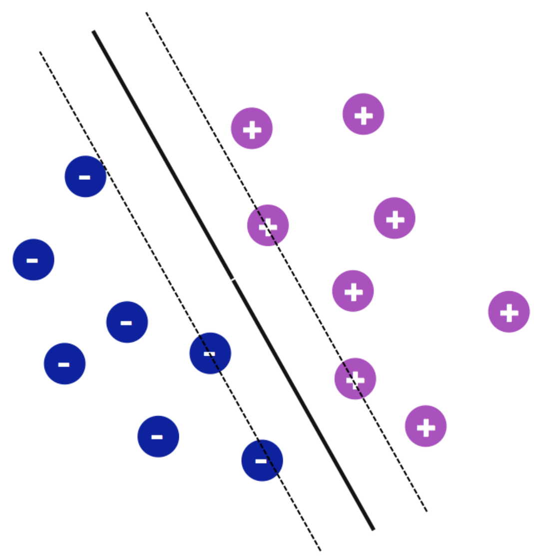
Artificial Intelligence

Machine Learning
Support Vector Machines



Ansaf Salleb-Aouissi

Columbia University - COMS 4701 - Summer 2016

Outline

1. History of Support Vector Machines (SVMs)
2. Basic Idea
3. Choice of the hyperplane: linearly separable case
4. Choice of the hyperplane: non-linearly separable case
5. SVM Primal Form
6. Lagrange Duality
7. SVM Dual Form
8. SVM with a Soft-margin
9. A hint of Kernels (more in the next lecture)
10. SVMs in practice
11. Demo
12. Non-linearity: Example
13. From linear models to non-linear models
14. Kernels
15. Examples of Kernels
16. Validity of Kernels
17. Composition of Kernels
18. Conclusion

History of SVMs

- SVMS: State-of-the-art classification method.
- Boser, Guyon and Vapnik 1992.
- Powerful and widely used in both academia and industry:
  1. Handles high-dimensional data
  2. Handles non-linear problems
  3. Allows overlap in the classes
- A kernel method that depend only on the data through inner products.
- Come with theoretical guaranteed about their performance.

Basic Idea

- Find the optimal hyperplane for linearly separable examples.
- For non linearly separable data, transform the original data using a *kernel function*.
- To allow for some overlap in the classes, use *slack variables*.
- The support vectors are the examples that are the closest to the decision surface.
- Support Vectors are the most difficult to classify.
- Output a discrete answer $\in \mathbb{Y} = \{-1, +1\}$.

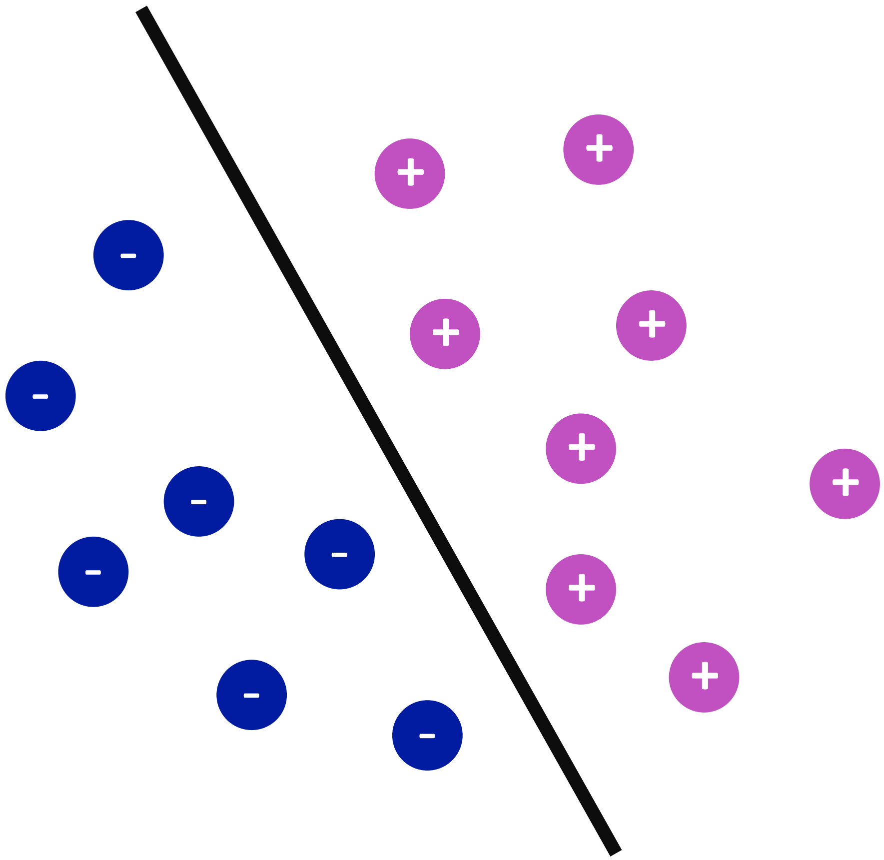
Choice of the hyperplane

Given: Training data: $(x_1, y_1), \ldots, (x_n, y_n) / x_i \in \mathbb{R}^d$ and y_i is discrete $y_i \in \mathbb{Y} = \{-1, +1\}$.

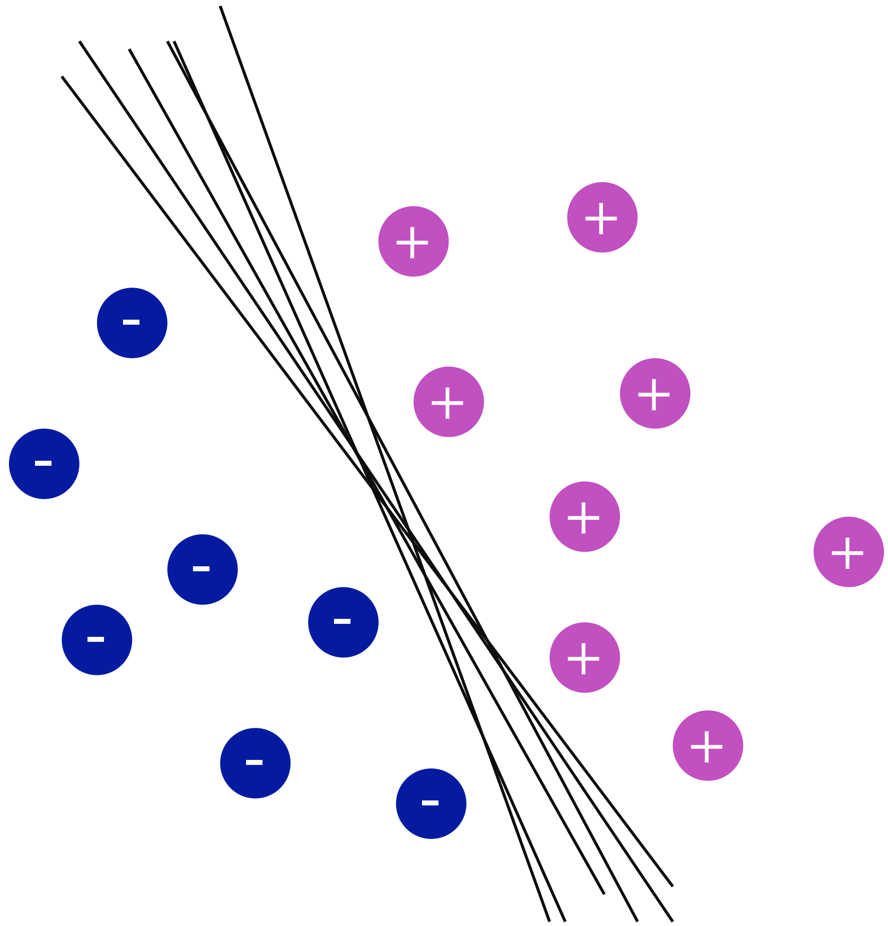
Task: Learn a classification function:
$f : \mathbb{R}^d \longrightarrow \mathbb{Y}$

