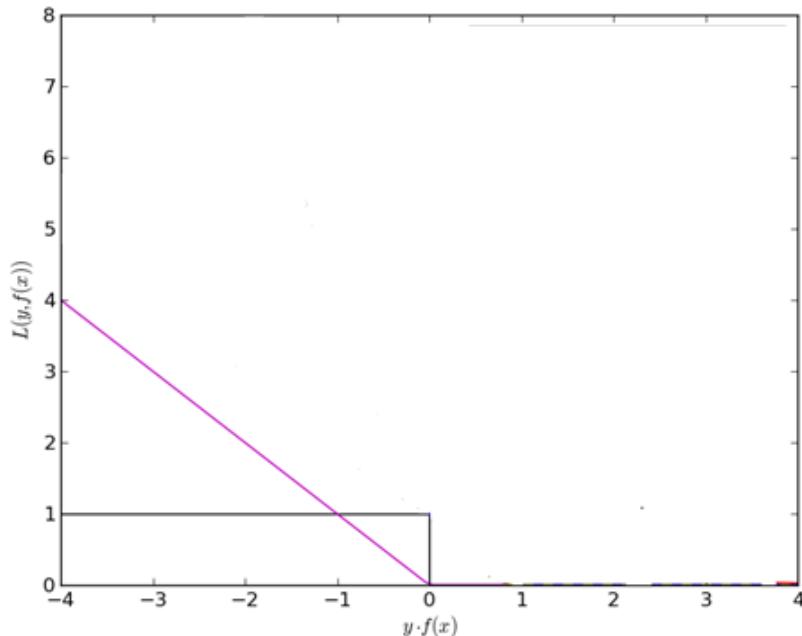
$$\text{zero/one loss} = \frac{1}{2} \sum_m |y^{(m)} - \text{sign}(w^T x^{(m)} + b)|$$

Not Cont
 Grad techniques
 won't work

- Perceptron Loss:

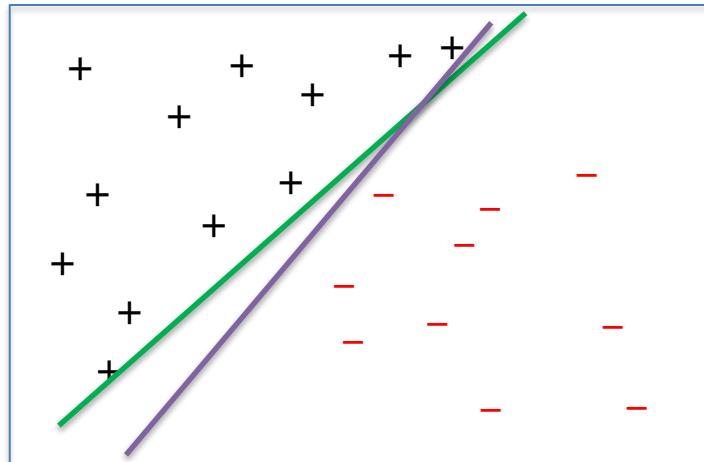
$$\text{perceptron loss} = \sum_m \max\{0, -y^{(m)}(w^T x^{(m)} + b)\}$$

NP hard



Perceptron Drawbacks

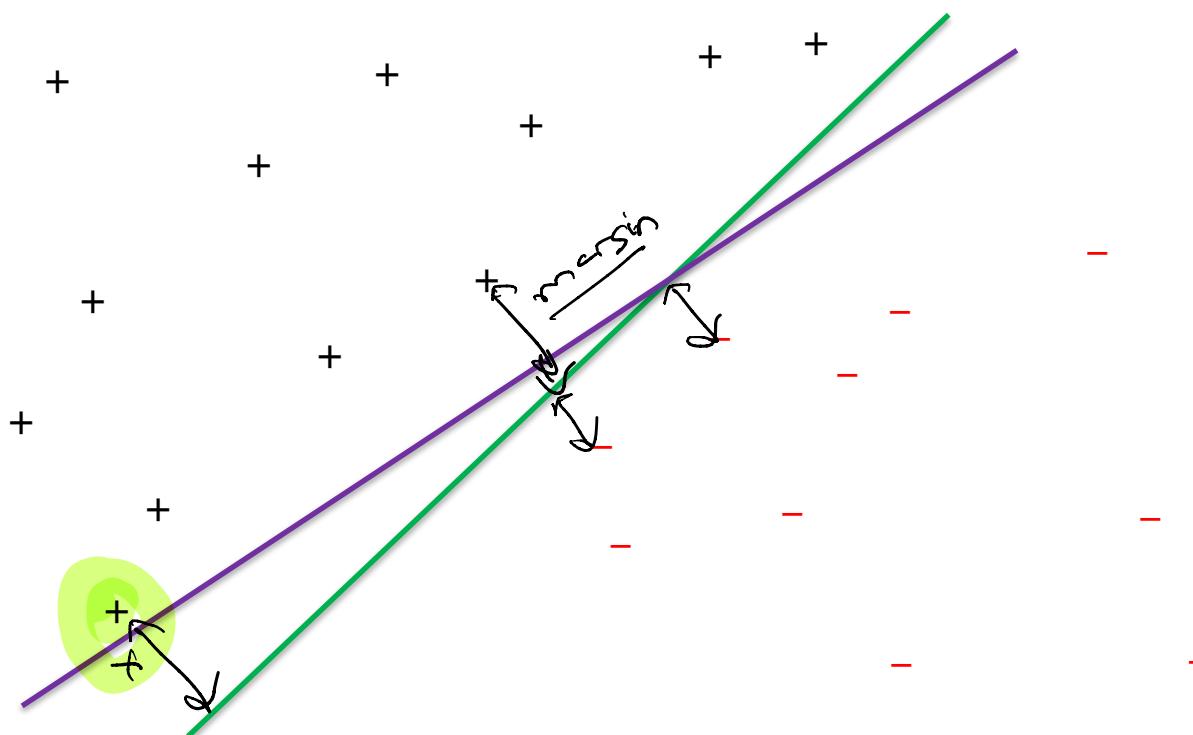
- No convergence guarantees if the observations are not linearly separable
- Can overfit
 - There can be a number of perfect classifiers, but the perceptron algorithm doesn't have any mechanism for choosing between them



Support Vector Machines



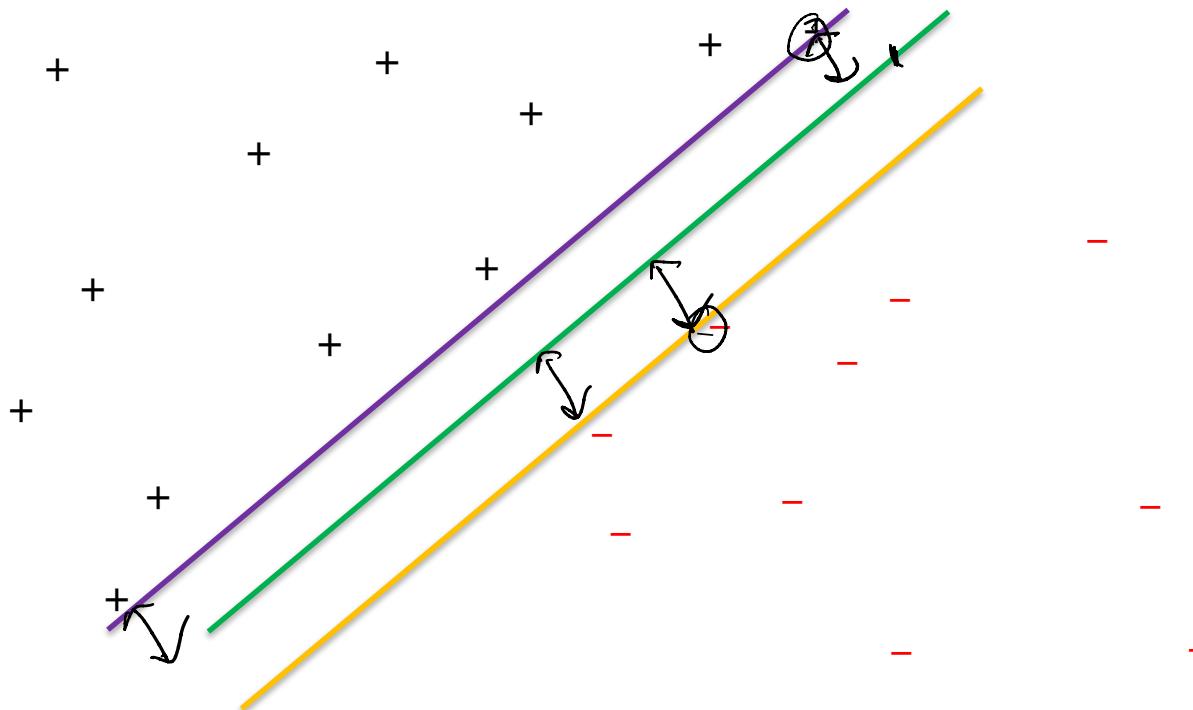
- How can we decide between perfect classifiers?



