Support Vector Machines

Nicholas Ruozzi
University of Texas at Dallas

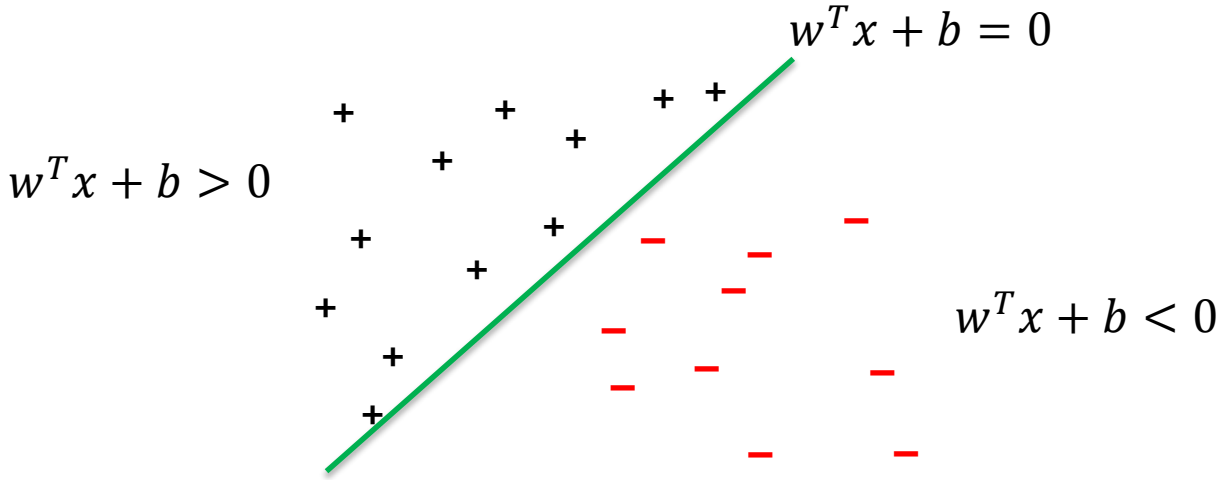
Slides adapted from David Sontag and Vibhav Gogate

Announcements

- Homework 1 is now available online
- Join the Piazza discussion group
- Reminder: my office hours are 11am-12pm on Tuesdays