$$f(x) = sign(\sum_{i=0}^{d} w_i x_i)$$

Choice of the hyperplane



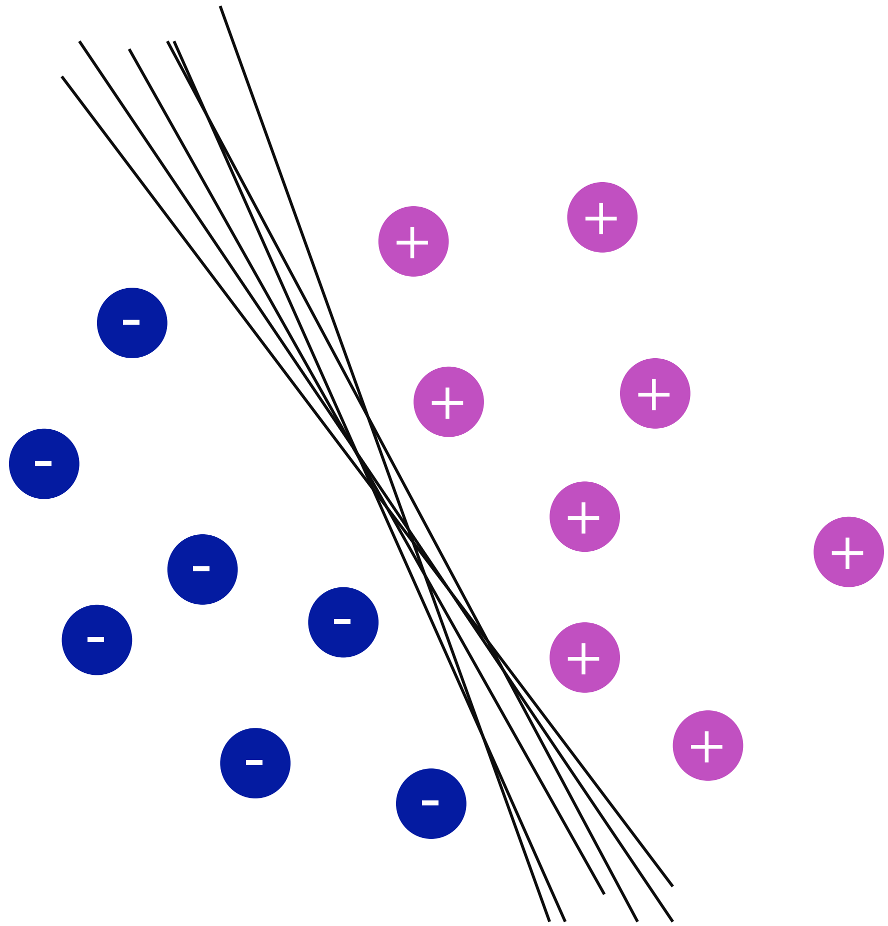
Choice of the hyperplane



Lots of possible solutions!

Copyright ©Ansaf Salleb-Aouissi: Summer 2016 COMS 4701 Section 1 ARTIFICIAL INTELLIGENCE

Choice of the hyperplane

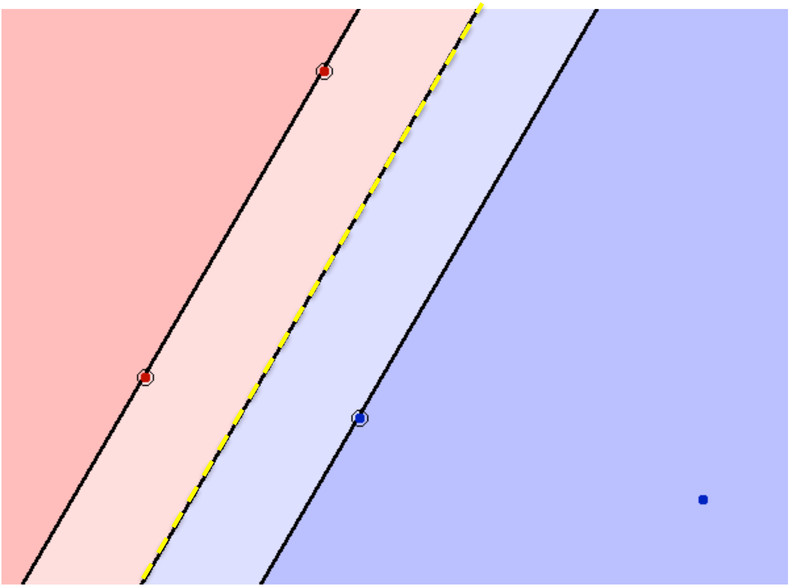


Lots of possible solutions!

Idea of SVM: find the "best" margin".

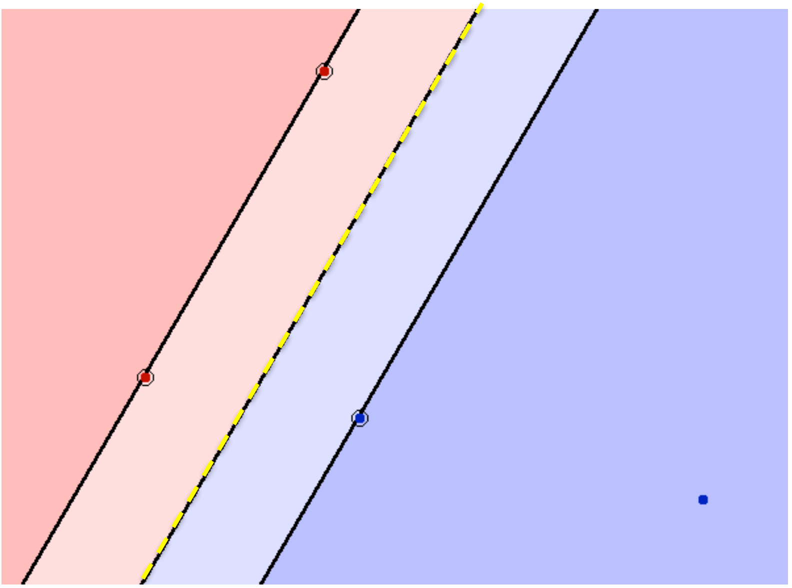
Maximum Margin: Intuition

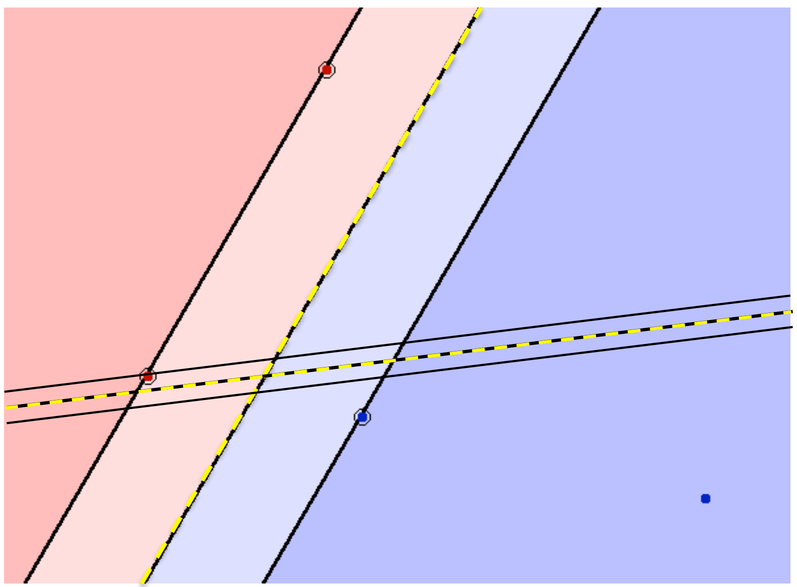
Why is a fat margin the best?



