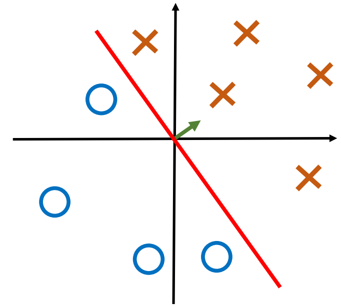
Support Vector Machines

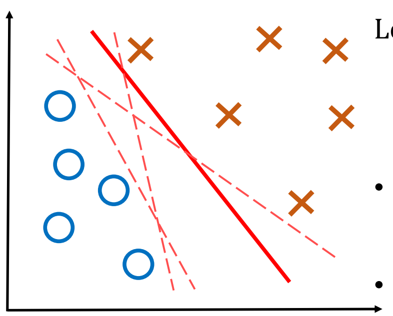
CISC 5800
Professor Daniel Leeds

Separating boundary, defined by **w**



- Separating **hyperplane** splits **class 0** and **class 1**
- Plane is defined by line **w** perpendicular to plan
- Is data point **x** in class 0 or class 1? $\mathbf{w}^T\mathbf{x}+b > 0$ class **1**
  $\mathbf{w}^T\mathbf{x}+b < 0$ class **0**

But, where do we place the boundary?



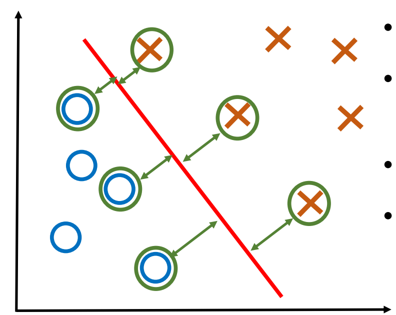
Logistic classifier:

$LL(y|x;w)$:

$$\sum_i (y^i - 1)\boldsymbol{w}^T\boldsymbol{x}^i - \log\left(1 + e^{-\boldsymbol{w}^T\boldsymbol{x}^i}\right)$$

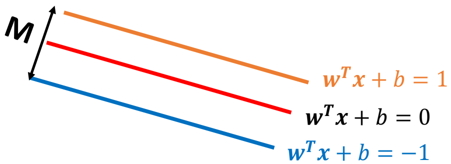
- Each data point $\boldsymbol{x}^i$ considered for boundary $\boldsymbol{w}$
- Outlier data pulls boundary towards it

Max margin classifiers



- Focus on boundary points
- Find largest margin between boundary points on both sides
- Works well in practice
- We can call the boundary points **"support vectors"**

Maximum margin definitions



Classify as +1
if $\boldsymbol{w}^T\boldsymbol{x} + b \geq 1$

Classify as -1
if $\boldsymbol{w}^T\boldsymbol{x} + b \leq -1$

Undefined
if $-1 < \boldsymbol{w}^T\boldsymbol{x} + b < 1$

- M is the margin width
- x^+ is a +1 point closest to boundary, x^- is a -1 point closest to boundary
- $\boldsymbol{x}^+ = \lambda \boldsymbol{w} + \boldsymbol{x}^-$
- $|\boldsymbol{x}^+ - \boldsymbol{x}^-| = \boldsymbol{M}$

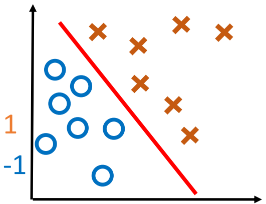
$$M = \frac{2}{\sqrt{\boldsymbol{w}^T\boldsymbol{w}}}$$

maximize M **minimize $\boldsymbol{w}^T\boldsymbol{w}$**

5

Support vector machine (SVM) optimization

argmin$_w$ $\boldsymbol{w}^T\boldsymbol{w}$
subject to
$\boldsymbol{w}^T\boldsymbol{x} + b \geq 1$ for **x** in class 1
$\boldsymbol{w}^T\boldsymbol{x} + b \leq -1$ for **x** in class -1



argmin$_w$ $\boldsymbol{w}^T\boldsymbol{w} + \left(\sum_{i \in +1} \lambda_i \left(1 - \left(\boldsymbol{w}^T\boldsymbol{x}^i + b\right)\right) + \right.$

9

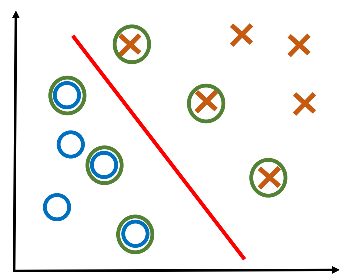
Support vector machine (SVM) optimization

argmin$_w$ $\boldsymbol{w}^T\boldsymbol{w}$
subject to
$\boldsymbol{w}^T\boldsymbol{x} + b \geq 1$ for **x** in class 1
$\boldsymbol{w}^T\boldsymbol{x} + b \leq -1$ for **x** in class -1

$$\frac{\partial}{\partial w_i}\left(\sum_j w_j^2 + \lambda_{i+}\sum_i\left(1 - \left(\sum_j w_j x_j^{i+} + b\right)\right) + \lambda_{i-}\sum_i\left(\left(\sum_j w_j x_j^{i-} + b\right) - 1\right)\right)$$