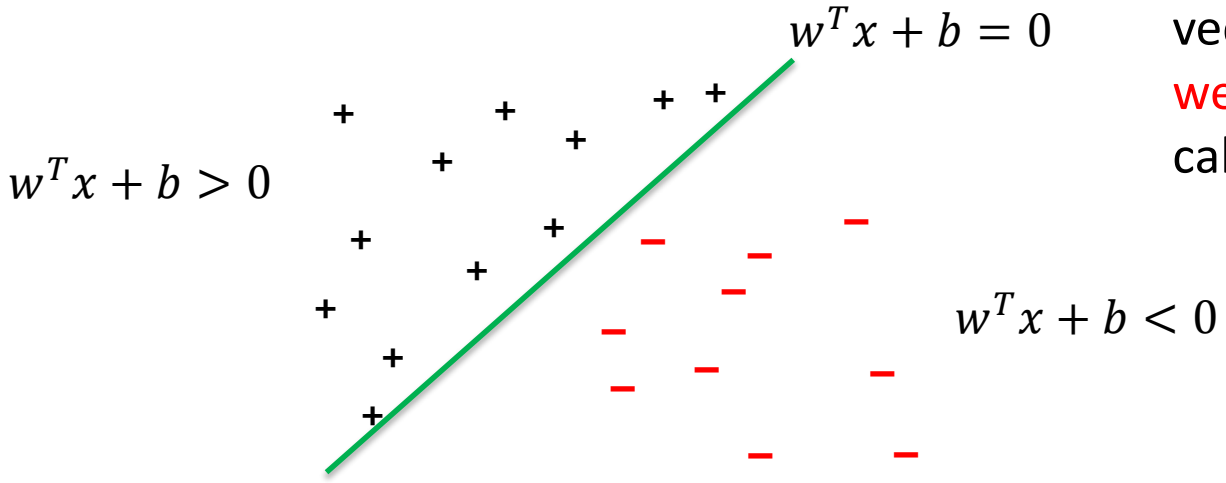
Binary Classification

- Input $\left(x^{(1)}, y_1\right), \ldots, \left(x^{(n)}, y_n\right)$ with $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$
- We can think of the observations as points in $\mathbb{R}^m$ with an associated sign (either +/- corresponding to 0/1)
- An example with $m = 2$



Binary Classification

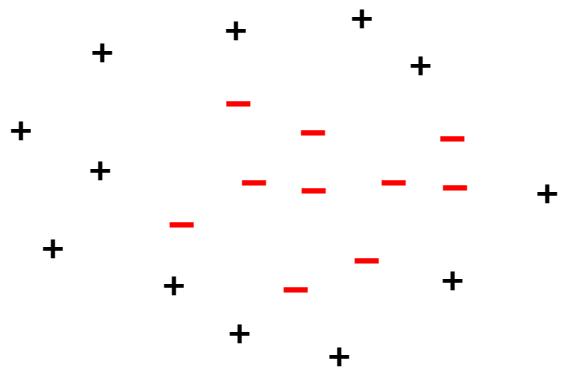
- Input $\left(x^{(1)}, y_1\right), \ldots, \left(x^{(n)}, y_n\right)$ with $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$
- We can think of the observations as points in $\mathbb{R}^m$ with an associated sign (either +/- corresponding to 0/1)
- An example with $m = 2$



w is called the vector of weights and b is called the bias

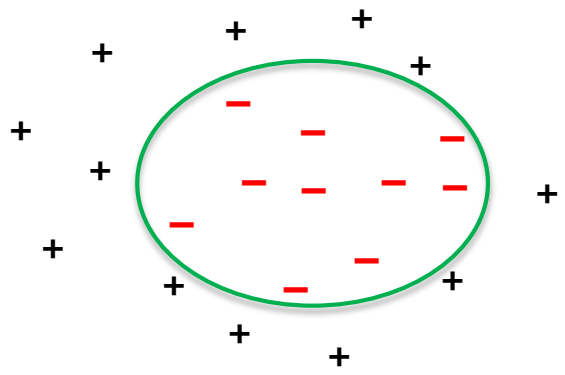
What If the Data Isn‘t Separable?

- Input $\left(x^{(1)}, y_1\right), \ldots, \left(x^{(n)}, y_n\right)$ with $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$
- We can think of the observations as points in $\mathbb{R}^m$ with an associated sign (either +/- corresponding to 0/1)
- An example with $m = 2$



What If the Data Isn‘t Separable?

- Input $\left(x^{(1)}, y_1\right), \ldots, \left(x^{(n)}, y_n\right)$ with $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$
- We can think of the observations as points in $\mathbb{R}^m$ with an associated sign (either +/- corresponding to 0/1)
- An example with $m = 2$



Adding Features

- The idea:
  - Given the observations $x^{(1)}, \ldots, x^{(n)}$, construct a feature vector $\phi(x)$
  - Use $\phi\left(x^{(1)}\right), \ldots, \phi\left(x^{(n)}\right)$ instead of $x^{(1)}, \ldots, x^{(n)}$ in the learning algorithm
  - Goal is to choose ϕ so that $\phi\left(x^{(1)}\right), \ldots, \phi\left(x^{(n)}\right)$ are linearly separable
  - Learn linear separators of the form $w^T \phi(x)$ (instead of $w^T x$)