Maximum Margin: Intuition

Why is a fat margin the best?





Choice of the hyperplane

Given: Training data: $(x_1, y_1), \ldots, (x_n, y_n) / x_i \in \mathbb{R}^d$ and y_i is discrete $y_i \in \mathbb{Y} = \{-1, +1\}$.

Task: Learn a classification function: $f : \mathbb{R}^d \longrightarrow \mathbb{Y}$

$$f(x) = sign(\sum_{i=0}^{d} w_i x_i)$$

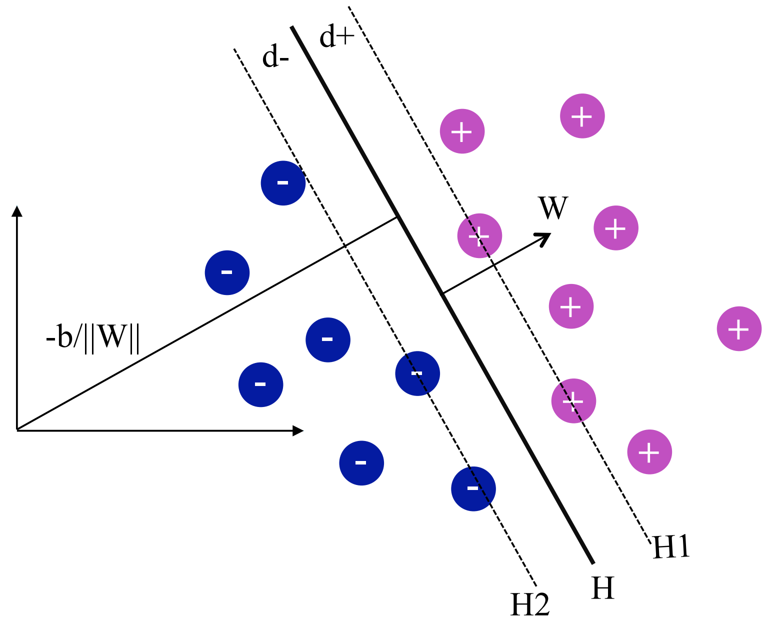
$$f(x) = sign(\sum_{i=1}^{d} w_i x_i + b)$$

$$f(x) = sign(w.x + b) \quad \text{(with "." is the dot product)}$$

Note: b corresponds to the intercept β_0 in the methods we have seen before. We use w and b as these are the most commonly used for SVMs. It will help in case you read SVMs literature.

Choice of the hyperplane

The separable case:



- The hyperplane satisfies: $w.x + b = 0$.
- w is the normal to the hyperplane. $||w||$ is its norm.
- $|b|/||w||$ is the perpendicular distance from the hyperplane to the origin.
- $d+$ is the shortest distance from the hyperplane to the closest positive example. $d-$ is the shortest from the hyperplane to the closest negative example.
- Margin: $d_+ + d_-$

The SVM algorithm looks for separating hyperplane with the **largest margin**.

The examples on H_1 and H_2 are called the Support Vectors (SVs) and have $w.x + b = 0$.

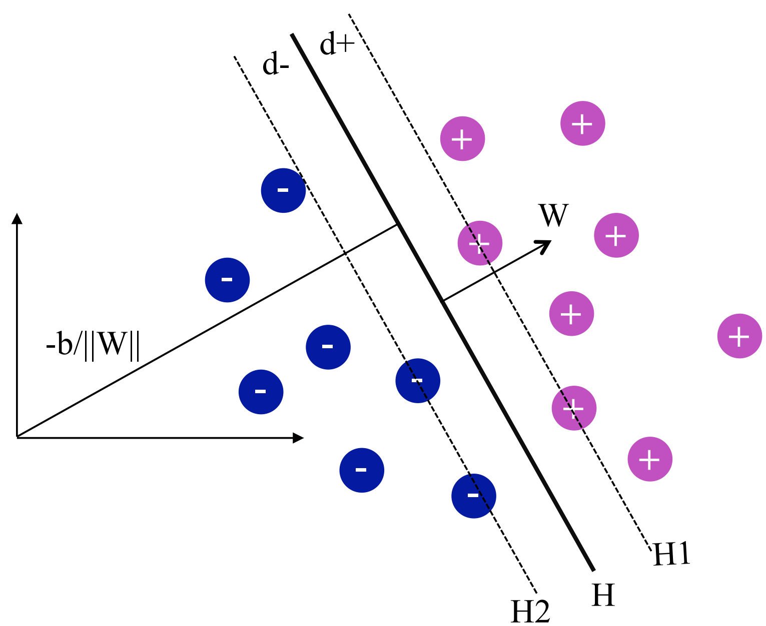
Choice of the hyperplane

The separable case:

- H_1 and H_2 are parallel.
- H_1: $w.x_i + b = +1$ with normal w and perpendicular distance from the origin $|1-b|/||w||$.
- H_2: $w.x_i + b = -1$ with normal w and perpendicular distance from the origin $|-1-b|/||w||$.
- $d_+ = d_- = 1/||w||$h
- $w.x_i + b \geq +1$ if $y_i = +1$

  $w.x_i + b \leq -1$ if $y_i = -1$
- These can be combined:

  $$y_i(w.x_i + b) \geq 1 \ \forall i = 1, \ldots, n$$



Choice of the hyperplane

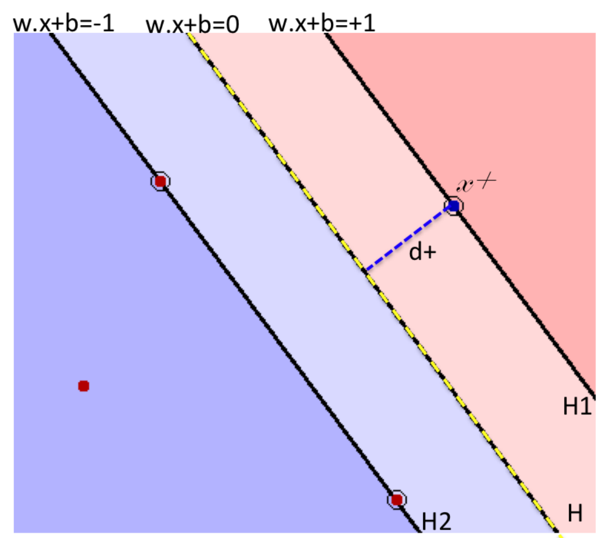
Why is $d_+ = d_- = \frac{1}{||w||}$?

