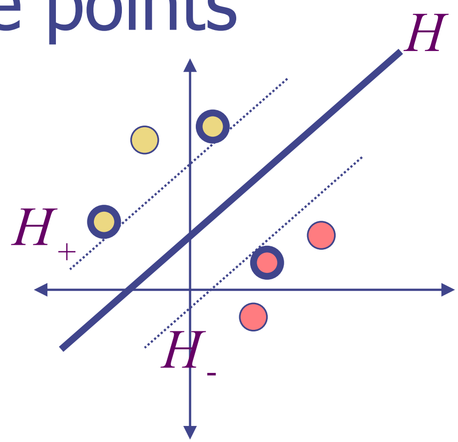
Machine Learning

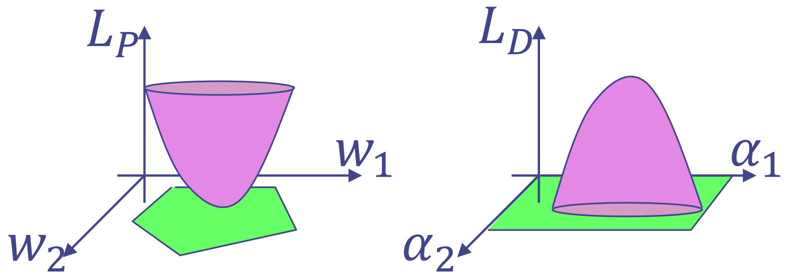
4771

Instructor: Itsik Pe’er

Reminder: SVM

- Linear classifier of separable points
- Maximizes margin
- QP + dual
- Has few support vectors





Duality

L_P

s.t.

$$\min_{w,b} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left(y_i (w^T x_i + b) - 1 \right) \geq 0$$

α_i KKT multipliers

$$w = \sum_i \alpha_i y_i x_i$$

L_D:

$$\max \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$\alpha_i \geq 0, \quad \sum_i \alpha_i y_i = 0$ $\quad \text{sign}\left(\sum \alpha_i y_i x_i^T x + b \right)$

Duality

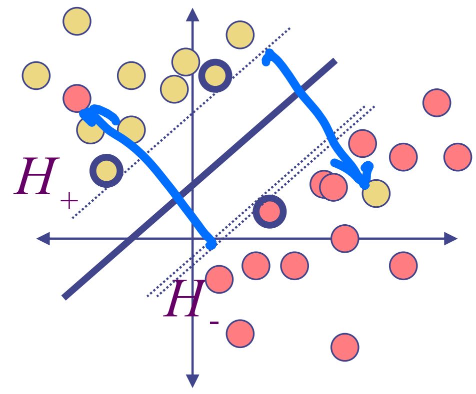
- Primal SVM problem L_P:
  minimize $\frac{1}{2}\|w\|^2$ s.t. $y_i(w^T x_i + b) - 1 \geq 0$
- ~~Lagrange~~ KKT multipliers α_i:
  $$\min_{w,b} \max_{\alpha \geq 0} \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left(y_i(w^T x_i + b) - 1\right)$$
- $w = \sum_i \alpha_i y_i x_i$; for $\alpha_i > 0$: $w^T \boldsymbol{x}_i + b = y_i$
- Dual: $L_D = \; max \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$
  s.t. $\sum_i \alpha_i y_i = 0,\ \alpha_i \geq 0$
- Classifier: $sign(\sum_i \alpha_i y_i x_i\, x + b)$

Class 9 – More SVMs

- Review
- Generalizations
  - Non-separable
  - Non-linear

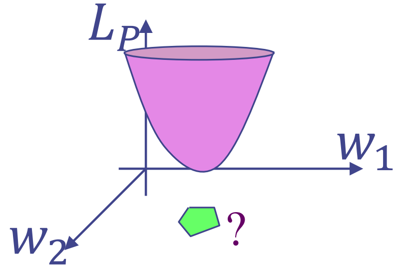
Non-Separable SVMs

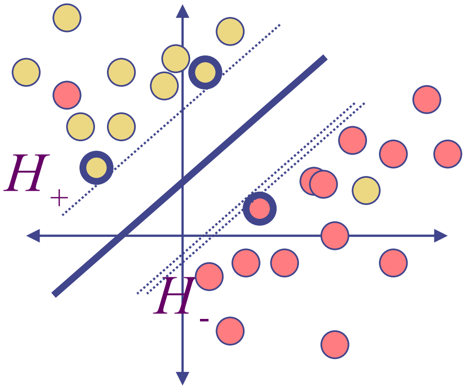
- What happens when non-separable?