10

Alternate SVM formulation



$$\boldsymbol{w} = \sum_i \alpha^i \boldsymbol{x}^i y^i$$

Support vectors $\boldsymbol{x_i}$ have $\alpha_i > 0$

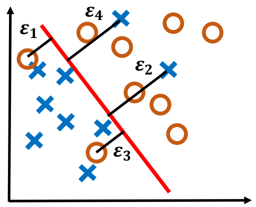
y_i are the data labels +1 or -1

$\boldsymbol{\alpha^i \geq 0 \ \forall i}$ $\quad \boldsymbol{\sum_i \alpha^i y^i = 0}$

11

Support vector machine (SVM) optimization
with slack variables

What if data not linearly separable?



$\text{argmin}_{w,b}\ \boldsymbol{w}^T\boldsymbol{w} + C\sum_i \varepsilon^i$

subject to

$\boldsymbol{w}^T\boldsymbol{x} + b \geq 1 - \varepsilon^i$ **for x in class 1**

$\boldsymbol{w}^T\boldsymbol{x} + b \leq -1 + \varepsilon^i$ **for x in class -1**

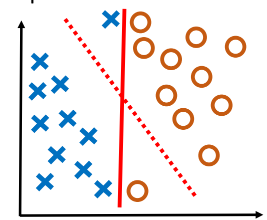
$\varepsilon^i \geq 0 \quad \forall i$

Each error ε^i is penalized based on distance from separator

14

Support vector machine (SVM) optimization
with slack variables

Example: Linearly separable but with narrow margins



$\text{argmin}_{w,b}\ \boldsymbol{w}^T\boldsymbol{w} + C\sum_i \varepsilon^i$

subject to

$\boldsymbol{w}^T\boldsymbol{x} + b \geq 1 - \varepsilon^i$ **for x in class 1**

$\boldsymbol{w}^T\boldsymbol{x} + b \leq -1 + \varepsilon^i$ **for x in class -1**

$\varepsilon_i \geq 0 \quad \forall i$

15

Hyper-parameters for learning

$\text{argmin}_{w,b}\ \boldsymbol{w}^T\boldsymbol{w} + C\sum_i \varepsilon_i$

Optimization constraints: ***C*** influences tolerance for label errors versus narrow margins

$$w_j \leftarrow w_j + \varepsilon \boldsymbol{x}_j^i\left(y^i - g(w^T\boldsymbol{x}^i)\right) - \frac{w_j}{\lambda}$$

Gradient ascent:

- ***ε*** influences effect of individual data points in learning
- ***T*** number of training examples, ***L*** number of loops through data – balance learning and over-fitting

Regularization: ***λ*** influences the strength of your prior belief

16

Hyper-parameters to learn

Each data point $\boldsymbol{x}^i$ has *N* features (presuming classify with $\boldsymbol{w}^T\boldsymbol{x}^i+b$)

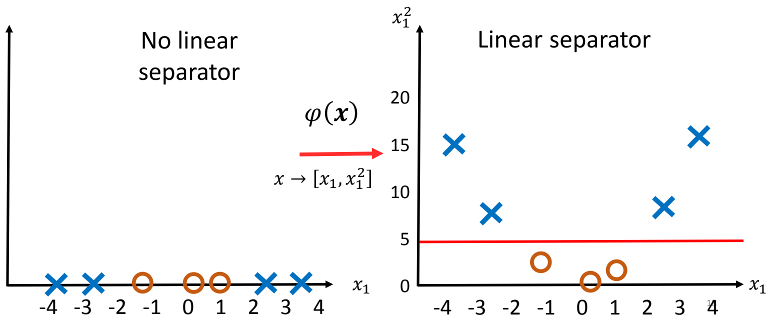
Separator: ***w*** and *b*

- *N* elements of **w**, 1 value for *b*: *N+1* parameters **OR**
- *t* support vectors -> *t* non-zero α^i, 1 value for *b*: *t+1* parameters

17

Classifying with additional dimensions

Note: More dimensions makes it easier to separate T training points: training error minimized, may risk over-fit



Quadratic mapping function (math)

$w^T x^k + b = \sum_i \alpha^i y^i (x^i)^T x^k + b$

x_1, x_2, x_3, x_4 -> $x_1, x_2, x_3, x_4, x_1^2, x_2^2, \ldots, x_1x_2, x_1x_3, \ldots, x_2x_4, x_3x_4$

N features -> $N + N + \frac{N\times(N-1)}{2} \approx N^2$ features

N^2 values to learn for w in higher-dimensional space

Or, observe: $(\boldsymbol{v}^T\boldsymbol{x}+1)^2 = \boldsymbol{v}_1^2 x_1^2 + \cdots + \boldsymbol{v}_N^2 x_N^2$
$+\boldsymbol{v}_1\boldsymbol{v}_2 x_1 x_2 + \cdots + \boldsymbol{v}_{N-1}\boldsymbol{v}_N x_{N-1} x_N$
$+\boldsymbol{v}_1 x_1 + \cdots + \boldsymbol{v}_N x_N$

v with N elements operating in quadratic space

19

Quadratic mapping function *Simplified*

x= $[x_1, x_2]$ -> $[\sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2, 1]$

