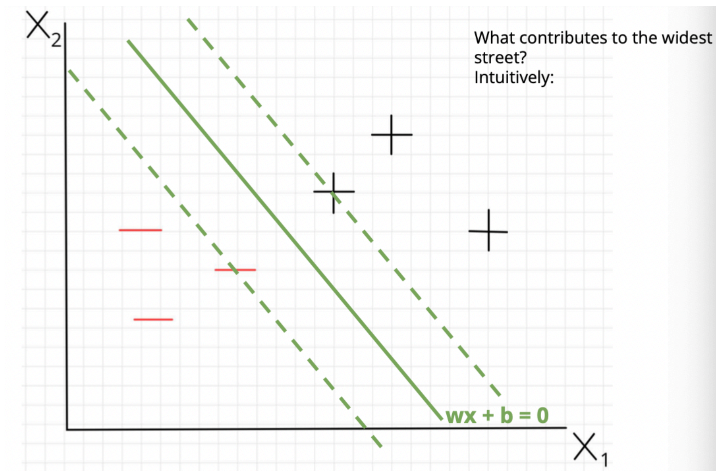
CAS CS 506
Lec 15

Support Vector Machine (cont.)

1. Continuation
   a. 

   b. What contributes to the widest street?
      i. The “+” that is on the dotted line
      ii. That point is called a support vector

2. Find the Widest Street Subject to…
   a. We want our samples to lie beyond the street. That is

$$\vec{w} \cdot \vec{x}_+ + b \geq 1$$

$$\vec{w} \cdot \vec{x}_- + b \leq -1$$

   b. Note: for an unknown u, we can have

$$-1 < \vec{w} \cdot \vec{u} + b < 1$$

   c. Let’s introduce a variable

$$y_i = \begin{cases} +1 & \text{if } x_i \text{ is a } + \text{sample} \\ -1 & \text{if } x_i \text{ is a } - \text{sample} \end{cases}$$

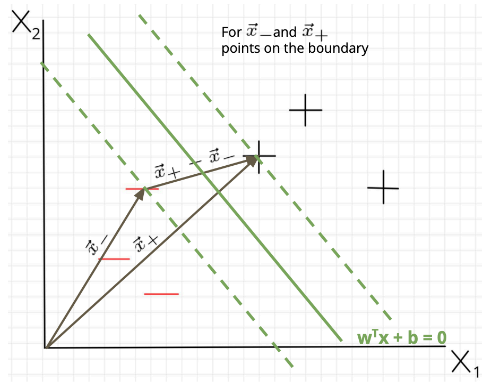
   d. Note: this is effectively the class label of x_i

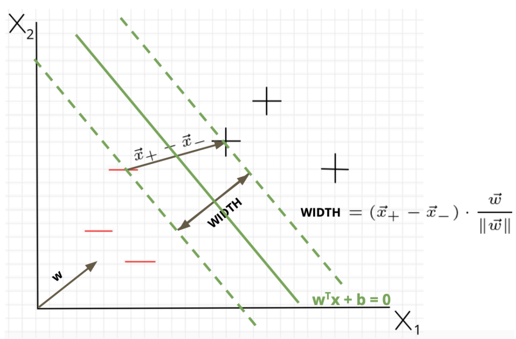
e. If we multiply our sample decision rules by this new variable:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1$$

f. Meaning, for xi on the decision boundary, we want:

$$y_i(\vec{w} \cdot \vec{x}_i + b) - 1 = 0$$

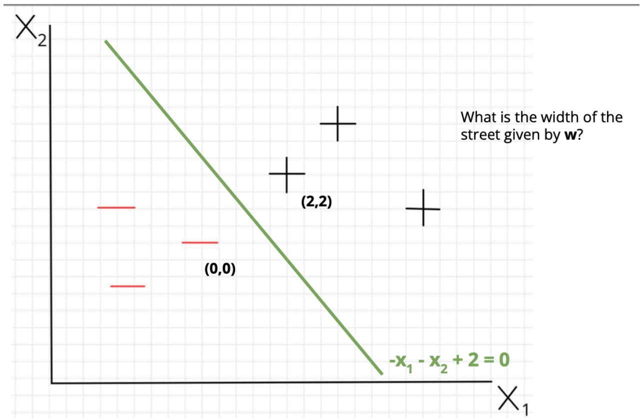
g. 

h. 

