Lecture 10

Optimal Margin Classifiers and SVM

Physical Address for Prof. Ilias Tagkopoulos
Computer Science:
Office: 3063 Kemper Hall
Phone: (530) 752-4821
Fax: (530) 752-4767

Genome and Biomedical Sciences Facility:
Office: 5313 GBSF
Phone: (530) 752-7707
Fax: (530) 754-9658

Instructor: Ilias Tagkopoulos
iliast@ucdavis.edu

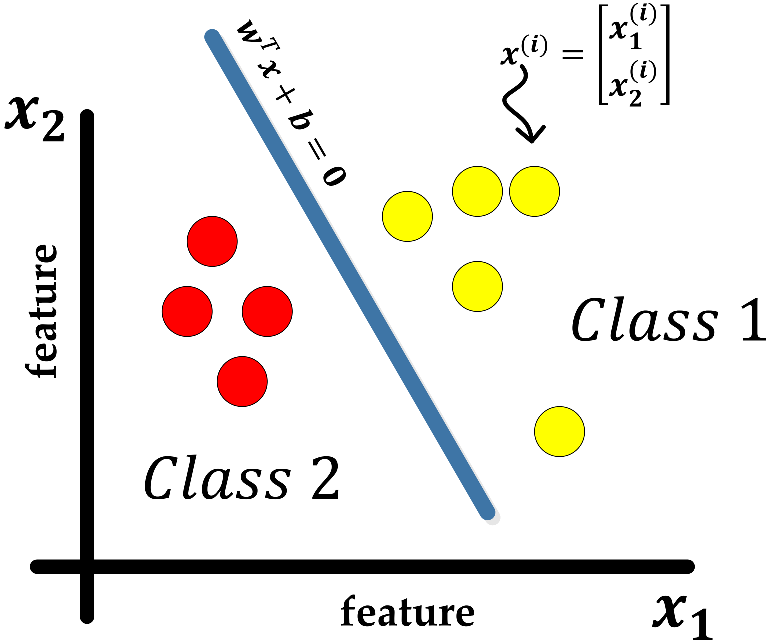
Classification: the quest for the optimal boundary

- In classification our task is to create a method that is able to accurately categorize new samples.

- We want to define a line (hyperplane, boundary) that separates the different classes.

- Many possibilities…which one is the best?

- Margin classifiers answer this question, by selecting the line that has the maximum distance from one or more samples from each class.



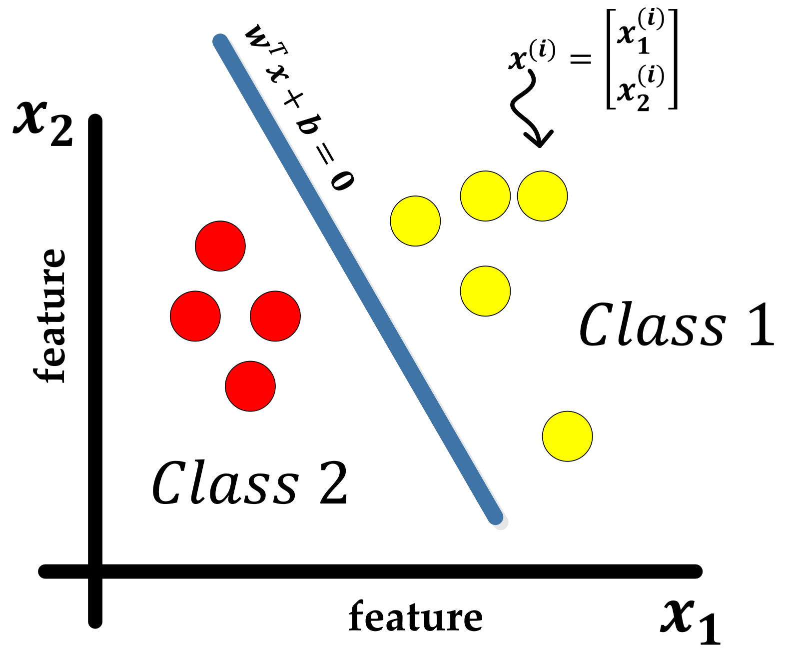
Assume 2 classes here but the same applies when there are K classes

Lines and Vectors

- The decision boundary (hyperplane) will have the form:

$$\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{0}$$

- All points were $y(x) = w^T x^{(i)} + b > 0$ will be class 1, all points with $y(x) = w^T x^{(i)} + b < 0$ will be class 2.