Support Vector Machines



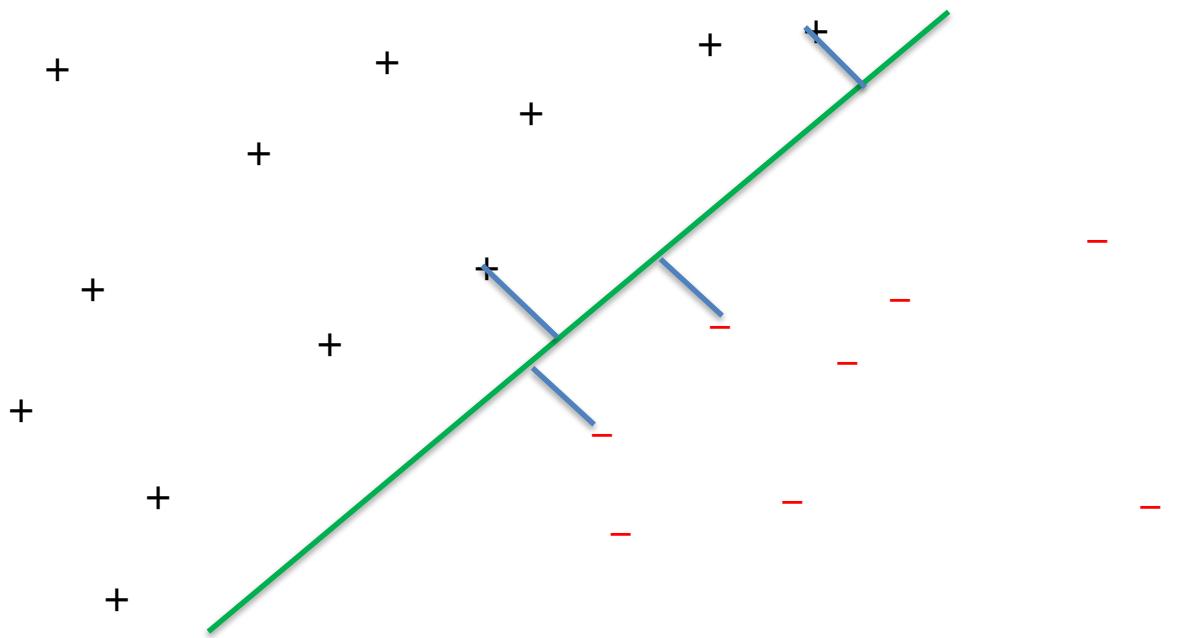
- How can we decide between perfect classifiers?



Support Vector Machines



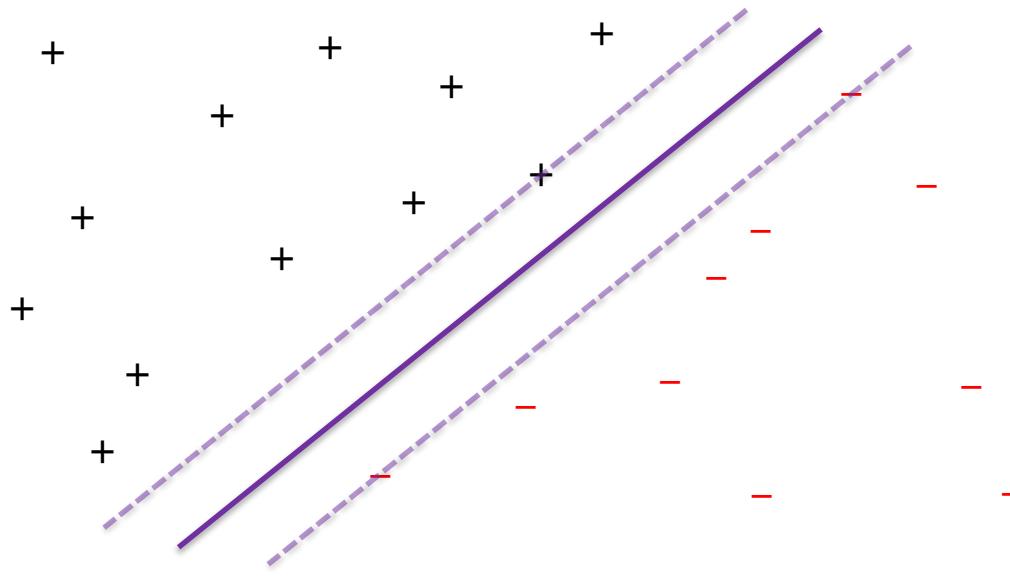
- Define the margin to be the distance of the closest data point to the classifier



Support Vector Machines



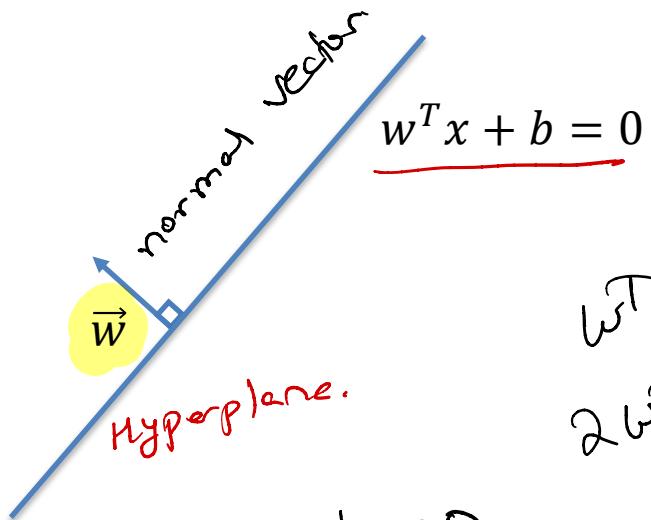
- Support vector machines (SVMs)



- Choose the classifier with the largest margin
 - Has good practical and theoretical performance