The distance from a point (x_0, y_0) to a line with equation $ax + by + c = 0$ is:

$$distance(ax + by + c = 0, (x_0, y_0)) = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}}$$

If we reason in 2D without loss of generality, we have vector w $= (w_1, w_2)$. We have $a = w_1,\ b = w_2,\ c = b$.

$\sqrt{w_1^2 + w_2^2} = ||w||$ that is norm of vector w



Distance $d+$ from any positive point x_+ in H_1 to the hyperplane H.

$$d+ = \frac{|w.x_+ + b|}{||w||}$$

x_+ being on H_1 verifies the equation $w.x_+ + b = 1$.
Hence: $d_+ = \frac{1}{||w||}$

One could do a similar calculation for a point x_- on H_2 to get d_-.

SVMs Primal Form

The maximum margin classifier is the function that maximizes the geometric margin $1/||w||$, equivalent to minimizing $||w||^2$.

Solve the constrained optimization problem:

$$\underset{w,b}{\text{Argmin}} \quad \frac{1}{2}||w||^2$$

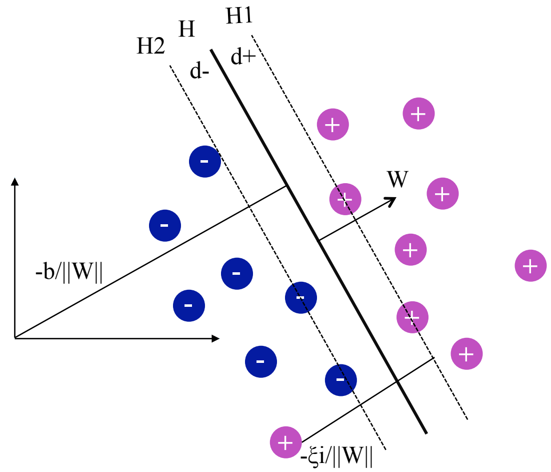
$$\text{subject to:} \quad y_i(w.x_i + b) \;\geq\; 1 \;\forall i = 1, \ldots, n$$

Inequality constraint

Choice of the hyperplane

The non separable case:

We allow errors but not too much!



$$\underset{w,b}{\text{argmin}} \quad \frac{1}{2}||w||^2 + C\sum_i \xi_i$$

$$\text{subject to:} \quad y_i(w.x_i + b) \;\geq\; 1 - \xi_i \qquad -\xi_i \geq 0 \;\forall i = 1, \ldots, n$$

A large C corresponds to assigning a higher penalty to errors.

Lagrange Duality

Solving constrained optimization problems:

$$\begin{cases} \underset{w}{\text{Argmin}} \quad f(w) \\ \\ \text{s.t.} \quad h_i(w) = 0 \ \forall i = 1, \ldots, n \end{cases}$$

Can be solved with **Lagrange multipliers**.

Lagrangian:

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{n} \beta_i h_i(w)$$

The βs are called Lagrange multipliers.

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

Lagrange Duality

$$\begin{cases} \underset{w}{\text{Argmin}} \quad f(w) \\ \\ \text{s.t.} \quad g_i(w) \leq 0 \ \forall i = 1, \ldots, k \\ \text{s.t.} \quad h_i(w) = 0 \ \forall i = 1, \ldots, l \end{cases}$$

Generalized Lagrangian:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Lagrange Duality

$$\begin{cases} \underset{w}{\text{Argmin}} \quad f(w) \\ \\ \text{s.t.} \quad g_i(w) \leq 0 \ \forall i = 1, \ldots, k \\ \text{s.t.} \quad h_i(w) = 0 \ \forall i = 1, \ldots, l \end{cases}$$

Generalized Lagrangian:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Solution w^* in the primal, α^* and β^* in the dual.

For a solution to exist (and hence the primal and dual problems are equivalent), the **Karush-Kuhn-Tucker KKT Conditions** must be fulfilled.

Lagrange Duality

Karush-Kuhn-Tucker KKT Conditions.

1. $\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \forall i = 1, \ldots, n$

2. $\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \forall i = 1, \ldots, l$

3. $\alpha_i^* g_i(w^*) = 0, , \forall i = 1, \ldots, k$

4. $g_i(w^*) \leq 0, \forall i = 1, \ldots, k$

5. $\alpha^* \geq 0, \forall i = 1, \ldots, k$

For more details: T. Rockarfeller (1970) Convex Analysis, Princeton University Press.

SVMs Dual Form

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [y_i(w.x_i + b) - 1]$$

SVMs Dual Form

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [y_i(w.x_i + b) - 1]$$

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

SVMs Dual Form

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [y_i(w.x_i + b) - 1]$$

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

SVMs Dual Form

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i[y_i(w.x_i + b) - 1]$$

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

By plugging in these 2 quantities back into $\mathcal{L}$:

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j$$

Copyright ©Ansaf Salleb-Aouissi: Summer 2016 COMS 4701 Section 1 ARTIFICIAL INTELLIGENCE

SVMs Dual Form

$$\underset{\alpha}{\text{Argmax}} \sum_{i=1}^{n} {}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \color{blue}{x_i x_j}$$

$$\text{s.t.} \qquad \alpha_i \geq 0, \forall i = 1, \cdots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Solve the dual problem to find the α's!

SVMs Dual Form

Few observations:

- Total dependence on the **dot product**.
- The dual form depends only on the inputs.
- Once we find the α's, we can find the optimal w's.

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

- We can find the optimal b, that is b^* but reconsidering the primal form.

  Teaser: calculate b^* using w^*.
- Except for Support vectors, all α's will be 0 (from KKT).
- How can we make a prediction given an example u (unknown)?

SVMs Dual Form

Few observations:

- Total dependence on the **dot product**.
- The dual form depends only on the inputs.
- Once we find the α's, we can find the optimal w's.

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

- We can find the optimal b, that is b^* but reconsidering the primal form.

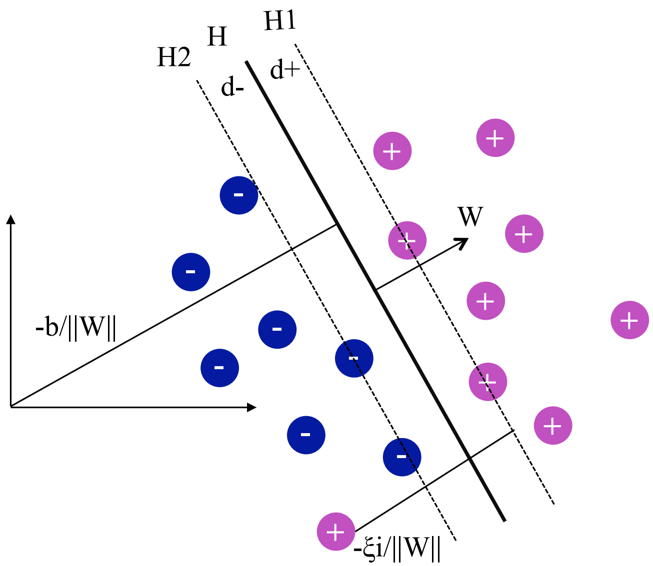
  Teaser: calculate b^* using w^*.
- Except for Support vectors, all α's will be 0 (from KKT).
- How can we make a prediction given an example u (unknown)?

$$\sum_{i=1}^{n} \alpha_i^* y_i x_i u + b^*$$

Copyright ©Ansaf Salleb-Aouissi: Summer 2016 COMS 4701 Section 1 ARTIFICIAL INTELLIGENCE

Soft Margin

The non separable case: We allow errors but not too much!



$$\underset{w,b}{\text{Argmin}} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

subject to: $\quad y_i(w.x_i + b) \ \geq \ 1 - \xi_i \quad \xi_i \geq 0 \ \forall i = 1, \ldots, n$

A large C corresponds to assigning a higher penalty to errors.