- The vector w is perpendicular to the decision boundary. Indeed, take any two points x_A and x_B on the boundary and the following should hold: $\boldsymbol{w}^T\boldsymbol{x}_A + \boldsymbol{b} = \boldsymbol{w}^T\boldsymbol{x}_B + \boldsymbol{b} = \boldsymbol{0}$

$$\boldsymbol{w}^T(\boldsymbol{x}_A - \boldsymbol{x}_B) = \boldsymbol{0}$$

The inner product of the vector $\boldsymbol{w}$ with any vector $(\boldsymbol{x}_A - \boldsymbol{x}_B)$ on the boundary is zero, hence the vector $\boldsymbol{w}$ is perpendicular to it.

Geometric margin

- Take any sample (i). Then we can write its input vector $x^{(i)}$ as a function of its projection $x_P^{(i)}$ and its distance γ to the boundary:

$$x^{(i)} = x_P^{(i)} + \gamma \frac{w}{\|w\|} y^{(i)}$$

- Or

$$\frac{\|w\|}{y^{(i)} w}(x^{(i)} - x_P^{(i)}) = \gamma \Rightarrow \gamma = \frac{w^T}{y^{(i)}\|w\|}(x^{(i)} - x_P^{(i)})$$

- Also, $x_P^{(i)}$ is in the boundary, so:

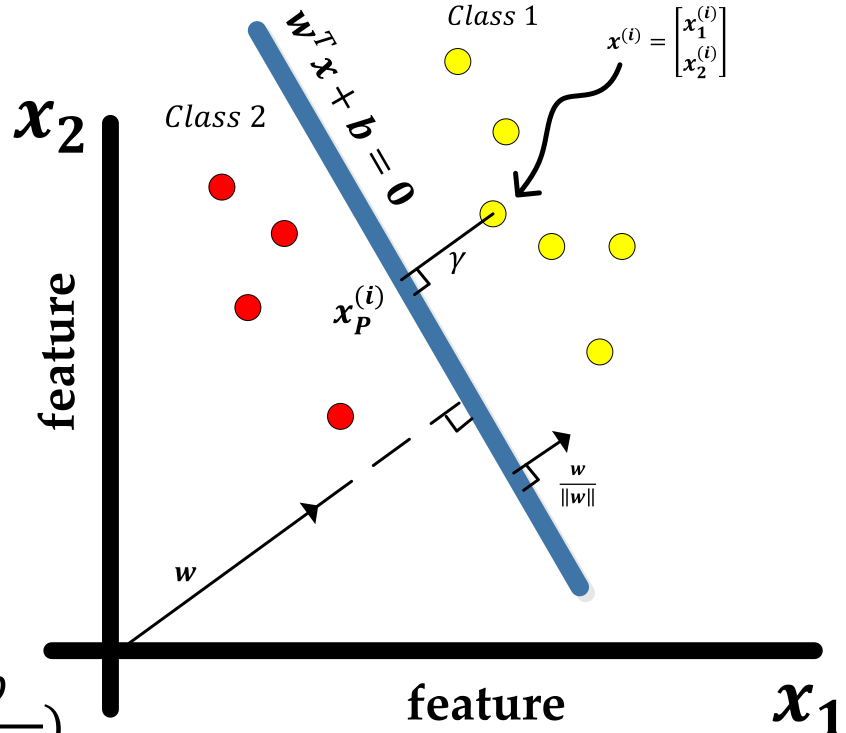
$$x_P^{(i)} = -\frac{b}{w^T}$$

- Substituting $x_P^{(i)}$, we get

$$\gamma = \frac{w^T}{y^{(i)}\|w\|}(x^{(i)} + \frac{b}{w^T})$$

- Or since $y^{(i)}$ ={-1,1}

$$\gamma = y^{(i)}(\frac{w^T}{\|w\|} x^{(i)} + \frac{b}{\|w\|})$$



Geometric margin

- We define the geometric margin for a specific training set D as the smallest geometric margin when considering all samples in the set:

$$\gamma = \min_{i=1\ldots m} \gamma^{(i)}$$

- We also define as functional margin as:

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b) = \gamma ||w||$$

$$\hat{\gamma} = \min_{i=1\ldots m} \hat{\gamma}^{(i)}$$

Forming the optimization problem

- In the last lecture, we defined the optimization problem that the **Optimal Margin Classifiers** have to solve:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$so\ that\ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 \quad for\ i = 1 \dots m$$