"Test Perf"
Generalization

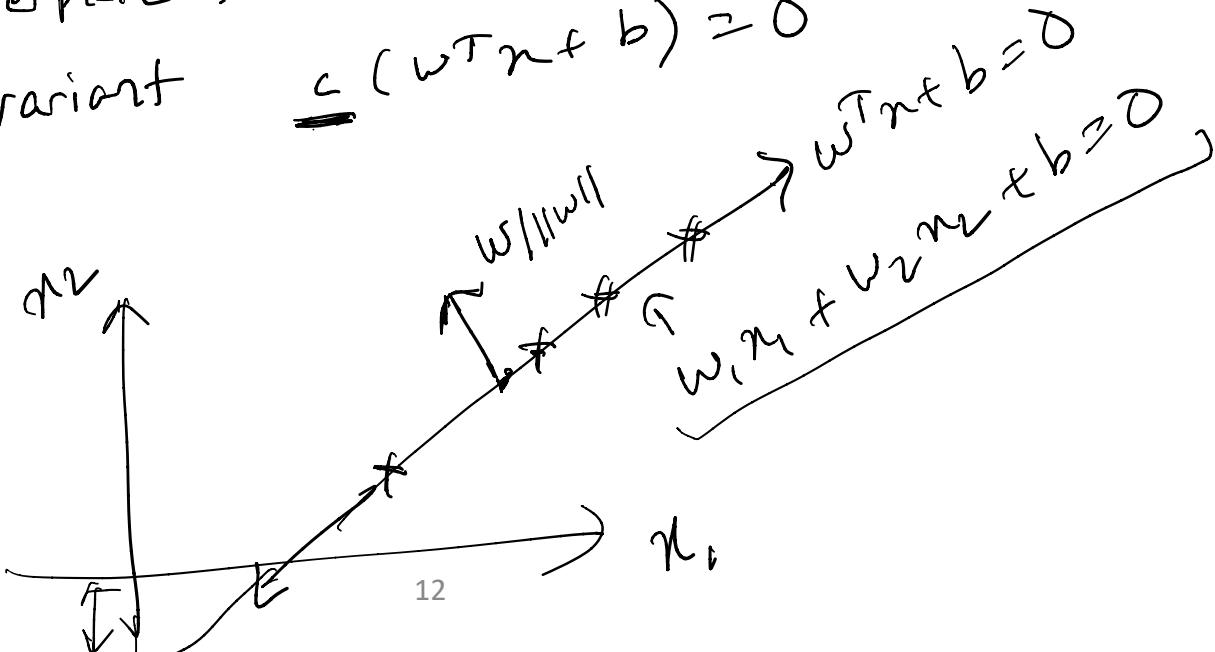
Some Geometry



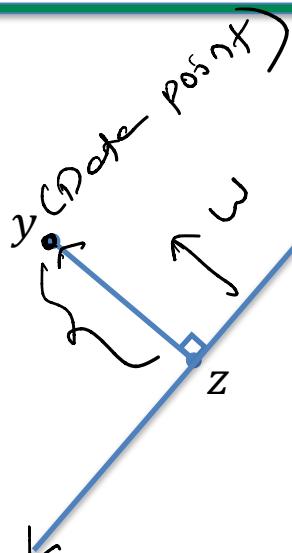
- ① Eq. of hyperplane :
- ② scale invariant

$$w^T n + b = 0$$
$$\underline{c} (w^T n + b) = 0$$

$$w^T n + b = 0$$
$$2w^T n + 2b = 0$$

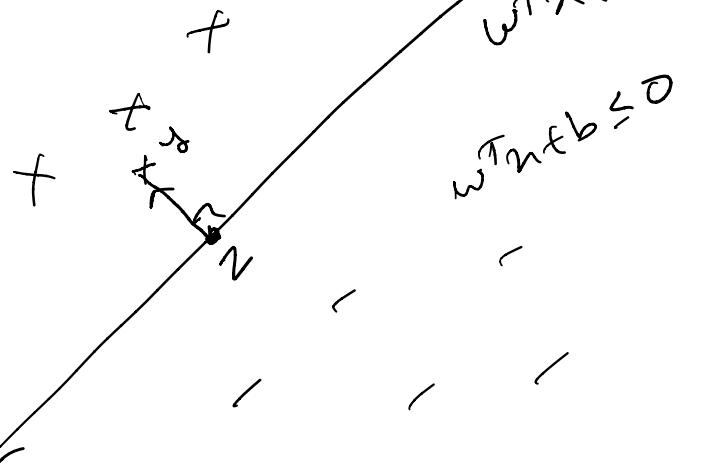


Some Geometry



$$w^T x + b = 0$$

$$w^T x + b \geq 0$$



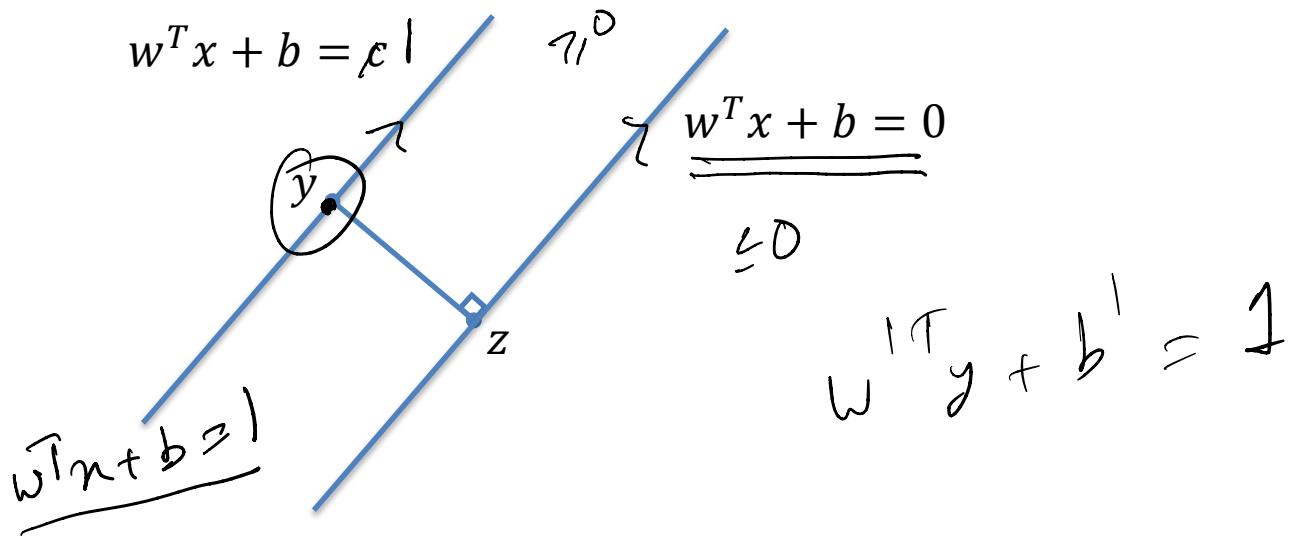
Claim:

$$\|y - z\| = \|y - z\| \frac{w}{\|w\|}$$

$$\Rightarrow \frac{\|y - z\|}{\|y - z\|} = \frac{w}{\|w\|}$$

$$\frac{\|y - z\|}{\|y - z\|}$$

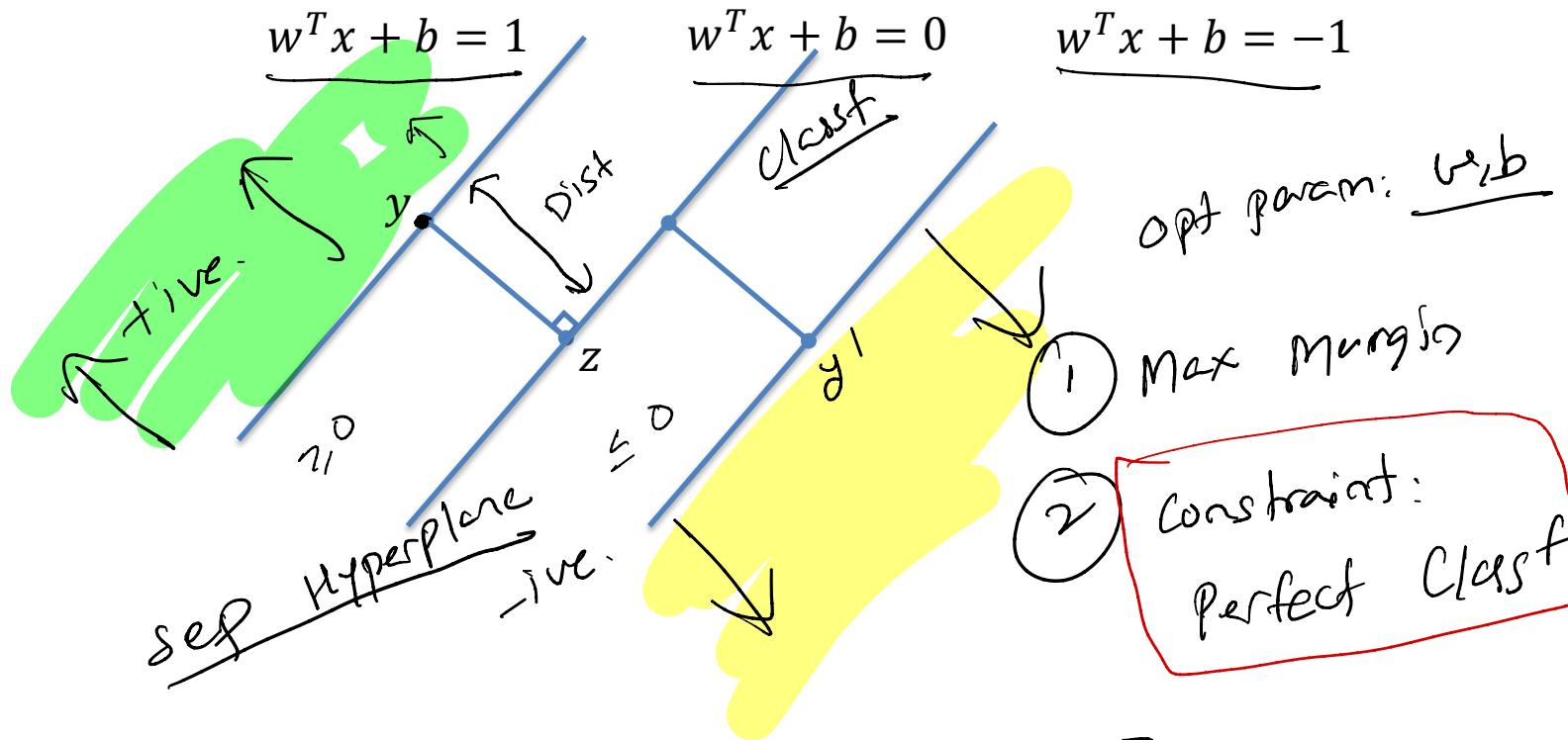
Scale Invariance



$$w^T x + b = w^T y + b = c$$

$$w' = \frac{w}{c}, \quad b' = \frac{b}{c}$$

Constraints

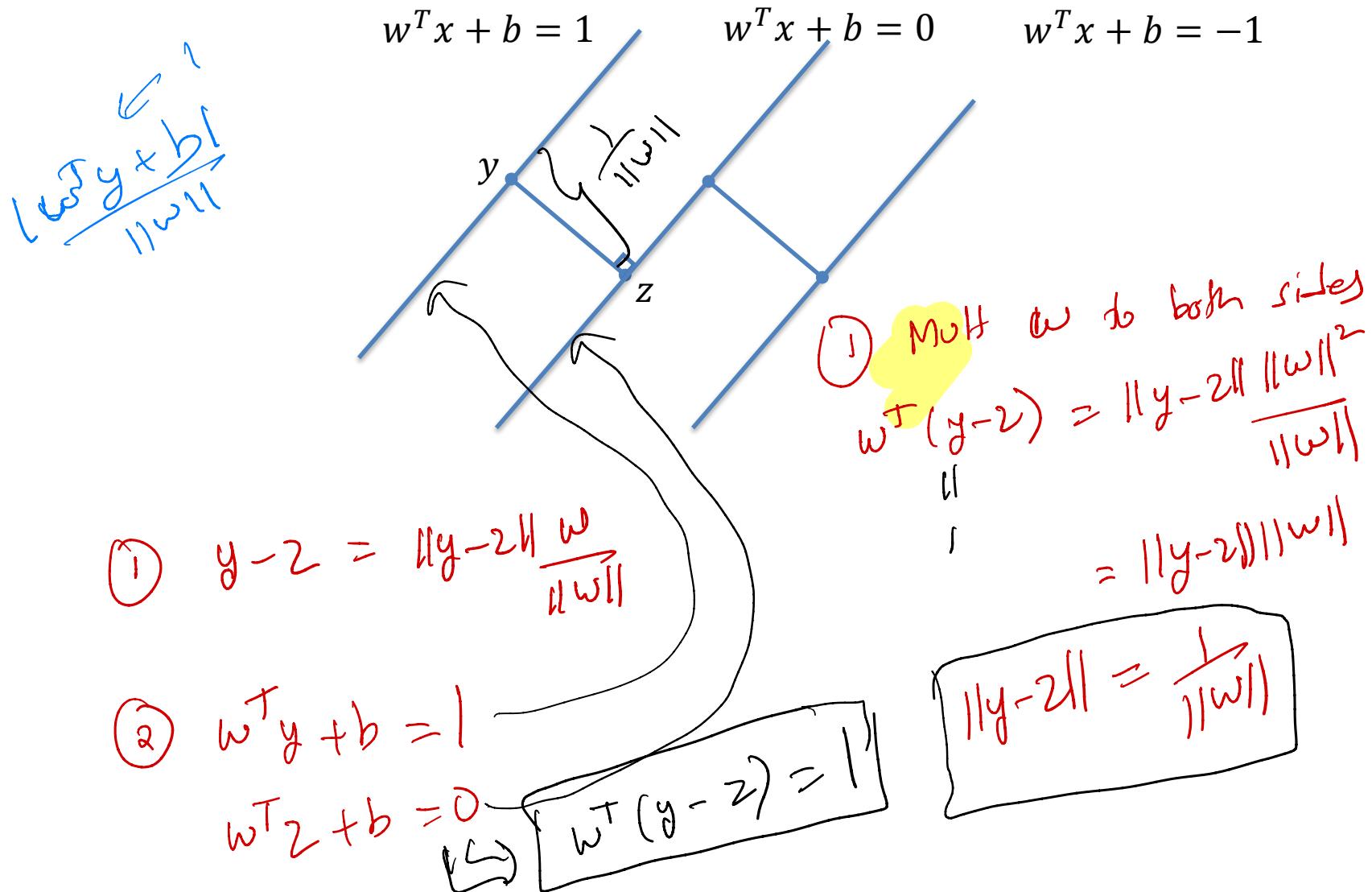


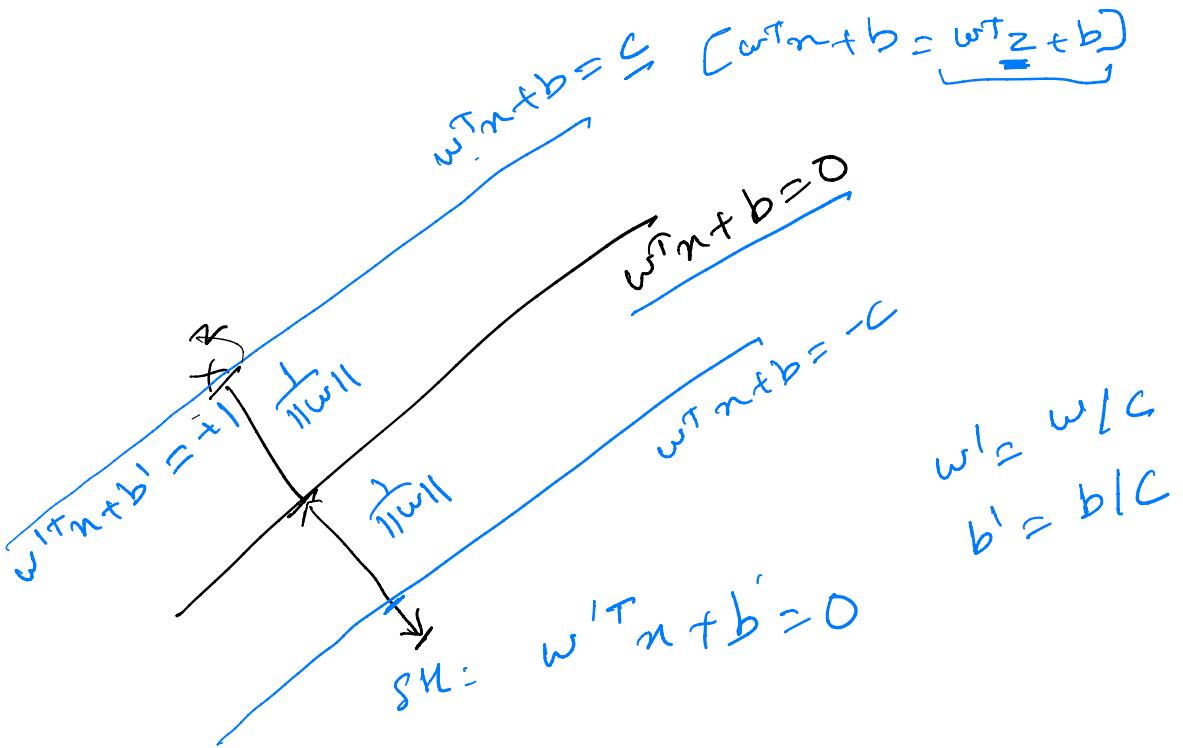
$$\# m = 1 : M \quad \# \text{Data} \quad y^{(m)} [w^T x^{(m)} + b] \geq 1$$

-1 $\rightarrow w^T x^{(m)} + b \leq -1$ ← -ives

+1 $\rightarrow w^T x^{(m)} + b \geq 1$ ← +ives.

What is the Margin?





$$\begin{aligned} w' &= w/C \\ b' &= b/C \end{aligned}$$

SVMs



- This analysis yields the following optimization problem

$$\max_{w,b} \frac{1}{\|w\|}$$

MARGIN

such that

$$y^{(i)}(w^T x^{(i)} + b) \geq 1, \text{ for all } i$$

[constraint]

- Or, equivalently,

$$\min_{w,b} \|w\|^2$$

Convex

Convex

such that

$$y^{(i)}(w^T x^{(i)} + b) \geq 1, \text{ for all } i$$

[Constrained opt problem]
Constrained Convex Opt Problem

↑
Perfect classifier
(0 error classif.)