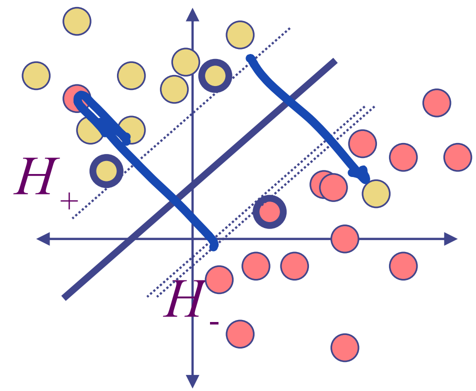
Non-Separable SVMs

- What happens when non-separable?
- There is no solution and convex hull shrinks to nothing

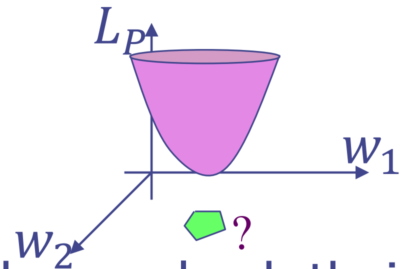




Non-Separable SVMs



- What happens when non-separable?
- There is no solution and convex hull shrinks to nothing



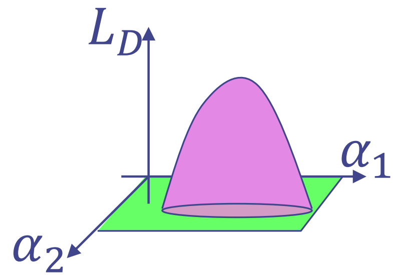
- Not all constraints can be resolved, their alphas go to ∞

$L_D = max \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$ subject to $\sum_i \alpha_i y_i = 0$, $\alpha_i \geq 0$

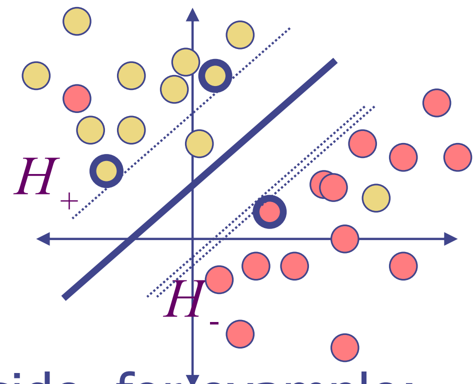
$w x_i + b \geq 1 - \xi_i$ $\quad y_i = +1$

$\xi_i \geq 0$

$\sum \xi_i \leq C$



Non-Separable SVMs

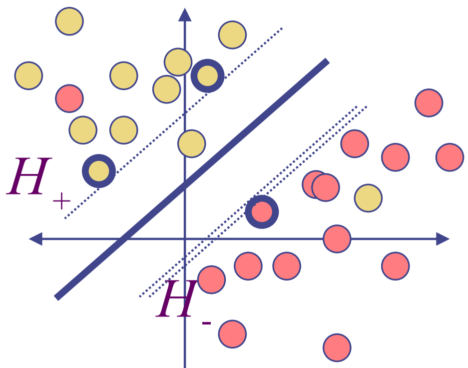


- Instead of perfectly classifying each point: $y_i(w^T x_i + b) \geq 1$ we "Relax" the problem with (positive) slack variables ξ's allow data to (sometimes) fall on wrong side, for example:

$$(w^T x_i + b) \geq 1 - 0.03 \text{ if } y_i = +1$$

Itsik Pe'er, Columbia University

Non-Separable SVMs



- Instead of perfectly classifying each point: $y_i(w^T x_i + b) \geq 1$
 we "Relax" the problem with (positive) slack variables ξ's
 allow data to (sometimes) fall on wrong side, for example:

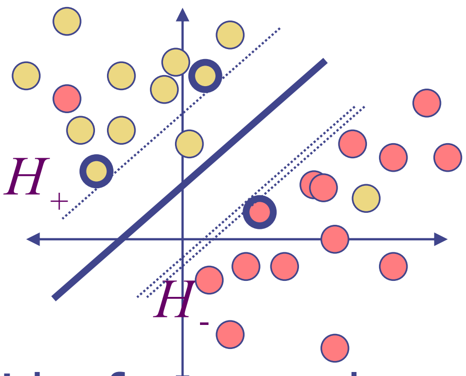
$$(w^T x_i + b) \geq 1 - 0.03 \text{ if } y_i = +1$$