- To solve this, we introduce the **Lagrange multipliers** and we formulate the **Langrangian** function:

$$L(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} a_i \left(y^{(i)}\left(w^T x^{(i)} + b\right) - 1\right)$$

- This we can solve by setting the **derivatives w.r.t. w, b, a to zero**. For example:

$$\frac{\partial L(w, b, a)}{\partial w_1} = w_1 - \sum_{i=1}^{m} a_i\, y^{(i)} x_1^{(i)} = 0 \Longrightarrow \quad w_1 = \sum_{i=1}^{m} a_i\, y^{(i)} x_1^{(i)}$$

Support Vectors

- Only a few of the points will have $a_i \neq 0$, otherwise we can minimize the Langrangian to our hearts delight with selecting arbitrary large a_i

$$L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} a_i \left(y^{(i)}\left(w^T x^{(i)} + b\right) - 1\right)$$

- Those points are called **support vectors** and they are on the margin boundary where $y^{(i)}\left(w^T x^{(i)} + b\right) = 1$

- We can re-write the optimization problem in its dual formulation that only needs the calculation of the inner product $\langle x^{(i)} x^{(j)} \rangle$

$$\begin{aligned} \max_\alpha \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \boxed{\langle x^{(i)}, x^{(j)} \rangle}. \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \end{aligned}$$

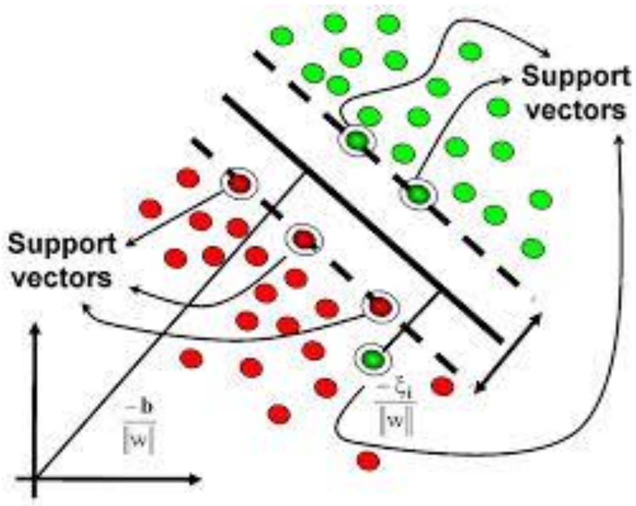
KERNEL

First cheat: Slack variables

- In some cases we want the classifier to be less sensitive to outliers to keep the margin sufficiently large. To do so, we introduce "slack variables" that allow samples to be misclassified or be within the margin.

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i$$

$$so\ that\ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i \quad for\ i = 1 \dots m$$
$$\xi_i \geq 0$$



Second cheat: Kernels and feature space

- We can replace the inner product $\langle x^{(i)} x^{(j)} \rangle$ with a kernel K(x,y)=$\langle \varphi(x^{(i)}), \varphi(x^{(j)}) \rangle$ to map the samples in a different "feature space" where they can be more separable.