SVMs



such that

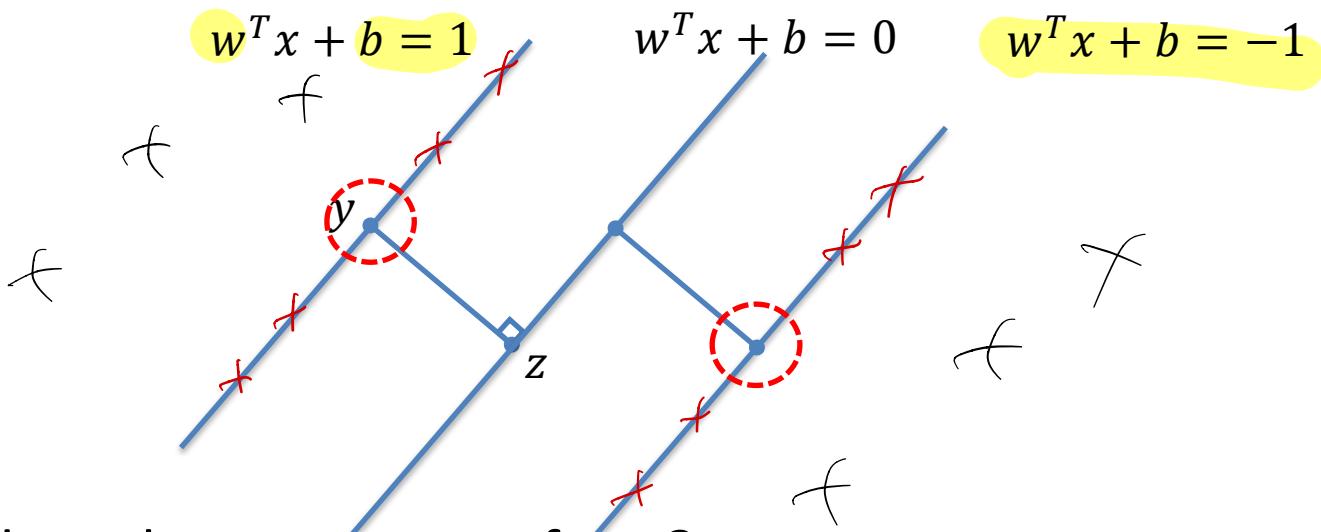
$$\min_{w,b} \|w\|^2 \quad \leftarrow \text{Objective.}$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1, \text{ for all } i \quad \leftarrow \text{Constraint.}$$

- This is a standard quadratic programming problem
 - Falls into the class of convex optimization problems
 - Can be solved with many specialized optimization tools (e.g., quadprog() in MATLAB)

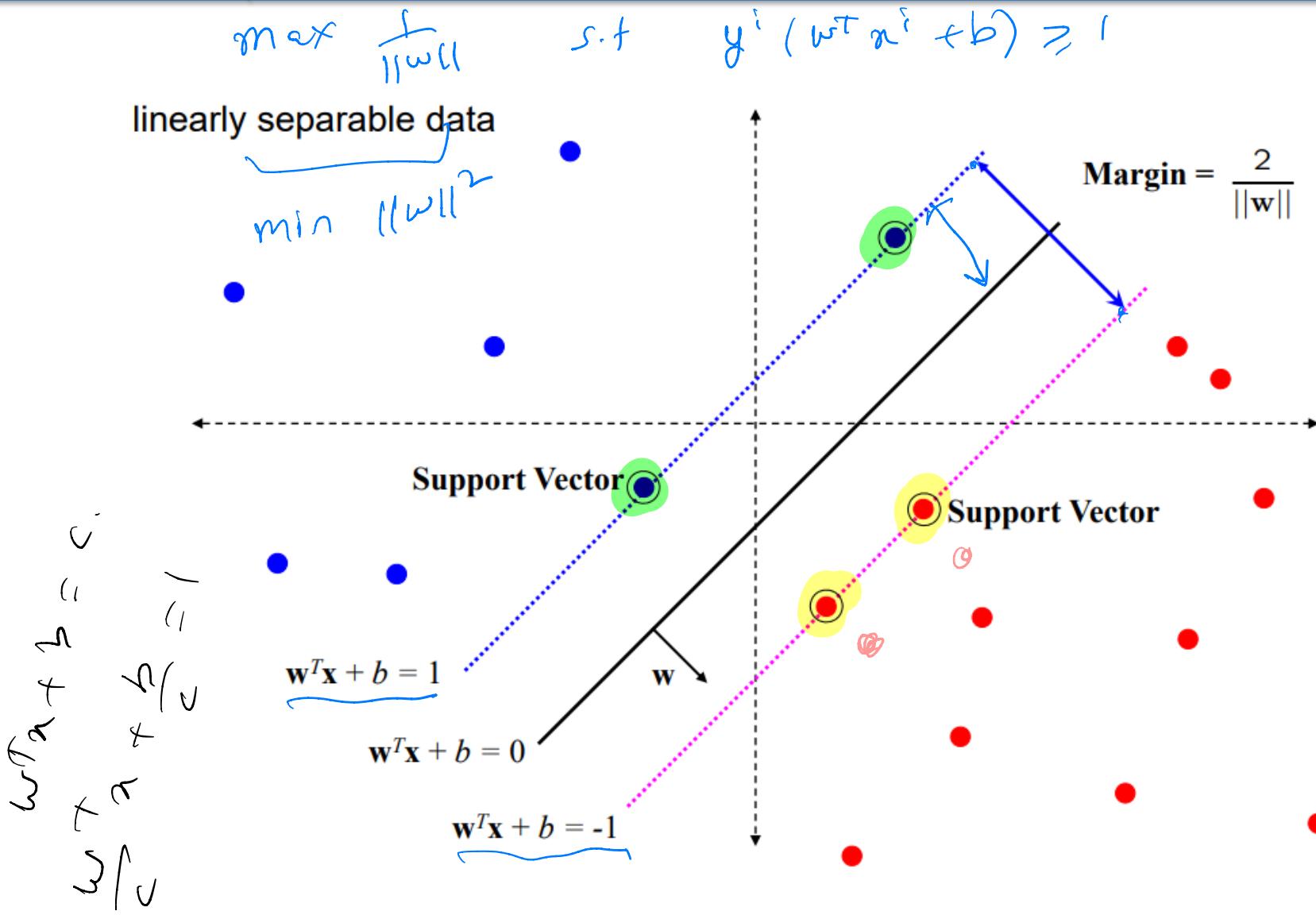
$$\sum_{i=1}^M L(f(x_i, w, b), y_i)$$

Support Vectors



- Where does the name come from?
 - The set of all data points such that $y^{(i)}(w^T x^{(i)} + b) = 1$ are called **support vectors**
 - The SVM classifier is completely determined by the support vectors (you could delete the rest of the data and get the same answer)

Putting Everything Together



SVMs

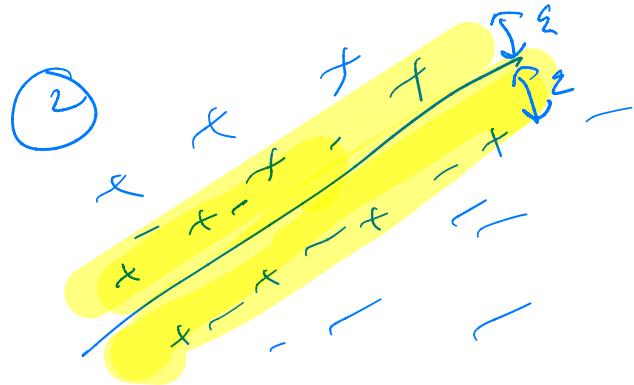


$$w^T \phi(x) + b$$

- What if the data isn't linearly separable?



$$\phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{bmatrix}$$



- What if we want to do more than just binary classification (i.e., if $y \in \{1,2,3\}$)?