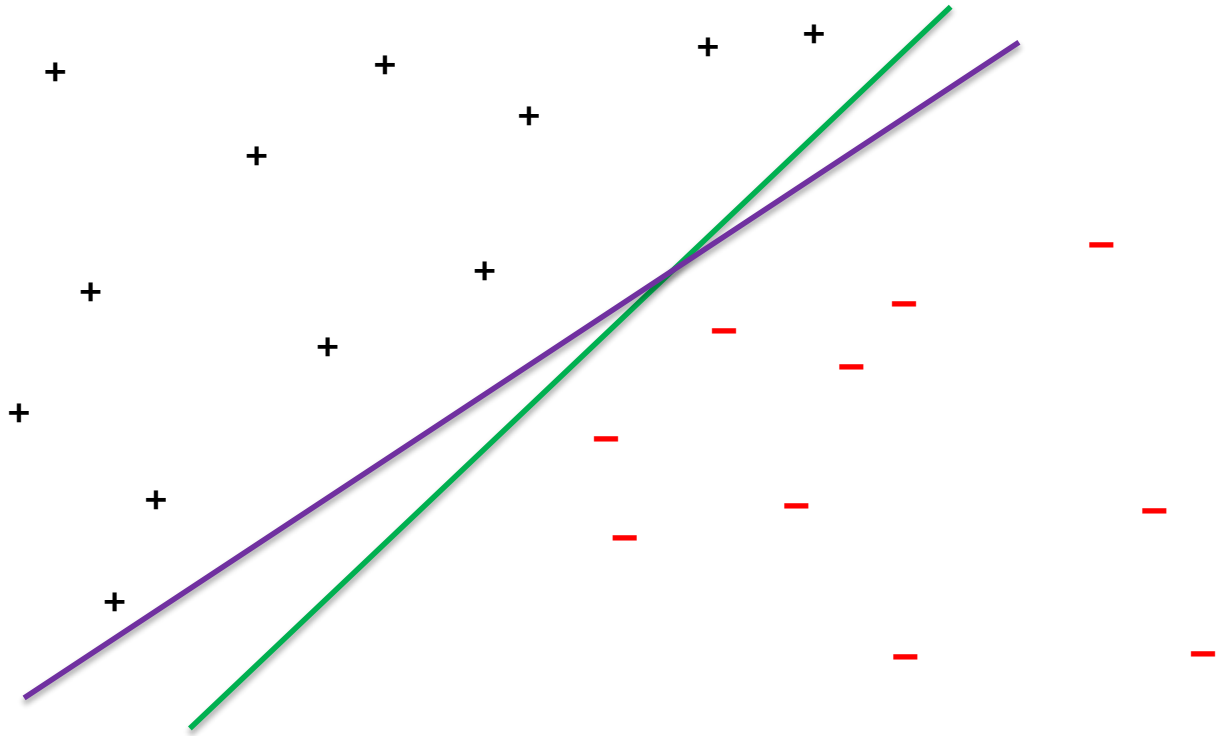
Adding Features

- Sometimes it is convenient to group the bias together with the weights
- To do this
  - Let $\phi(x_1, x_2) = \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$ and $\tilde{w} = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix}$,
  - This gives