- New constraints: $w^T x_i + b \geq +1 - \xi_i \;\; if \;\; y_i = +1 \; where \; \xi_i \geq 0$
 $w^T x_i + b \leq -1 + \xi_i \;\; if \;\; y_i = -1 \; where \; \xi_i \geq 0$

$$\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum \alpha_i\left(y_i(w^T x_i + b) - 1 + \xi_i\right) - \sum \mu_i \xi_i$$

Itsik Pe'er, Columbia University

Non-Separable SVMs



- Instead of perfectly classifying each point: $y_i(w^T x_i + b) \geq 1$ we "Relax" the problem with (positive) slack variables ξ's allow data to (sometimes) fall on wrong side, for example:

$$(w^T x_i + b) \geq 1 - 0.03 \text{ if } y_i = +1$$

- New constraints: $w^T x_i + b \geq +1 - \xi_i \; if \; y_i = +1 \; where \; \xi_i \geq 0$

$$w^T x_i + b \leq -1 + \xi_i \; if \; y_i = -1 \; where \; \xi_i \geq 0$$

- But too much ξ's means too much slack, so penalize them

$$L_p\!: \min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \;\; subject \; to \;\; y_i(w^T x_i + b) - 1 + \xi_i \geq 0$$

Non-Separable SVMs

- This new problem is still convex, still qp()!
- User chooses scalar C (or cross-validates) which controls how much slack ξ to use (how non-separable) and how robust to outliers or bad points on the wrong side

L_p:

Non-Separable SVMs

- This new problem is still convex, still qp()!
- User chooses scalar C (or cross-validates) which controls how much slack ξ to use (how non-separable) and how robust to outliers or bad points on the wrong side

Large margin **Low slack** **On right side** **For ξ positivity**

$$L_p: \min_{w,i,\xi} \frac{1}{2}||w||^2 + C\sum_i \xi_i - \sum_i \alpha_i\left(y_i(w^T x_i + b) - 1 + \xi_i\right) - \sum_i \beta_i \xi_i$$

max $\alpha_i \geq 0$ $\beta_i \geq 0$

$$\frac{\partial}{\partial w} L_p =$$

$$\frac{\partial}{\partial b} L_p =$$

$$\frac{\partial}{\partial \xi_i} L_p = C - \alpha_i - \beta_i = 0 \qquad \alpha_i + \beta_i = C$$

Non-Separable SVMs

- This new problem is still convex, still qp()!
- User chooses scalar C (or cross-validates) which controls how much slack ξ to use (how non-separable) and how robust to outliers or bad points on the wrong side

Low slack **On right side** **For ξ positivity**

Large margin

$$L_p\!: \min \frac{1}{2}\left\|w\right\|^2 + C\sum_i \xi_i - \sum_i \alpha_i\left(y_i(w^T x_i + b) - 1 + \xi_i\right) - \sum_i \beta_i \xi_i$$

$$\frac{\partial}{\partial w} L_p \; and \; \frac{\partial}{\partial b} L_p \;\; as\ before \ldots$$

$$\frac{\partial}{\partial \xi_i} L_p = C - \alpha_i - \beta_i = 0 \qquad \Rightarrow \alpha_i = C - \beta_i$$

$$but\ \alpha_i \,\&\, \beta_i \geq 0 \Rightarrow 0 \leq \alpha_i \leq C$$

Non-Separable SVMs

$$L_p: \min \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_i \xi_i - \sum_i \alpha_i\left(y_i(w^T x_i + b\ ) - 1 + \xi_i\right) - \sum_i \beta_i \xi_i$$

$$\frac{\partial}{\partial w} L_p \ and \frac{\partial}{\partial b} L_p \ \ as\ before \ldots$$

$$\frac{\partial}{\partial \xi_i} L_p = C - \alpha_i - \beta_i = 0 \quad \Rightarrow \alpha_i = C - \beta_i$$

$$but\ \alpha_i \ \&\ \beta_i \geq 0 \Rightarrow 0 \leq \ \alpha_i \leq C$$

- Can now write dual problem (to maximize):

L_D:

Non-Separable SVMs

$$L_p: \min \frac{1}{2}||w||^2 + C\sum_i \xi_i - \sum_i \alpha_i\,(y_i(w^T x_i + b\,) - 1 + \xi_i) - \sum_i \beta_i \xi_i$$

$$\frac{\partial}{\partial w} L_p \; and \; \frac{\partial}{\partial b} L_p \;\; as\ before \ldots$$

$$\frac{\partial}{\partial \xi_i} L_p = C - \alpha_i - \beta_i = 0 \qquad \Rightarrow \alpha_i = C - \beta_i$$

$$but\ \alpha_i \,\&\, \beta_i \geq 0 \Rightarrow 0 \leq \ \alpha_i \leq C$$

- Can now write dual problem (to maximize):

$$L_D: \max \sum_i \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i{}^T\, x_j$$

$$subject\ to\ \sum_i \alpha_i y_i = 0 \;\; and\ \alpha_i \in [0, C]$$

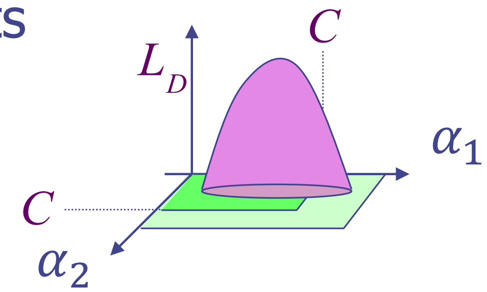
- Same dual as before but alphas can't grow beyond C

Non-Separable SVMs

- As we try to enforce a classification for a data point its KKT multiplier alpha keeps growing endlessly
- Clamping alpha to stop growing at C makes SVM "give up" on those non-separable points

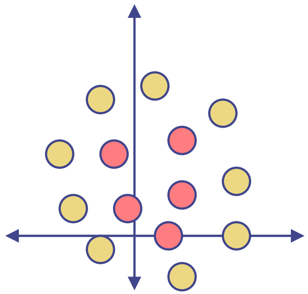
Non-Separable SVMs

- As we try to enforce a classification for a data point its KKT multiplier alpha keeps growing endlessly
- Clamping alpha to stop growing at C makes SVM “give up” on those non-separable points
- The dual program is now:
- Solve as before with extra constraints that alphas positive AND less than C… gives alphas… from alphas get $w = \sum_i \alpha_i y_i x_i$



Nonlinear SVMs

- What if the problem is not linear?



$$\begin{bmatrix} x(1) \\ x(2) \end{bmatrix} \rightarrow \begin{bmatrix} x(1) \\ x(2) \\ x(1)^2 + x(2)^2 \end{bmatrix}$$

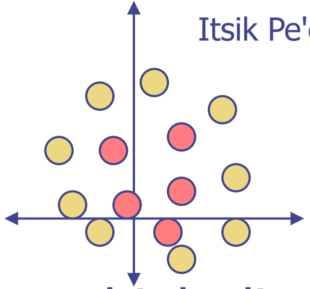
$x(1)^2$

$x(1)\,x(2)$

Classifying Antarctica

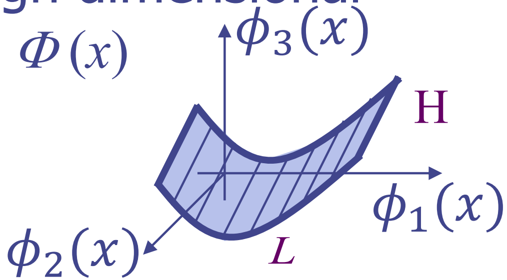
Itsik Pe'er, Columbia University

Nonlinear SVMs



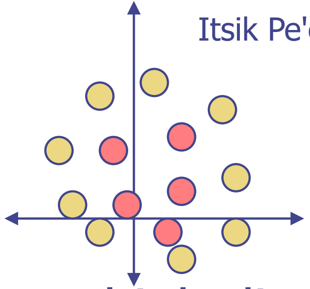
- What if the problem is not linear?
- We can use our old trick…
- Map d-dimensional x data from L-space to high dimensional H (Hilbert) feature-space via basis functions $\Phi(x)$
- For example, quadratic classifier:

$$x_i \rightarrow \phi(x_i) \;\; via \;\; \phi(\vec{x}) = \begin{bmatrix} \vec{x} \\ vec(\vec{x}\vec{x}^T) \end{bmatrix}$$