SVMs



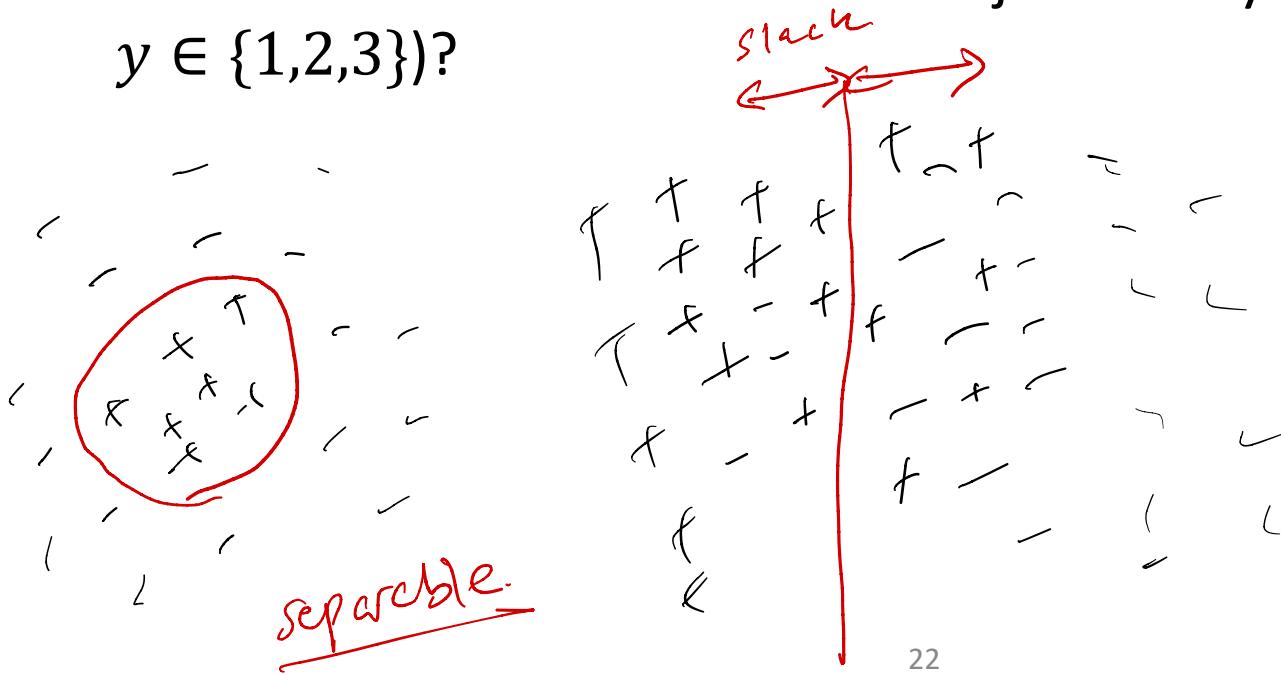
- What if the data isn't linearly separable?

- Higher order (polynomial features)

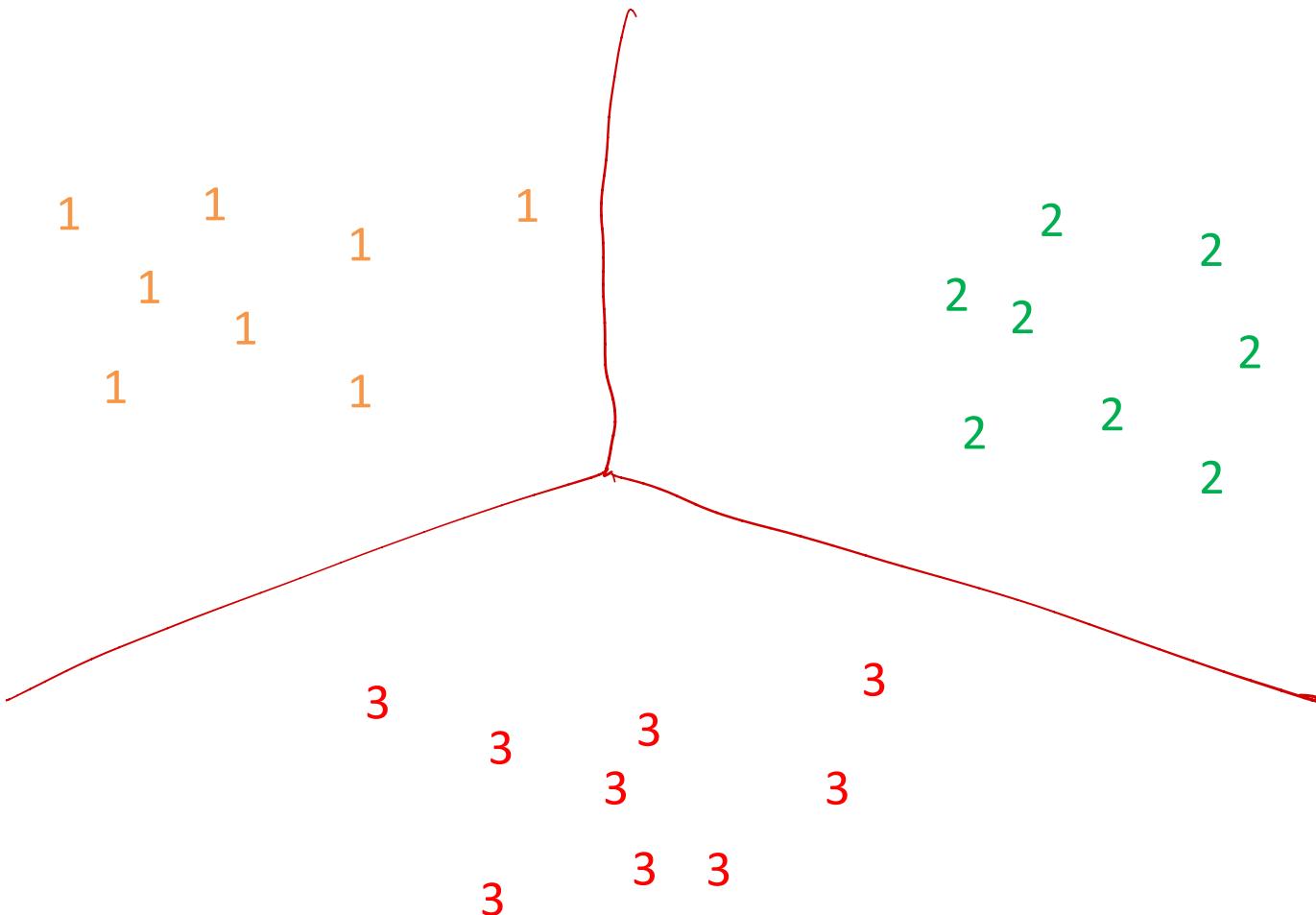
Linear Regression
Perception

- Relax the constraints (coming soon)

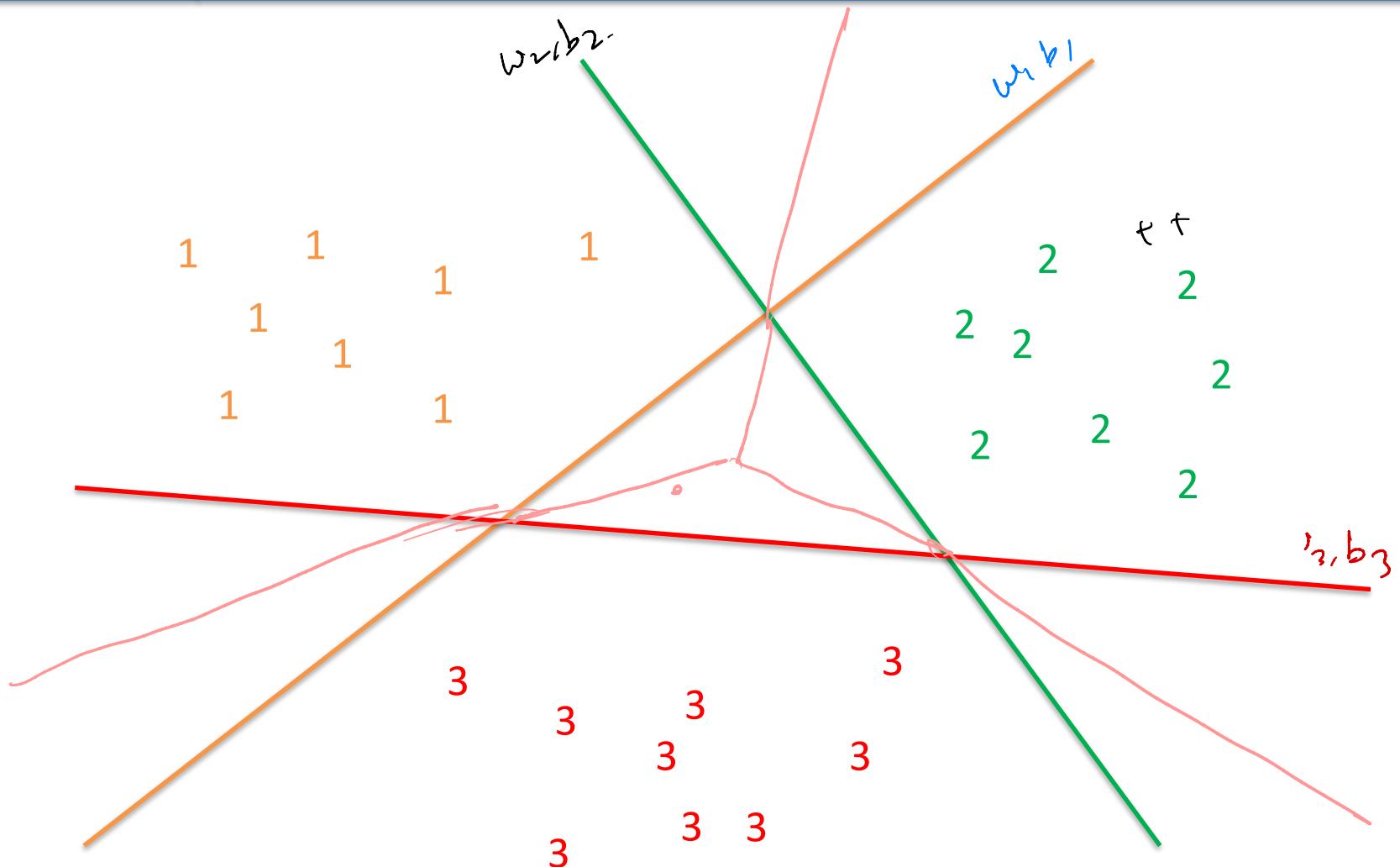
- What if we want to do more than just binary classification (i.e., if $y \in \{1,2,3\}$)?