$$\tilde{w}^T \phi(x_1, x_2) = w_1 x_1 + w_2 x_2 + b = w^T x + b$$

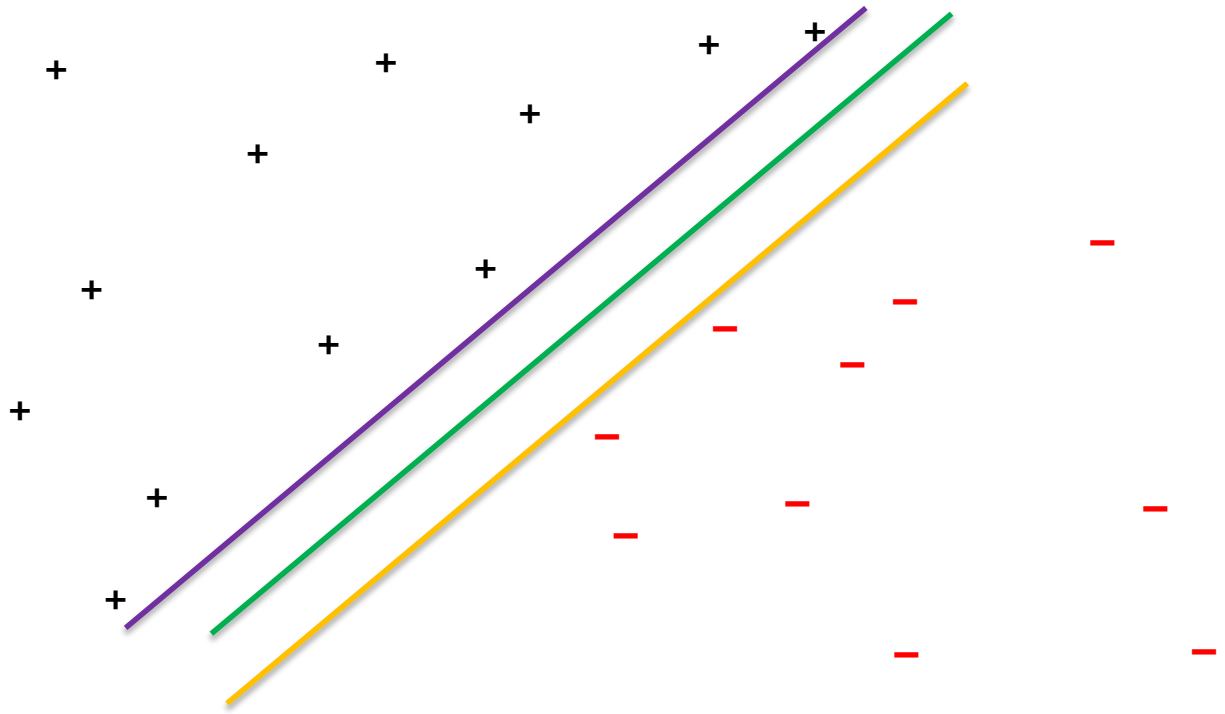
Support Vector Machines

- How can we decide between perfect classifiers?