Soft Margin: dual form

$$\underset{\alpha}{\text{Argmax}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j$$

$$\text{s.t.} \qquad 0 \leq \alpha_i \leq C, \forall i = 1, \cdots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

SVM in Practice

- Normalization is important.

SVM in Practice

- Normalization is important.
- Do model selection by searching the parameters.

SVM in Practice

- Normalization is important.
- Do model selection by searching the parameters.
- RBF (Gaussian kernel) is an effective and mostly used kernel.

SVM in Practice

- Normalization is important.
- Do model selection by searching the parameters.
- RBF (Gaussian kernel) is an effective and mostly used kernel.
- SVM with unbalanced datasets:

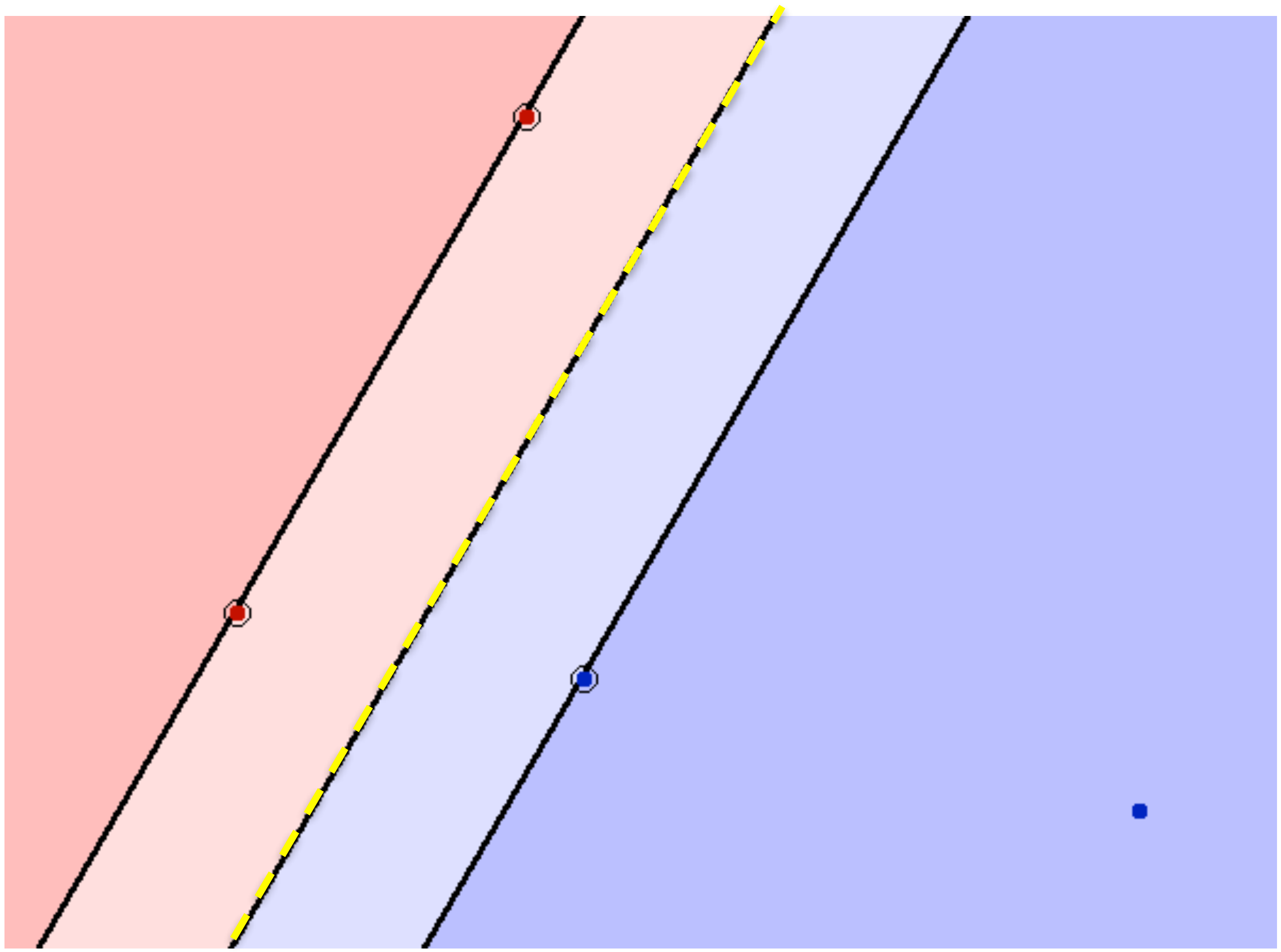
$$
\begin{aligned}
\underset{w,\xi}{\text{Argmin}} \quad & \frac{1}{2}||w||^2 + C_- \sum_{y_i=-1} \xi_i + C_+ \sum_{y_j=+1} \xi_j \\
\text{s.t.} \quad & y_k[w^\top x_k + b] \geq 1 - \xi_k, \quad \forall k,
\end{aligned}
$$

$$C_+ \; n_+ = C_- \; n_-$$

SVM in Practice

- Free Implementations: LibSVM, SVMLight.
- http://www.kernel-machines.org/
- Demo (at the end of the lecture): http://las.ethz.ch/courses/ml-f13/applets/JSupportVectorApplet.html
- More on kernels in a bit...

Maximum Margin



SVMs primal and dual Forms

Separable case:

$$\underset{w,b}{\text{Argmin}} \quad \frac{1}{2}||w||^2$$

$$\text{subject to:} \quad y_i(w.x_i + b) \; \geq \; 1 \; \forall i = 1, \ldots, n$$

$$\underset{\alpha}{\text{Argmax}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j$$

$$\text{s.t.} \qquad \alpha_i \geq 0, \forall i = 1, \cdots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

SVMs Dual Form

- Solve the dual problem to find the α's!
- Calculate w's as follows:

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

- How can we make a prediction given an example u (unknown)?