$\mathbf{x}^i=[5,-2]$ -> $\mathbf{x}^k=[3,-1]$ ->

$\varphi(\boldsymbol{x}^i)^T\varphi(\boldsymbol{x}^k) =$

Or, observe: $\left(\boldsymbol{x}^{i^T}\boldsymbol{x}^k + 1\right)^2 =$

21

Mapping function(s)

- Map from low-dimensional space $\boldsymbol{x} = (x_1, x_2)$ to higher dimensional space $\varphi(\boldsymbol{x}) = (\sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2, 1)$
- N data points guaranteed to be separable in space of N-1 dimensions or more

$$\boldsymbol{w} = \sum_i \alpha_i \varphi(\boldsymbol{x}^i) y^i$$

Classifying $\boldsymbol{x}^k$:

$$\sum_i \alpha_i y^i \varphi(\boldsymbol{x}^i)^T \varphi(\boldsymbol{x}^k) + b$$

22

Kernels

Classifying $\boldsymbol{x}^{\boldsymbol{k}}$:

$$\sum_i \alpha_i y^i \varphi(\boldsymbol{x}^{\boldsymbol{i}})^T \varphi(\boldsymbol{x}^{\boldsymbol{k}}) + b$$

Kernel trick:

- Estimate high-dimensional dot product with function
- $K(\boldsymbol{x}^{\boldsymbol{i}}, \boldsymbol{x}^{\boldsymbol{k}}) = \varphi(\boldsymbol{x}^{\boldsymbol{i}})^T \varphi(\boldsymbol{x}^{\boldsymbol{k}})$

Now classifying $\boldsymbol{x}^k$

$$\sum_i \alpha_i y^i K(\boldsymbol{x}^{\boldsymbol{i}}, \boldsymbol{x}^{\boldsymbol{k}}) + b$$

23

Radial Basis Kernel

Try projection to infinite dimensions

$$\varphi(\boldsymbol{x}) = [x_1 \ , \cdots, x_n \ , x_1^2, \cdots, x_n^2, \cdots, x_1^\infty \cdots, x_n^\infty]$$

Taylor expansion: $e^x = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^\infty}{\infty!}$

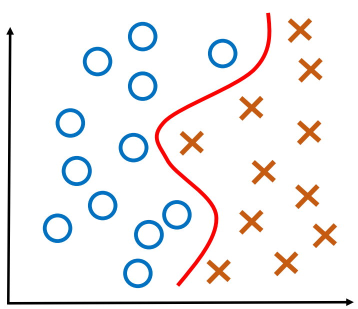
$$K(\boldsymbol{x}^{\boldsymbol{i}}, \boldsymbol{x}^{\boldsymbol{k}}) = \exp\left(-\frac{(\boldsymbol{x}^{\boldsymbol{i}} - \boldsymbol{x}^{\boldsymbol{k}})^2}{2\sigma^2}\right)$$

Note: $(\boldsymbol{x}^{\boldsymbol{i}} - \boldsymbol{x}^{\boldsymbol{k}})^2 = (\boldsymbol{x}^{\boldsymbol{i}} - \boldsymbol{x}^{\boldsymbol{k}})^T (\boldsymbol{x}^{\boldsymbol{i}} - \boldsymbol{x}^{\boldsymbol{k}})$

Draw separating plane to curve around all support vectors

24

Example RBF-kernel separator

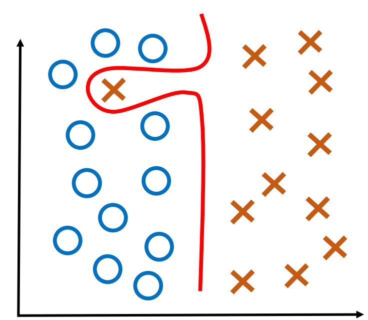


Large margin

Non-linear separation

25

Potential dangers of RBF-kernel separator



Small margin - **overfitting**

Non-linear separation

26

The power of SVM (+kernels)

Boundary defined by a few support vectors

- Caused by: maximizing margin
- Causes: less overfitting
- Similar to: regularization

Kernels keep number of learned parameters in check

27

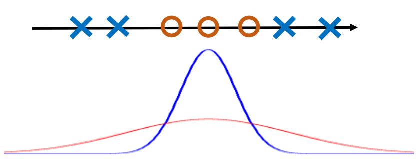
Binary -> *M*-class classification

- Learn boundary for class *m* vs all other classes
  - Only need M-1 separators for M classes – M^{th} class is for data outside of classes 1, 2, 3, …, M-1
- Find boundary that gives highest margin for data points $\mathbf{x}^i$

28

Benefits of generative methods

- $P(\boldsymbol{D}|\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}|\boldsymbol{D})$ can generate non-linear boundary
- E.g.: Gaussians with multiple variances



29