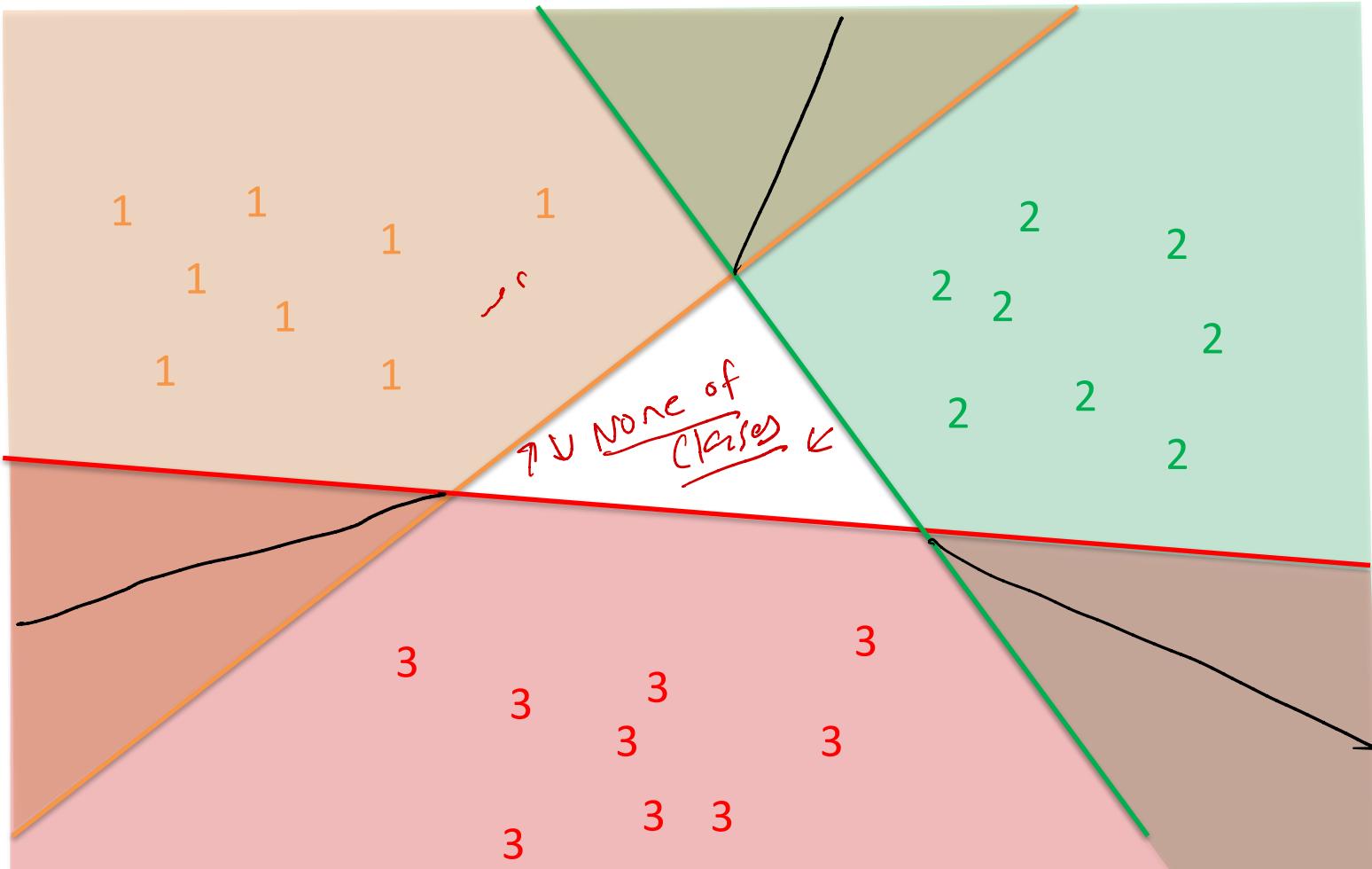
Multiclass Classification



One-Versus-All SVMs



One-Versus-All SVMs



Regions correctly classified by exactly one classifier

One-Versus-All SVMs

- Compute a classifier for each label versus the remaining labels (i.e., an SVM with the selected label as plus and the remaining labels changed to minuses)

- Let $f^k(x) = w^{(k)^T}x + b^{(k)}$ be the classifier for the k^{th} label

- For a new datapoint x , classify it as

$$k' \in \operatorname{argmax}_k f^k(x)$$

$n:$

$$\begin{aligned}f_1 &: w_1 \cdot x + b_1 \\f_2 &: w_2 \cdot x + b_2 \\f_3 &: w_3 \cdot x + b_3\end{aligned}$$

- Drawbacks:

- If there are L possible labels, requires learning L classifiers over the entire data set