3. How to Find the Width of the Street
   a. We know that $\textbf{WIDTH} = (\vec{x}_+ - \vec{x}_-) \cdot \frac{\vec{w}}{\|\vec{w}\|}$ for $\vec{x}_-$ and $\vec{x}_+$ points on the boundary
   b. Since they are on the boundary, we know that

$$y_i(\vec{w} \cdot \vec{x}_i + b) - 1 = 0$$

   c. Hence, $\textbf{WIDTH} = \frac{2}{\|\vec{w}\|}$
4. What does that Mean?
   a. Size of w is inversely proportional to the width of the street
   b. Aligns with what we found previously

5. Example
    a. 

    b. Width = 2 / sqrt(2) = sqrt(2)
6. How to Find the Widest Street
    a. Goal is to maximize the width

$$\max\left(\frac{2}{\|\vec{w}\|}\right)$$

    b. Subject to:

$$y_i(\vec{w} \cdot \vec{x}_i + b) - 1 = 0$$

    c. Can use Lagrange multipliers to form a single expression to find the extremum of

$$L = \frac{1}{2}\|\vec{w}\|^2 - \sum_i \alpha_i \left[y_i(\vec{x}_i \cdot \vec{w} + b) - 1\right]$$

Where α_i is 0 for x_i not on the boundary.

    d. Take the partial derivative of L wrt to w to see what w looks like at the extremum of L
    e.
$$\frac{\partial L}{\partial \vec{w}} = \vec{w} - \sum_i \alpha_i y_i \vec{x}_i = 0$$
$$\implies \vec{w} = \sum_i \alpha_i y_i \vec{x}_i$$

    f. Means w is a linear sum of vectors in our sample/training set

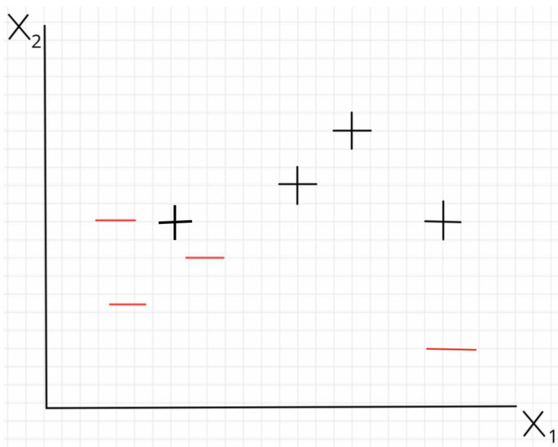
g. $$\boxed{\sum_i \alpha_i < x_i, x > +b \geq 0 \quad \text{then } +}$$

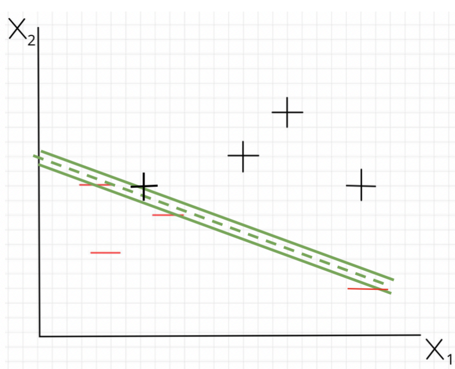
7. To Move the Street in the Direction of a Point

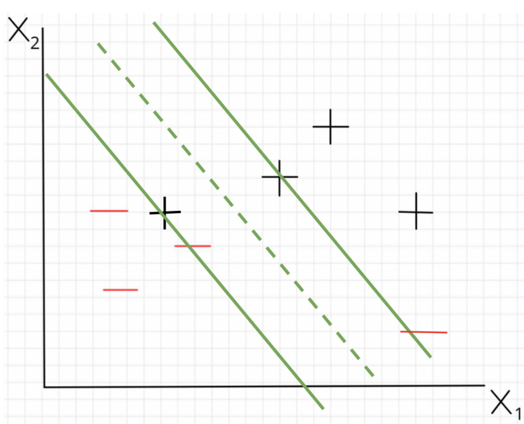
   a. $$\alpha_{i,new} = \alpha_{i,\,old} + y_i * a$$

   $$b_{new} = b_{old} + y_i * a$$

8. Trade-off Between Width and Error

   a. 

   b. 

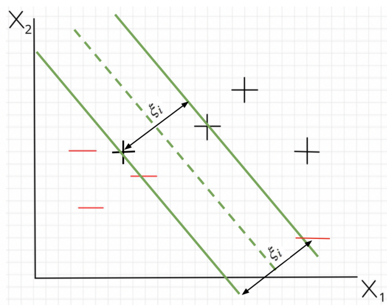
   c. 