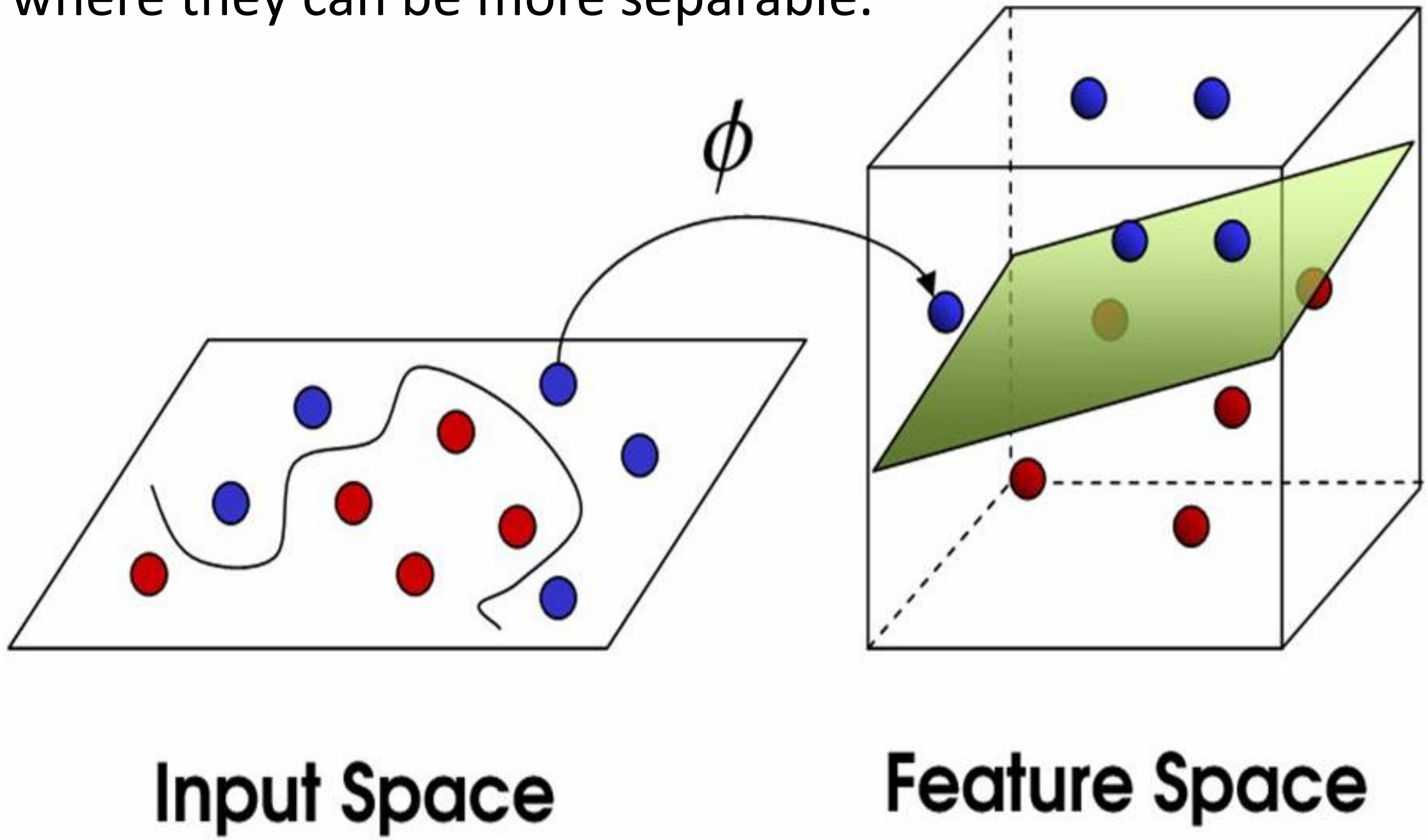
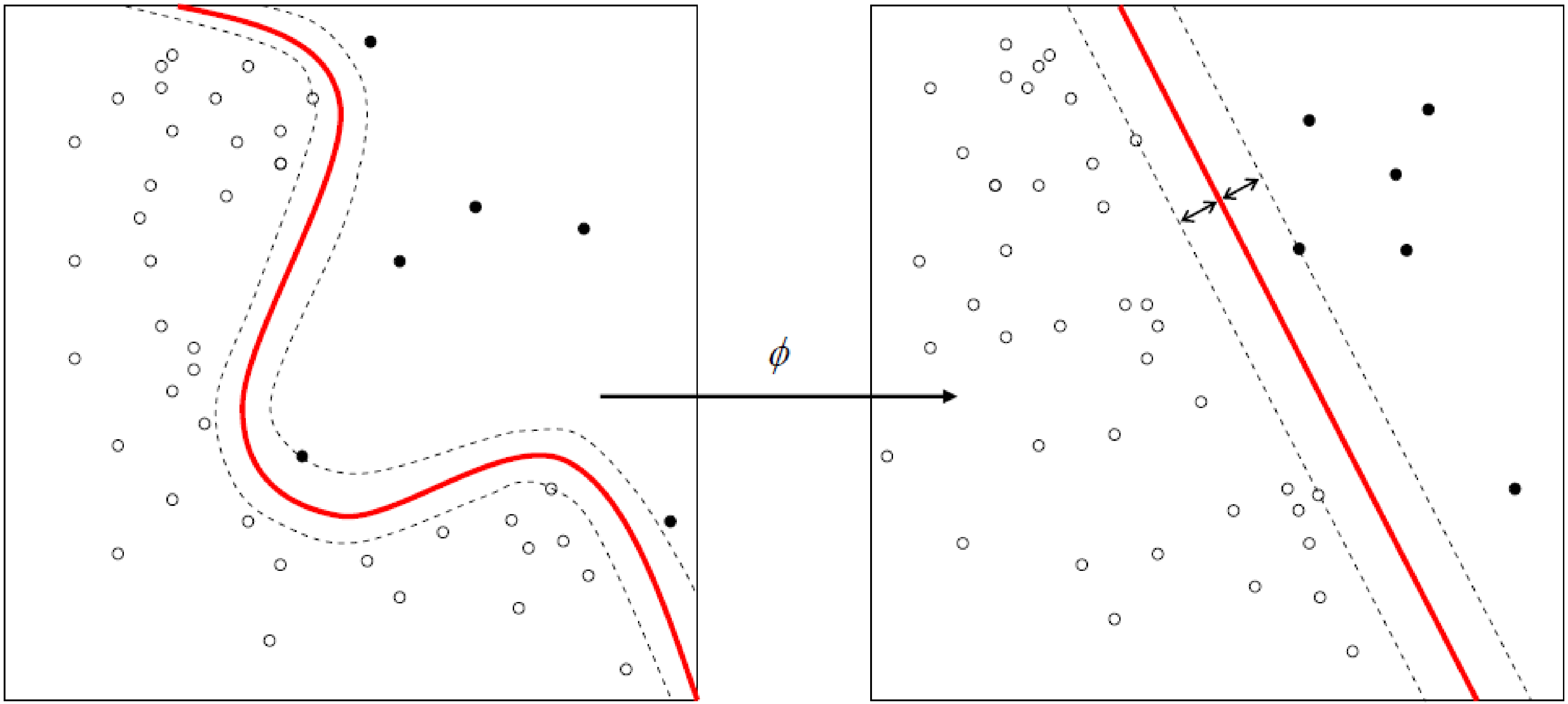