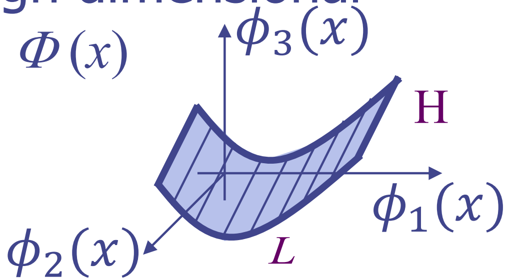
Itsik Pe'er, Columbia University

Nonlinear SVMs



- What if the problem is not linear?
- We can use our old trick…
- Map d-dimensional x data from L-space to high dimensional H (Hilbert) feature-space via basis functions $\Phi(x)$
- For example, quadratic classifier:

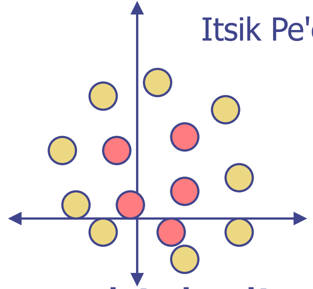
$$x_i \to \phi(x_i) \;\; via \;\; \phi(\vec{x}) = \begin{bmatrix} \vec{x} \\ vec(\vec{x}\vec{x}^T) \end{bmatrix}$$



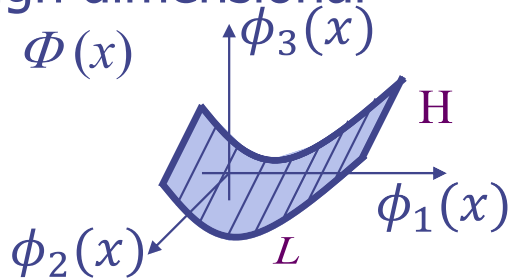
- Call ϕ's feature vectors computed from original x inputs

Itsik Pe'er, Columbia University

Nonlinear SVMs



- What if the problem is not linear?
- Map d-dimensional x data from L-space to high dimensional H (Hilbert) feature-space via basis functions $\Phi(x)$



- For example, quadratic classifier:

$$x_i \to \phi(x_i) \;\; via \;\; \phi(\vec{x}) = \begin{bmatrix} \vec{x} \\ vec(\vec{x}\vec{x}^T) \end{bmatrix}$$

- Call ϕ's feature vectors computed from original x inputs

- Dual qp used to be:

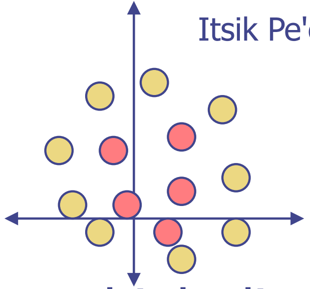
$$L_D\!: \max \sum_i \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j {x_i}^T x_j \;\; s.t.\, \alpha_i \geq 0 \,, \sum_i y_i \alpha_i = 0$$

- With linear classifier in original space:

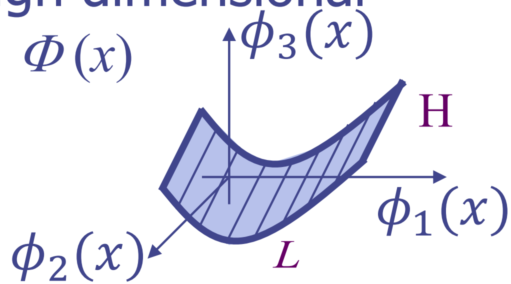
$$f(x) = sign\left(\sum_i \alpha_i y_i {x_i}^T x + b\right)$$

Itsik Pe'er, Columbia University

Nonlinear SVMs



- What if the problem is not linear?
- Map d-dimensional x data from L-space to high dimensional H (Hilbert) feature-space via basis functions $\Phi(x)$
- For example, quadratic classifier:



$$x_i \rightarrow \phi(x_i) \;\; via \;\; \phi(\vec{x}) = \begin{bmatrix} \vec{x} \\ vec(\vec{x}\vec{x}^T) \end{bmatrix}$$

- Call ϕ's feature vectors computed from original x inputs
- Replace all x's in the SVM equations with ϕ's
- Now solve the following learning problem:

$$L_D : \max \sum_i \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j \underline{\phi(x_i)^T \phi(x_j)} \;\; s.t. \, \alpha_i \geq 0 \, , \sum_i y_i \alpha_i = 0$$

- Which gives a nonlinear classifier in original space:

$$f(x) = sign\left(\sum_i \alpha_i y_i \underline{\phi(x_i)^T \phi(x)} + b \right)$$

Summary

- Nonseparable SVM
- Kernels