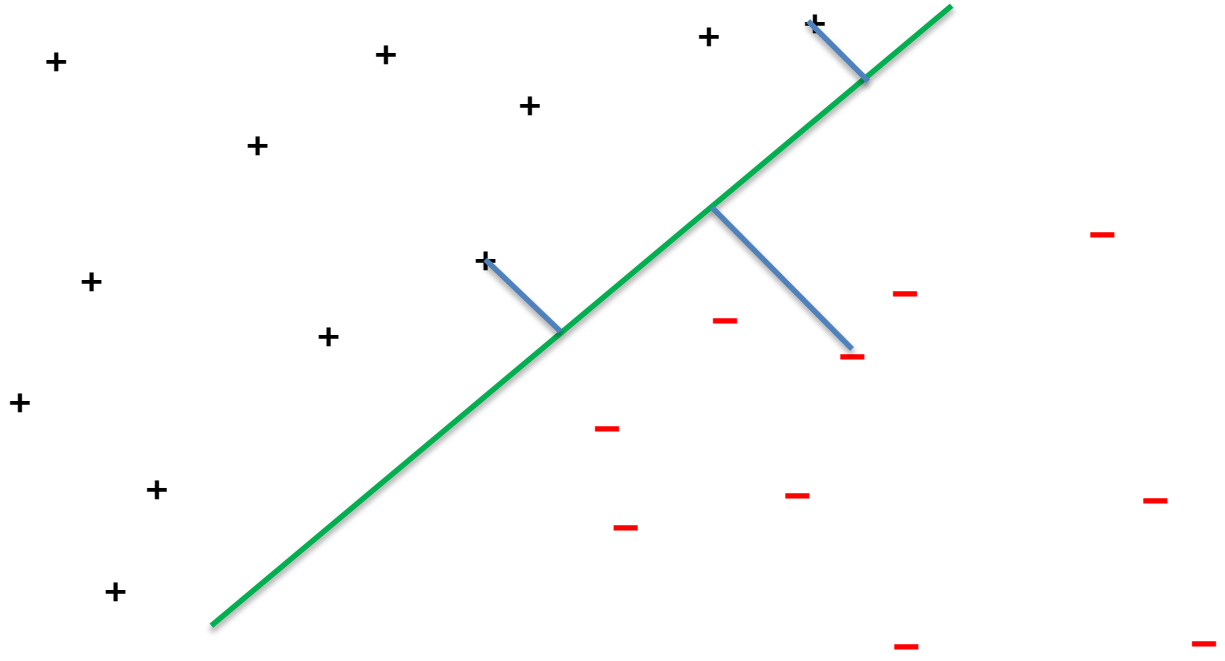
Support Vector Machines

- How can we decide between perfect classifiers?



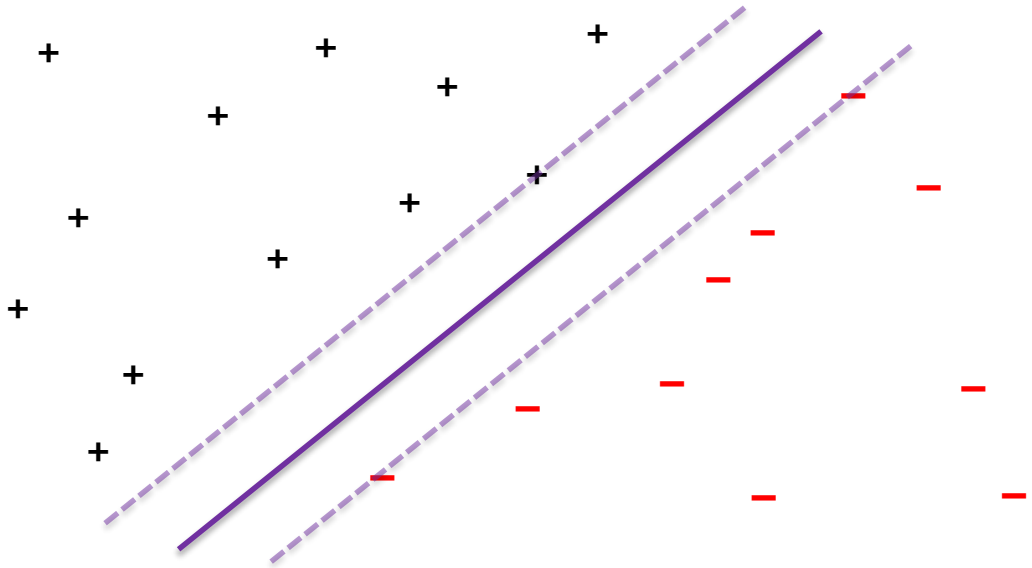
Support Vector Machines

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



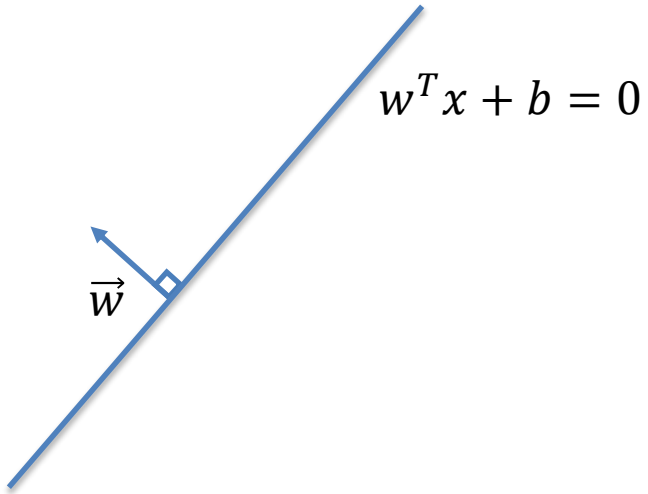
Support Vector Machines

- Support vector machines (SVMs)