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i^* y_i x_i u + b^*)$$

Soft Margin

The non separable case:

$$\underset{w,b}{\text{Argmin}} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

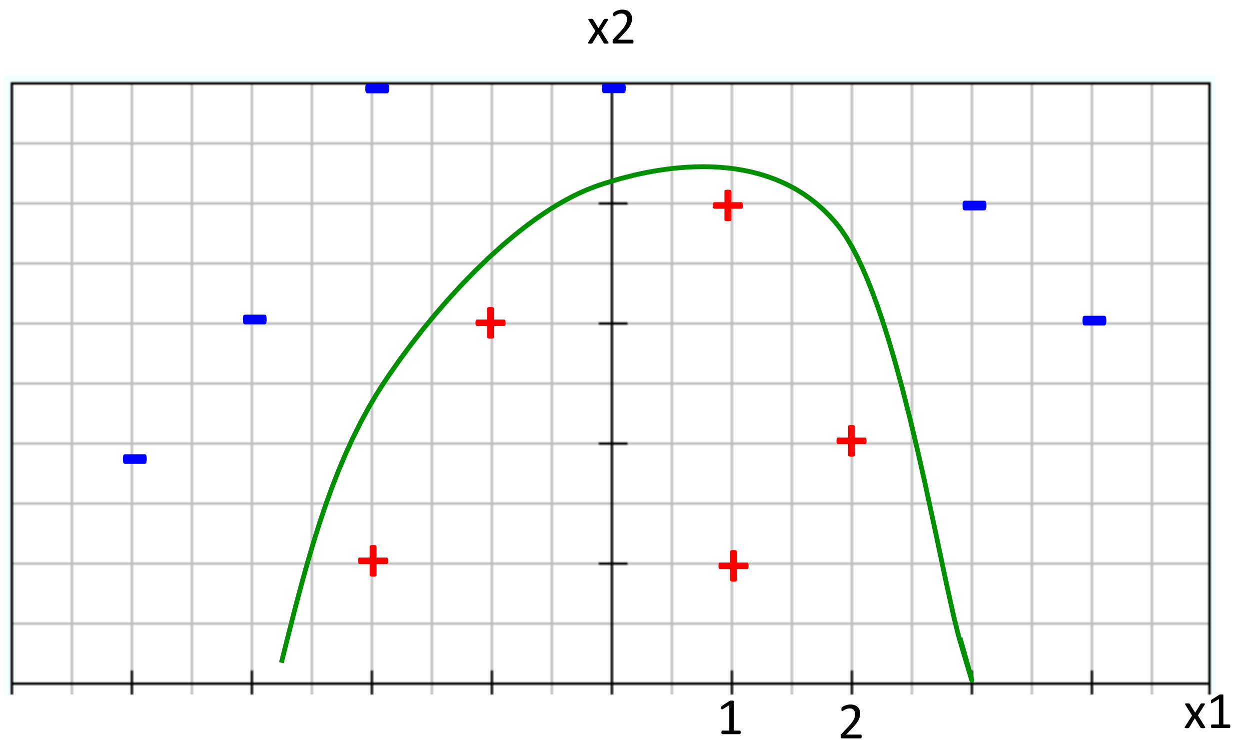
$$\text{subject to:} \quad y_i(w.x_i + b) \; \geq \; 1 - \xi_i \quad \xi_i \geq 0 \; \forall i = 1, \ldots, n$$

$$\underset{\alpha}{\text{Argmax}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \color{blue}{x_i x_j}$$

$$\text{s.t.} \qquad 0 \leq \alpha_i \leq C, \forall i = 1, \cdots, n$$

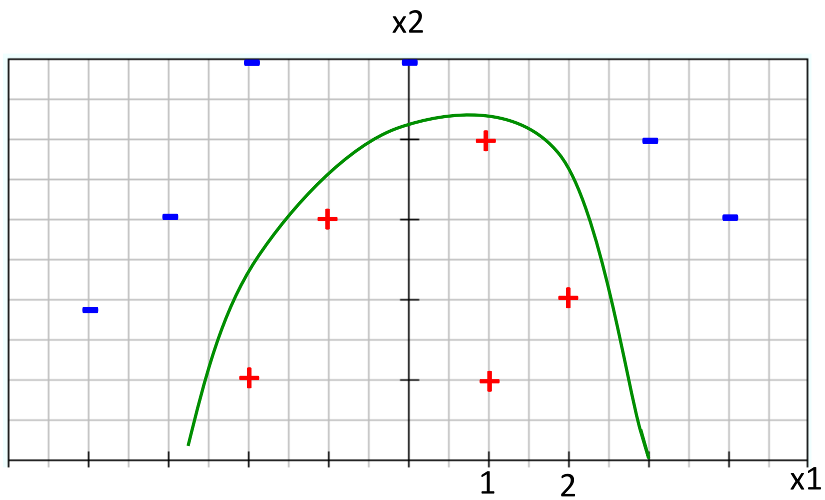
$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