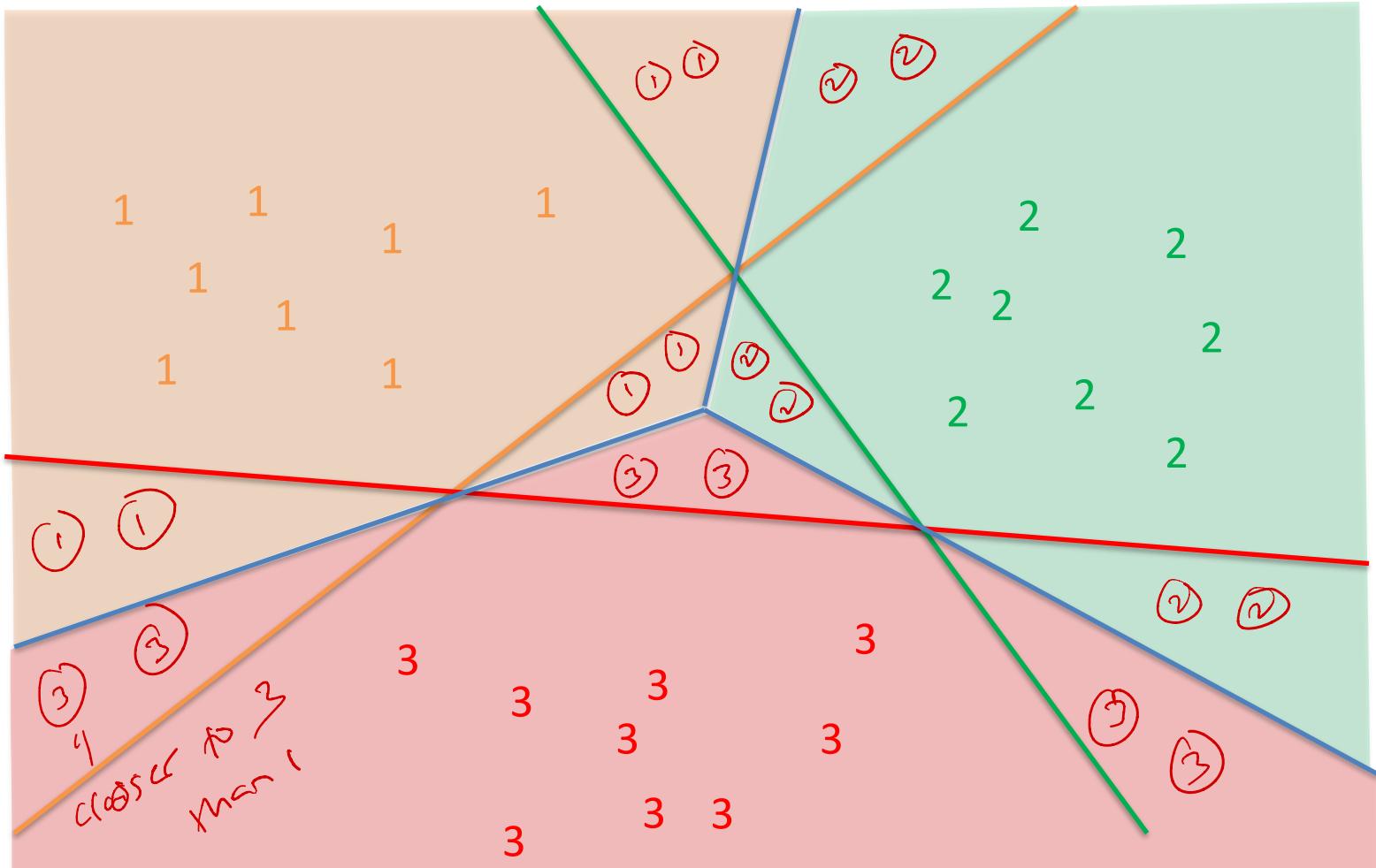
$$L = 10^0$$

$$\text{cost} \approx L^{\frac{N \cdot n}{D} \text{feat}}$$

Extreme Classif⁰

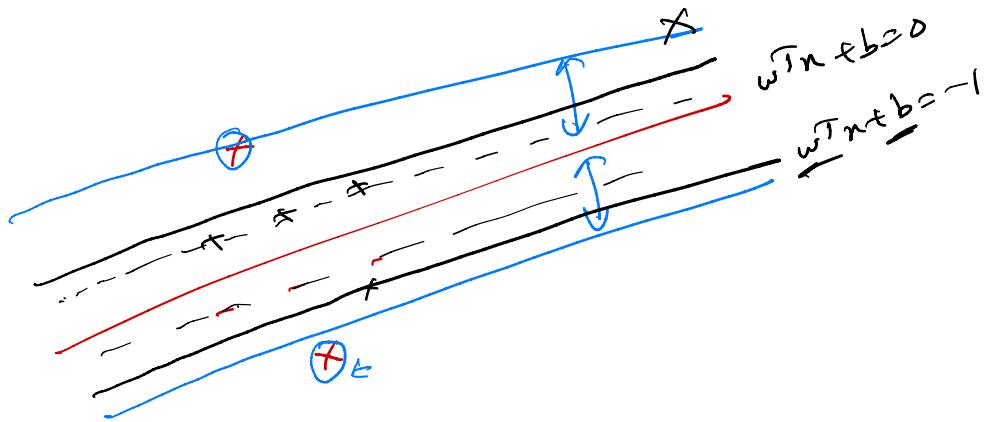
$$L = 10^6$$

One-Versus-All SVMs



Regions in which points are classified by highest value of $w^T x + b$

$$\arg \max_k w^{(k)T} x + b^{(k)}$$



One-Versus-One SVMs (Class)



- Alternative strategy is to construct a classifier for all possible pairs of labels
- Given a new data point, can classify it by majority vote (i.e., find the most common label among all of the possible classifiers)
- If there are L labels, requires computing $\binom{L}{2}$ different classifiers each of which uses only a fraction of the data
- Drawbacks: Can overfit if some pairs of labels do not have a significant amount of data (plus it can be computationally expensive)

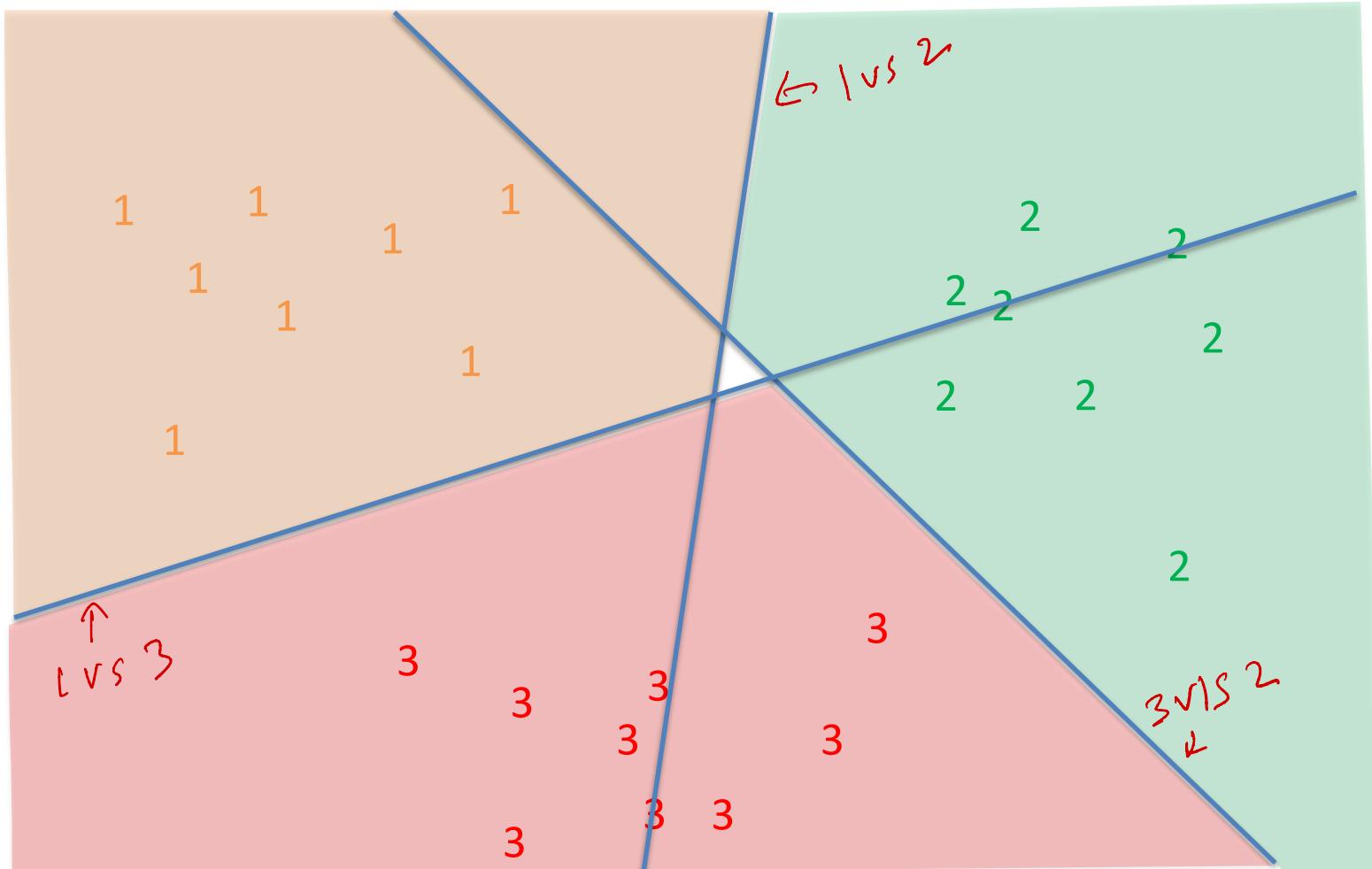
$$\frac{2M}{L}$$

$$\binom{L}{2} = \frac{L(L-1)}{2}$$

1, 2, ..., k
1 v 2, 1 v 3, 1 v 4, ..., $v k$
2 v 3, ..., $v k$

$\binom{k}{2}$

One-Versus-One SVMs



Regions determined by majority vote over the classifiers