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

Some Geometry



- In n dimensions, a hyperplane is a solution to the equation

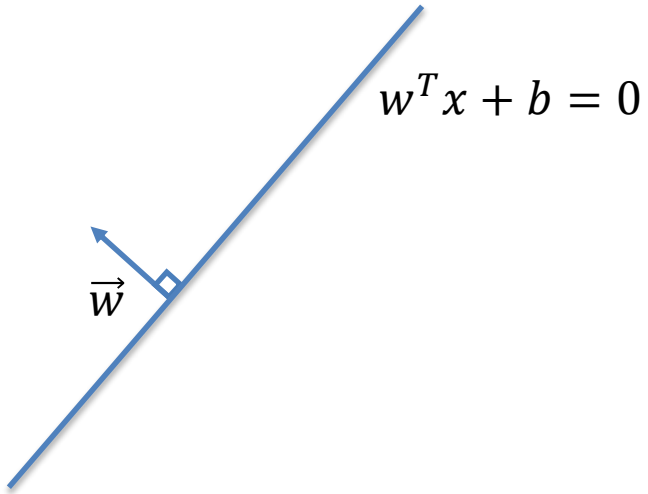
$$w^T x + b = 0$$

with $w \in \mathbb{R}^n, b \in \mathbb{R}$

- The vector w is sometimes called the normal vector of the hyperplane

UTD

Some Geometry



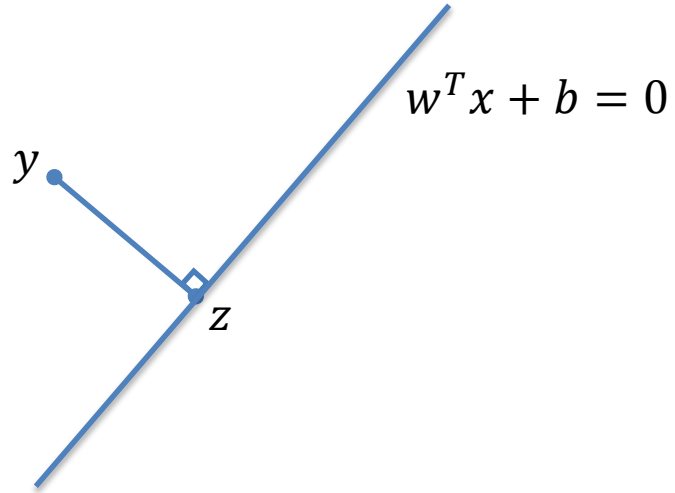
- In n dimensions, a hyperplane is a solution to the equation

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

- Note that this equation is scale invariant for any scalar c

$$c \cdot (w^T x + b) = 0$$

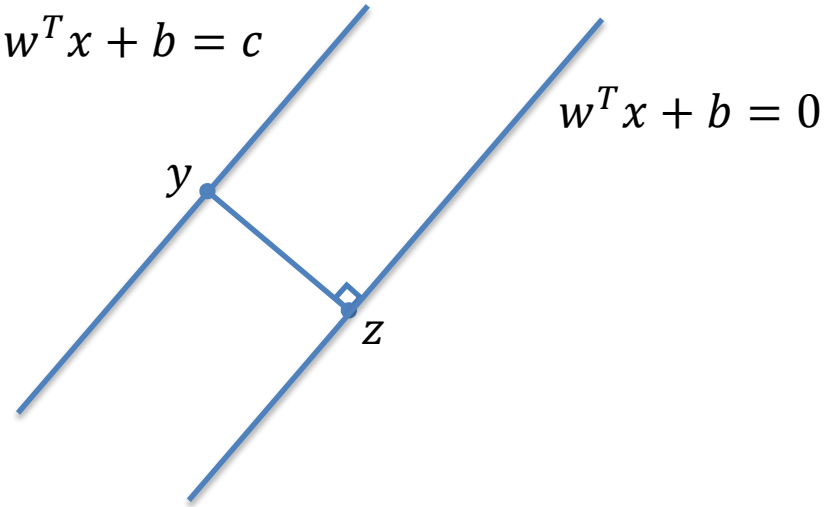
Some Geometry



- The distance between a point y and a hyperplane $w^T + b = 0$ is the length of the vector perpendicular to the line through the point y