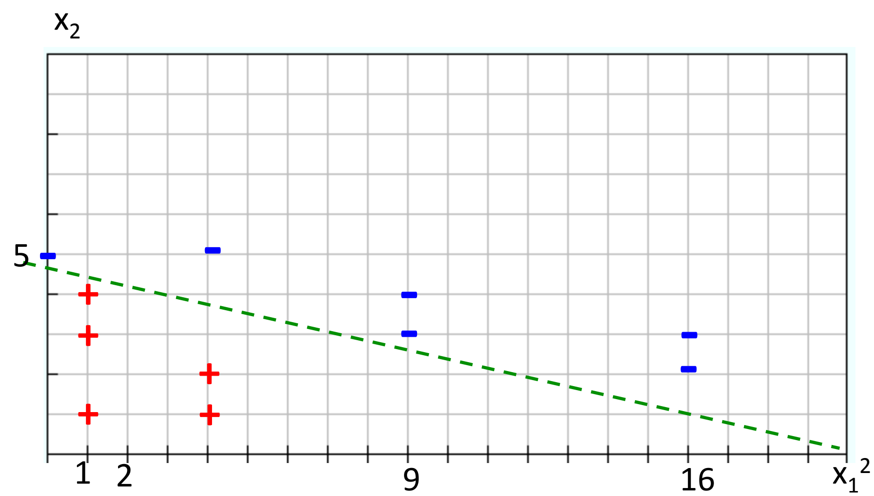
Non-linear problems



Non-linear problems

$$\phi(x) = (x_1^2, x_2)$$

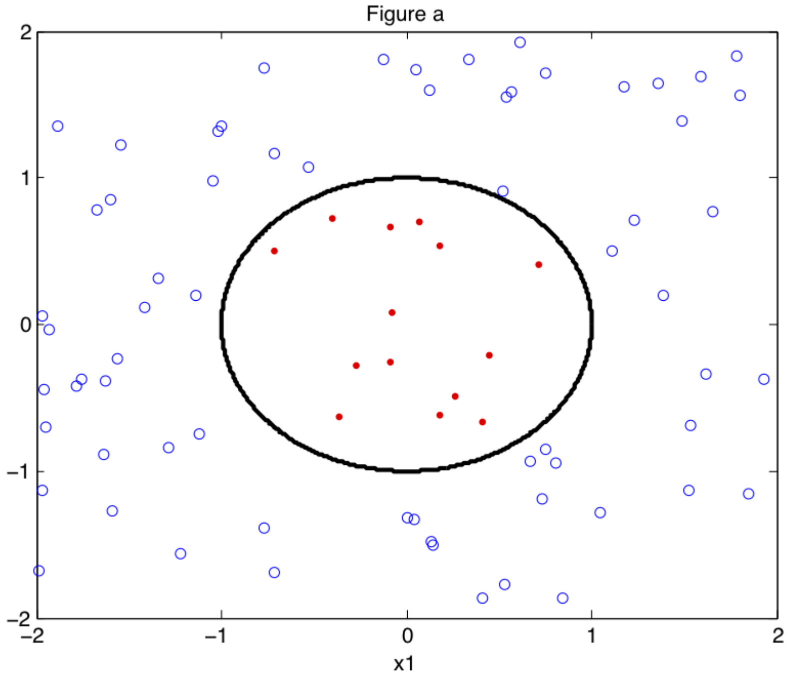


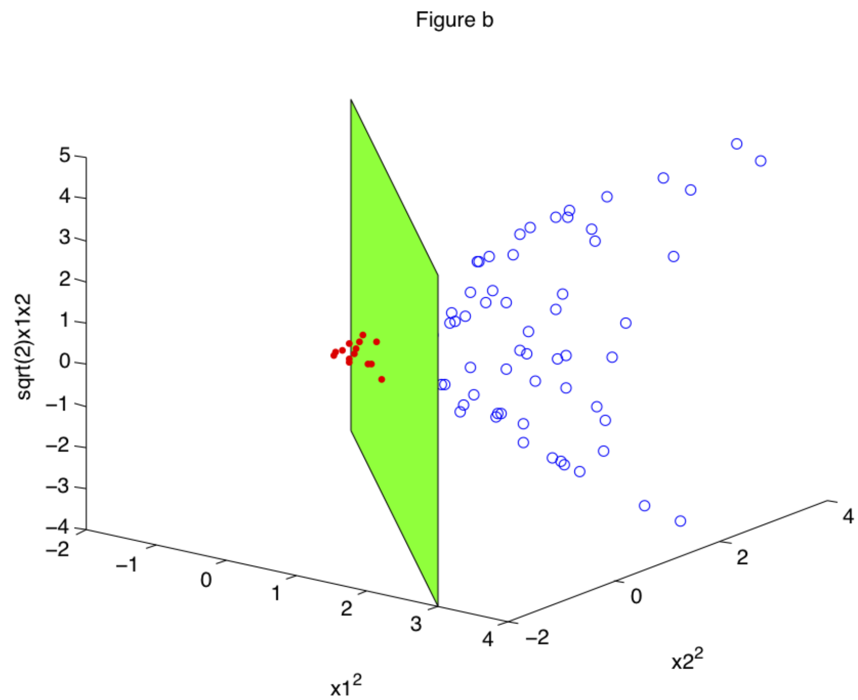


$$f(x) = w.\phi(x) + b$$

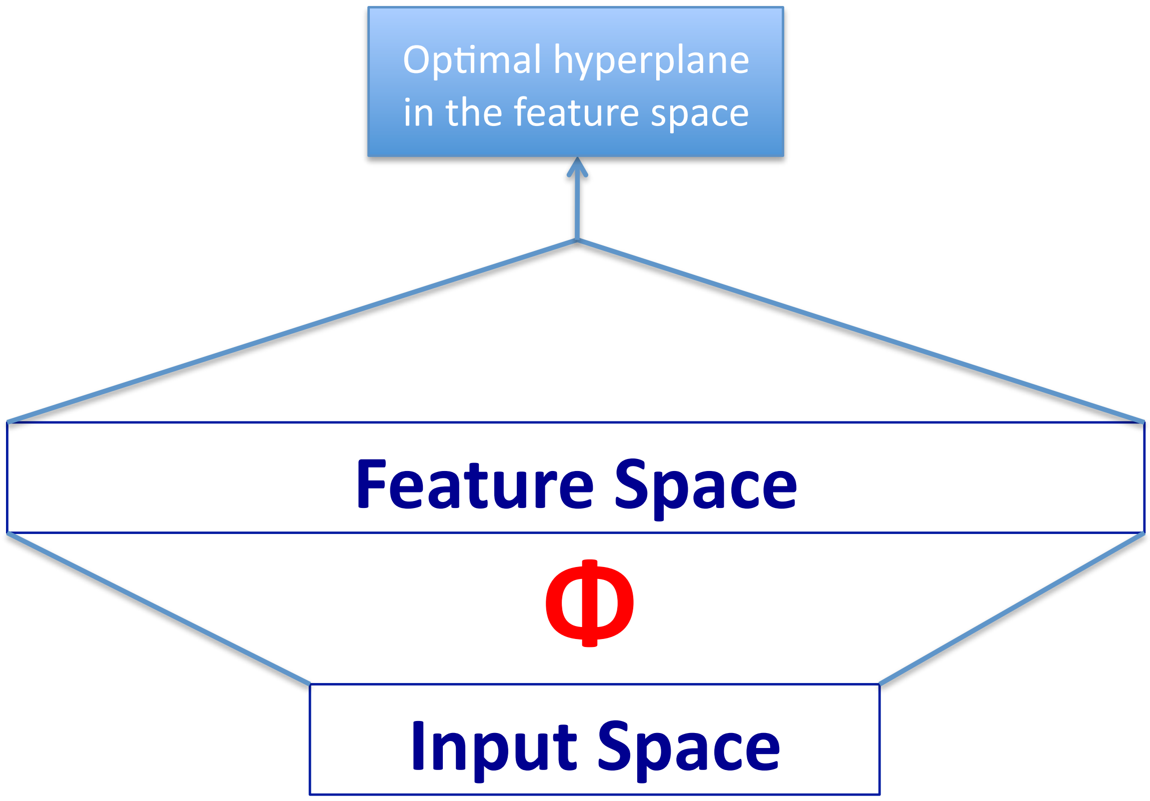
Non-linear problems

$$\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$





Beyond the input space



Plug in ϕ into the dual?

$$\underset{w,b}{\text{Argmin}} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

subject to: $y_i(w.x_i + b) \ \geq \ 1 - \xi_i \quad \xi_i \geq 0 \ \forall i = 1, \ldots, n$

$$\underset{\alpha}{\text{Argmax}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j$$

s.t. $0 \leq \alpha_i \leq C, \forall i = 1, \cdots, n$

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

$$w^* = \sum_{i=1}^{n}\alpha_i^* y_i x_i$$

Replace all x_i by $\phi(x_i)$!

Problems with plugging ϕ

Exponential number of features in the feature space!

Problems with plugging ϕ

Exponential number of features in the feature space!

Do we have to create and represent all these features?

Problems with plugging ϕ

Exponential number of features in the feature space!

Do we have to create and represent all these features?

No need to do it explicitly. Instead, use kernels!

Problems with plugging ϕ

Exponential number of features in the feature space!

Do we have to create and represent all these features?

No need to do it explicitly. Instead, use kernels!

$$K(x, x') = \phi(x).\phi(x')$$

We can do so because the dual form relies on inner products!

Example

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$$K(x, x') = \phi(x).\phi(x')$$

$$K(x, x') = [x.x' + 1]^2$$

Given two points $x^T = (x_1, x_2)$ and $x'^T = (x'_1, x'_2)$

$K(x, x') = [x.x' + 1]^2$
$K(x, x') = (x_1x'_1 + x_2x'_2 + 1)^2$
$K(x, x') = x_1^2{x'_1}^2 + x_2^2{x'_2}^2 + 2x_1x'_1x_2x'_2 + 2x_1x'_1 + 2x_2x'_2$

Which is the inner product $\phi(x).\phi(x')$.

