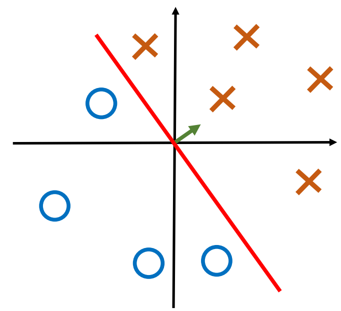
Support Vector Machines

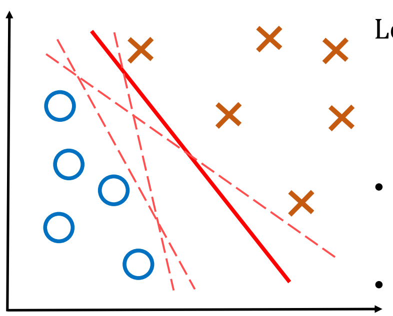
CISC 5800
Professor Daniel Leeds

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?



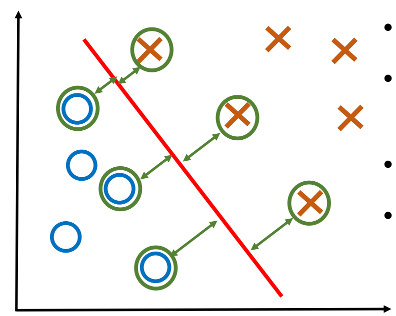
Logistic classifier:

$LL(y|x;w)$:

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

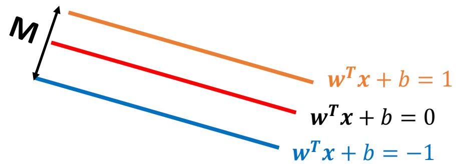
- Each data point $\boldsymbol{x}^i$ considered for boundary $\boldsymbol{w}$
- Outlier data pulls boundary towards it

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 **<u>"support vectors"</u>**

Maximum margin definitions



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

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

Undefined
if $-1 < w^T x + b < 1$

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

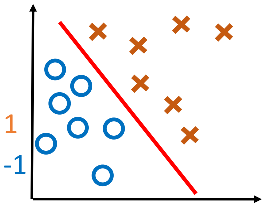
$$M = \frac{2}{\sqrt{w^T w}}$$

maximize M **minimize $w^T w$**

5

Support vector machine (SVM) optimization

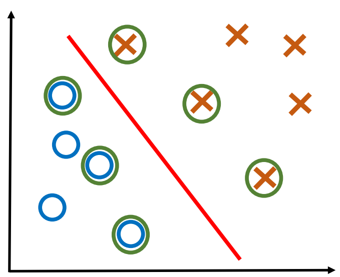
argmin$_w$ $w^T w$
subject to
$w^T x + b \geq 1$ for **x** in class 1
$w^T x + b \leq -1$ for **x** in class -1



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

9

Alternate SVM formulation



$$w = \sum_i \alpha^i x^i y^i$$

Support vectors x_i have $\alpha_i > 0$

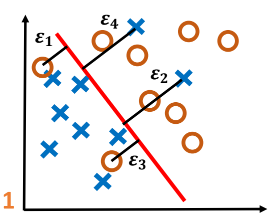
y_i are the data labels +1 or -1

$\alpha^i \geq 0 \ \forall i$ $\sum_i \alpha^i y^i = 0$

10

Support vector machine (SVM) optimization
with slack variables

What if data not completely ~~linearly~~ separable?



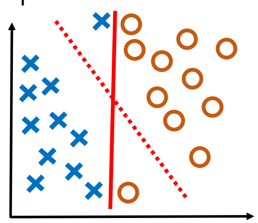
argmin$_{w,b}$ $w^T w + C \sum_i \varepsilon^i$
subject to
$w^T x + b \geq 1 - \varepsilon^i$ **for x in class 1**
$w^T x + b \leq -1 + \varepsilon^i$ **for x in class -1**
$\varepsilon^i \geq 0 \ \forall i$

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

13

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**

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

14

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

15

Parameter counts

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

16

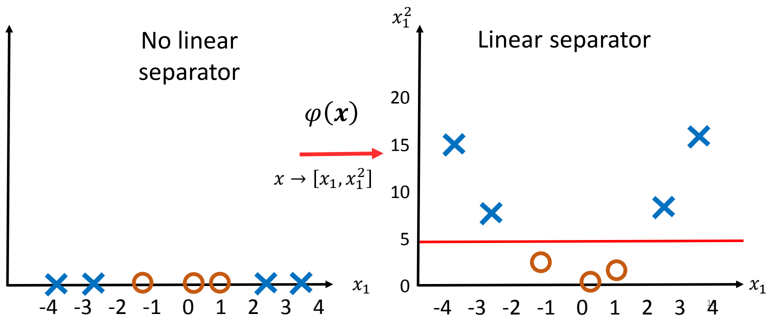
Binary -> *M*-class classification

- 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 $\mathbf{x}^i$

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)

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

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: $(\boldsymbol{v}^T\boldsymbol{x} + 1)^2 = v_1^2 x_1^2 + \cdots + v_N^2 x_N^2$
$+ v_1 v_2 x_1 x_2 + \cdots + v_{N-1} v_N x_{N-1} x_N$
$+ v_1 x_1 + \cdots + v_N x_N$

v with N elements operating in quadratic space

19