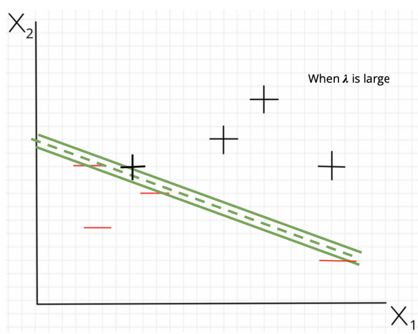
9. How to Find the Widest Street
    a. Goal is to maximize the width

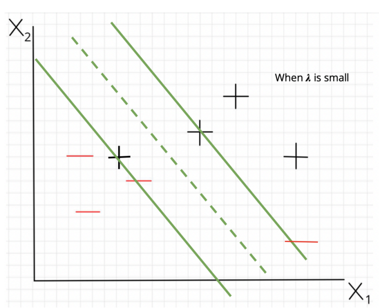
$$min(\frac{1}{2}\|w\|^2 + \lambda \sum_i \xi_i)$$

    b. Subject to:

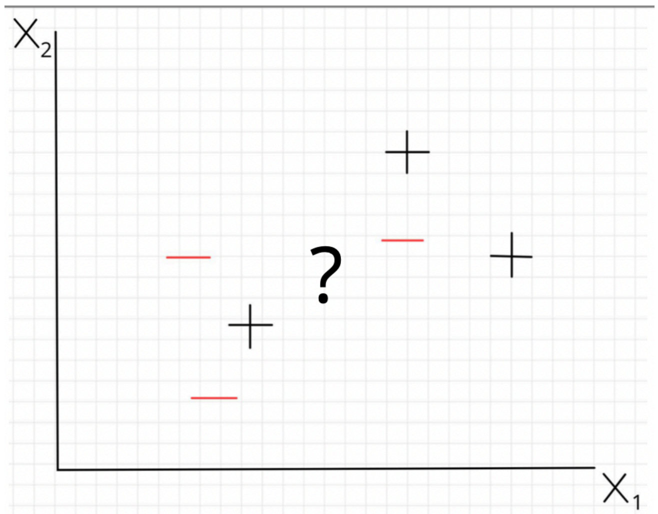
$$y_i(\vec{w}.\vec{x}_i + b) \geq 1 - \xi_i$$

    c. 

    d. 

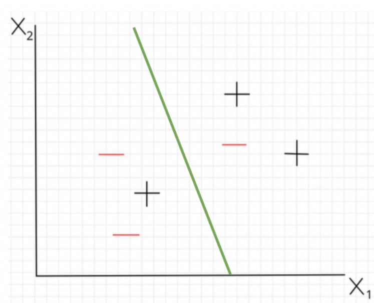
    e. 

10. What if There is No Line?

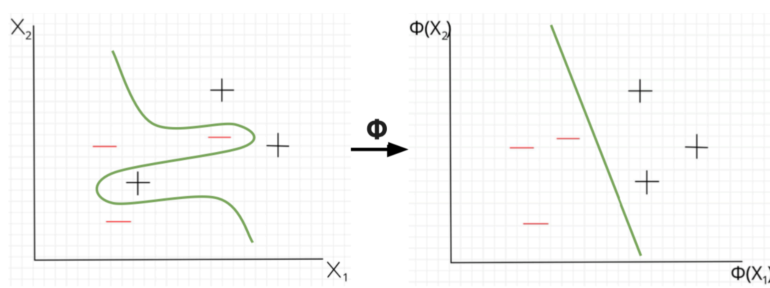
a. 

b. Option 1: Soft Margins

   i. Can allow for some points in the dataset to be misclassified

   ii. 

c. Option 2: Change Perspective

   i. 

11. But How to Find O?

a. Turns out we don't need to find or define a transformation Φ

b. Recall:

$$\boxed{\sum_i \alpha_i < x_i, x > + b \geq 0 \quad \text{then } +}$$

c. We only need to define

$$K(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$$

d. Called a Kernel function. This is often referred to as the “kernel trick”

$$\boxed{\sum_i \alpha_i K(x_i, x) + b \geq 0 \quad \text{then } +}$$

12. Example Kernel Functions

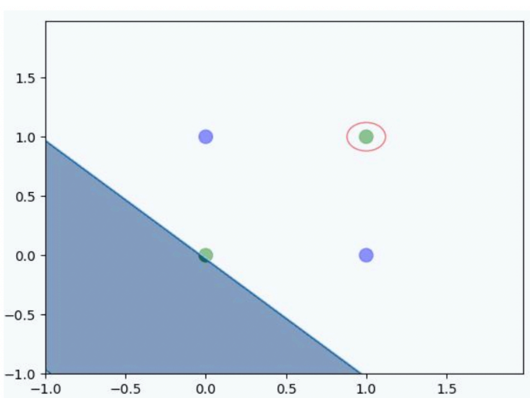
Polynomial Kernel

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + 1)^n$$

Radial Basis Function Kernel

$$K(\vec{x}_i, \vec{x}_j) = e^{\frac{\left\|\vec{x}_i - \vec{x}_j\right\|}{\sigma}}$$

a.

b. 

13. Kernel Function (Intuition)
   a. The inner product of a space describes how close/similar points are
   b. Kernel functions allow for specifying the closeness / similarity of points in a hypothetical transformed space
   c. The hope is that with that new notion of closeness, points in the dataset are linearly separable