Image stolen from a random internet location

Second cheat: Kernels and feature space



Second cheat: Kernels and feature space

Gaussian RBF

$$k(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{z}\|^2}{c}\right)$$

Polynomial

$$k(\boldsymbol{x}, \boldsymbol{z}) = \left(\left(\boldsymbol{x}^\top \boldsymbol{z}\right) + \theta\right)^d$$

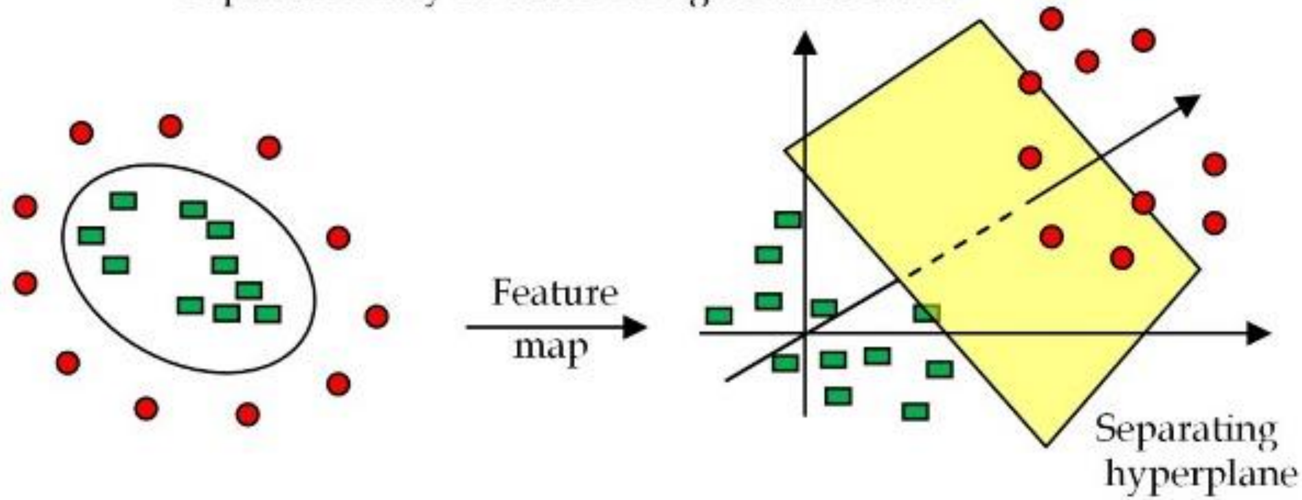
Sigmoidal

$$k(\boldsymbol{x}, \boldsymbol{z}) = \tanh\left(\kappa\left(\boldsymbol{x}^\top \boldsymbol{z}\right) + \theta\right)$$

Inverse multi-quadric

$$k(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{\sqrt{\|\boldsymbol{x} - \boldsymbol{z}\|^2 + c^2}}$$

Separation may be easier in higher dimension



Complex in low dimensions

Simple in higher dimensions

End of Lecture 10