Dual with Kernel

$$\underset{\alpha}{\text{Argmax}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{s.t.} \qquad 0 \leq \alpha_i \leq C, \forall i = 1, \cdots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, u) + b^*)$$

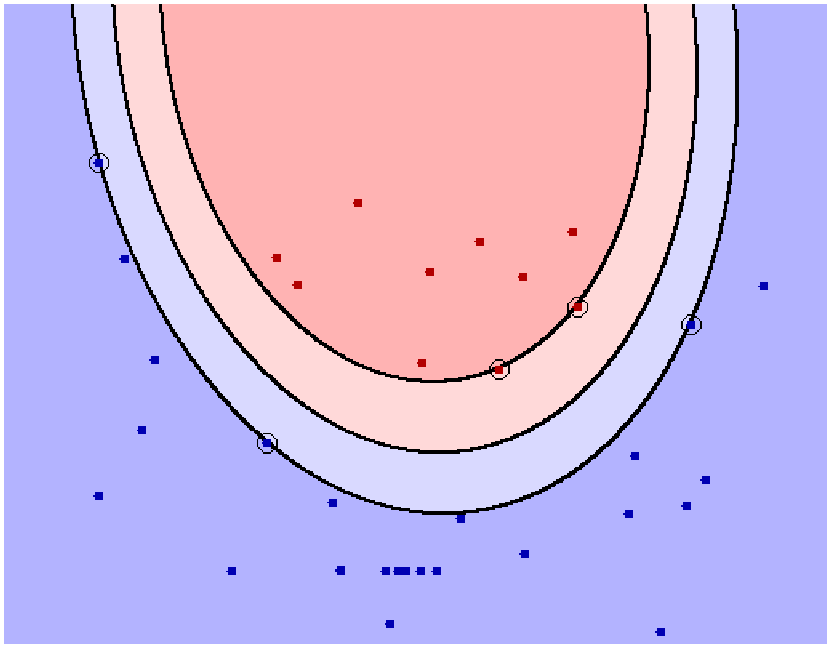
Examples of Kernels

- Linear: $K(x, x') = x.x'$
- Polynomial: $K(x, x') = [x.x' + 1]^d$
- Radial Basis Function (RBF): $exp(-\gamma[x - x']^2)$

Kernels can compute inner products efficiently.

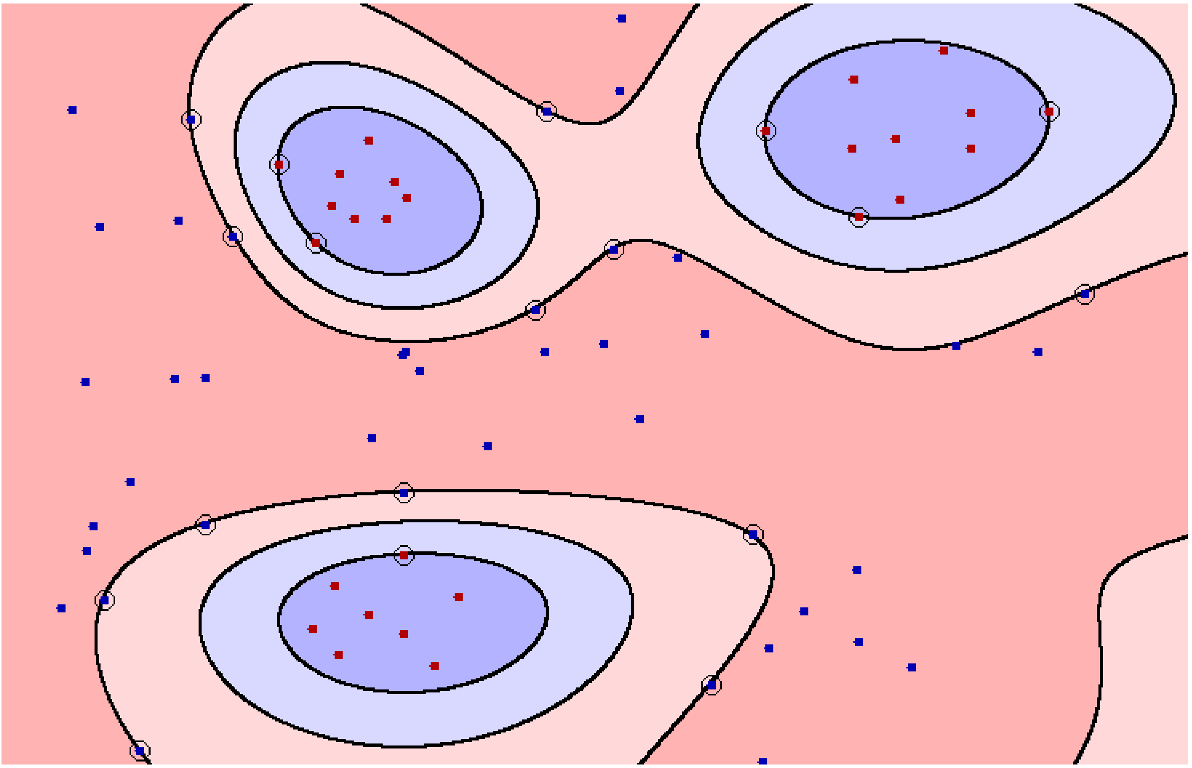
Note: In the polynomial kernel, d refers to the degree of the polynomial, not the number of features.

Example of polynomial kernel



$$K(x, x') = [x.x' + 1]^2$$

Example of RBF kernel



$$exp(-\gamma[x - x']^2)$$

Demo

http://las.ethz.ch/courses/ml-f13/applets/JSupportVectorApplet.html

Validity of kernels

A kernel $K(x, x')$ is a valid kernel iff for all example $x_1, x_2, \ldots, x_n$, it produces a **Gram** matrix:

$$G_{ij} = K(x_i, x_j)$$

1. Symmetric: .

$$G = G^T$$

2. positive semi-definite:

$$\alpha^T G \alpha \geq 0 \quad \forall \alpha$$

These are **Mercer conditions**. It ensures convexity of the dual form.

Composition of kernels

Given two valid kernels K_1 and K_2, $\alpha > 0$, $0 \leq \lambda \leq 1$, f a real-valued function, a mapping ϕ, K a positive semi-definite matrix, then the following functions are valid kernels:

1. $K(x,z) = \lambda K_1(x,z) + (1-\lambda)K_2(x,z)$
2. $K(x,z) = \alpha K_1(x,z)$
3. $K(x,z) = K_1(x,z)K_2(x,z)$
4. $K(x,z) = f(x)f(z)$
5. $K(x,z) = K_3(\phi(x), \phi(z))$
6. $K(x,z) = x^T K z$

Copyright ©Ansaf Salleb-Aouissi: Summer 2016 COMS 4701 Section 1 ARTIFICIAL INTELLIGENCE

Conclusion

SVM with kernels:

- Independent of the dimensionality of feature space.
- Has one global optima.
- Can represent any boolean function and reasonably any arbitrary smooth decision boundary.
- Need to choose a kernel type and its parameters.
- Setting the hyper-parameter is crucial but non-trivial.
- In practice, they are usually set using cross validation.
- RBF kernel is a reasonable first choice.
- There are very specific kernels depending on the applications (e.g., tree kernels, graph kernels, etc.).

Credit

- A User's guide to Support Vector Machines. Benhur and Weston 2010.
- A Tutorial on Support Vector Machines for Pattern Recognition. Burges, Christopher Data Mining and Knowledge Discovery 2, no. 2 (June 1998): 121-167.
- Statistical Learning Theory. Vapnik, 1998.
- Check out also "The practical guide to Support Vector Classification" Hsu and al. 2010, available online.