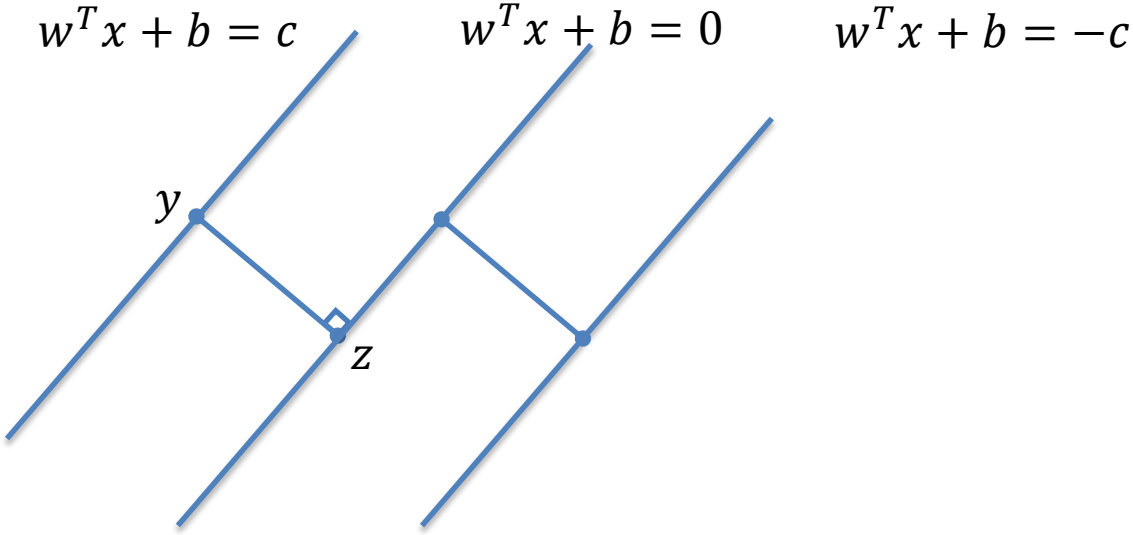
$$y - z = \|y - z\| \frac{w}{\|w\|}$$

Scale Invariance



- By scale invariance, we can assume that $c = 1$
- The maximum margin is always attained by choosing $w^T x + b = 0$ so that it is equidistant from the closest data point classified as +1 and the closest data point classified as -1

Scale Invariance

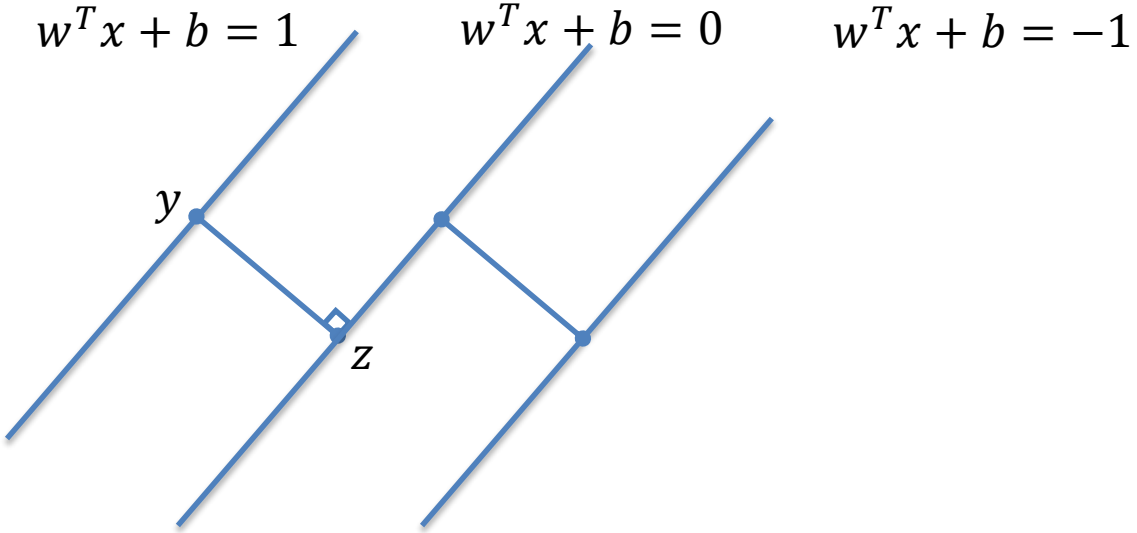


- We want to maximize the margin subject to the constraints that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1$$

- But how do we compute the size of the margin?

Some Geometry



Putting it all together

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

and

$$w^T y + b = 1$$
$$w^T z + b = 0$$

$$w^T (y - z) = 1$$

and

$$w^T (y - z) = \|y - z\| \|w\|$$

which gives

$$\|y - z\| = 1/\|w\|$$

SVMs

- This analysis yields the following optimization problem

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

such that

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

- Or, equivalently,

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

such that

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

SVMs

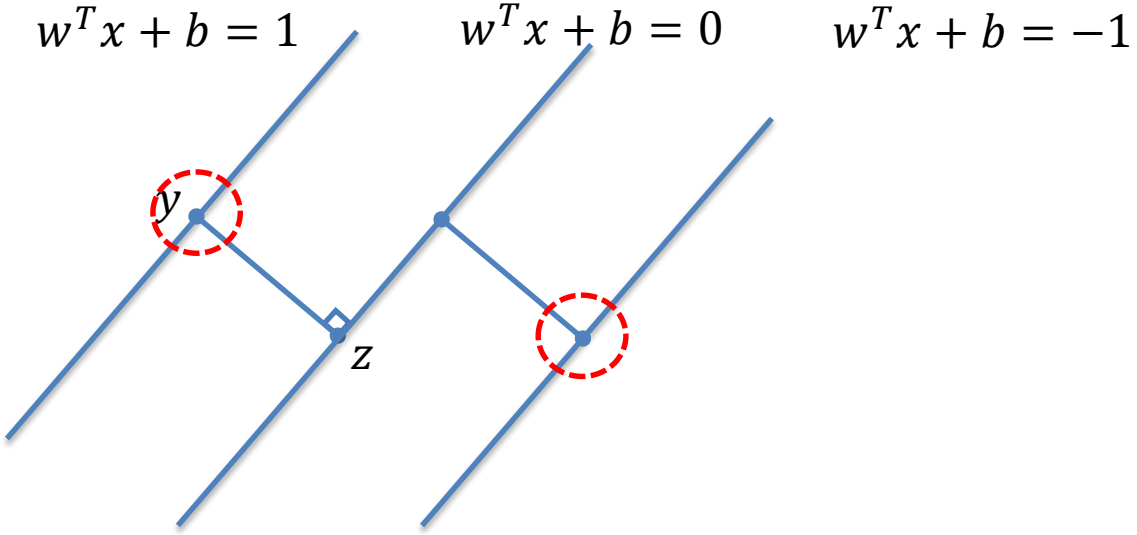
$$\min_{w} \|w\|^2$$

such that

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

- 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)

SVMs



- 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

SVMs

- What if the data isn't linearly separable?
  - Use feature vectors
- What if we want to do more than just binary classification (i.e., if $y \in \{1,2,3\}$)?
  - One versus all: for each class, compute a linear separator between this class and all other classes
  - All versus all: for each pair of classes, compute a linear separator