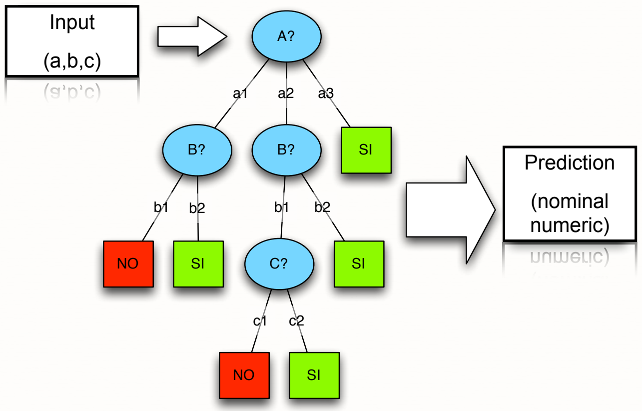
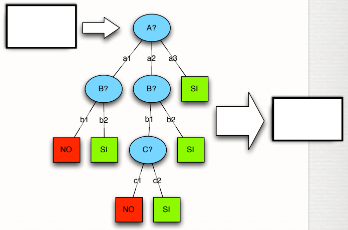
DECISION TREES AND RULES

DECISION TREES



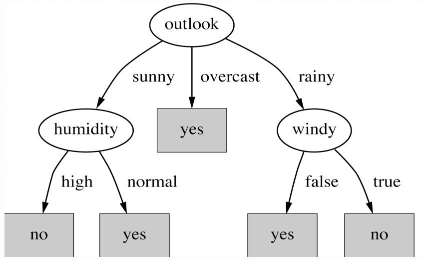
DECISION TREES

- A decision tree is built by
  - **Nodes** labeled by questions or **tests** about the value of one attribute
    - nominal: Which is the value?
    - continuous: Is the value less than or equal than a threshold?
  - **Leaves** labeled by prediction tags



WEATHER

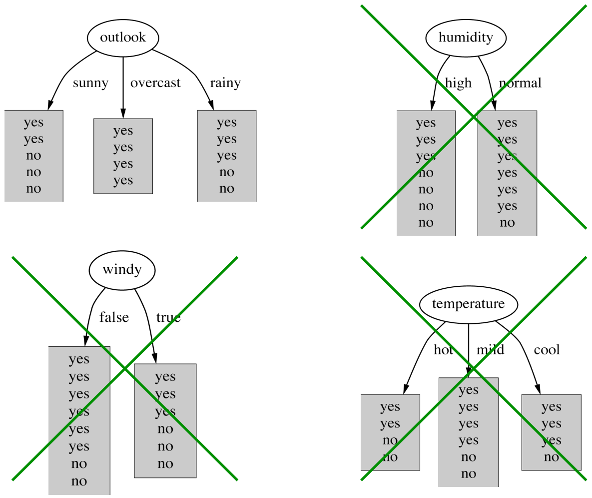
Classify a Saturday as right or not to play tennis according to weather conditions



Constructing decision trees

- Strategy: top down
  Recursive divide-and-conquer fashion
  - First: select attribute for root node
    Create branch for each possible attribute value
  - Then: split instances into subsets
    One for each branch extending from the node
  - Finally: repeat recursively for each branch, using only instances that reach the branch
- Stop if all instances have the same class

Which attribute to select?



Criterion for attribute selection

- Which is the best attribute?
  - Want to get the smallest tree
  - Heuristic: choose the attribute that produces the "purest" nodes
- Popular impurity criterion: information gain
  - Information gain increases with the average purity of the subsets
- Strategy: choose attribute that gives greatest information gain

C4.5

(Quinlan, 1993)

Computing information

- ID3 first, then C4.5
- Measure information in bits
  - Given a probability distribution, the info required to predict an event is the distribution's entropy
  - Entropy gives the information required in bits (can involve fractions of bits!)

C4.5: SELECTING THE BEST TEST

- C4.5 (Quinlan, 1993) employs the **entropy**

$$\text{info}(E) = -\sum_{j=1}^{k} p_j \cdot \log_2 p_j = -\sum_{j=1}^{k} \frac{\text{freq}(C_j, E)}{|E|} \cdot \log_2\left(\frac{\text{freq}(C_j, E)}{|E|}\right)$$

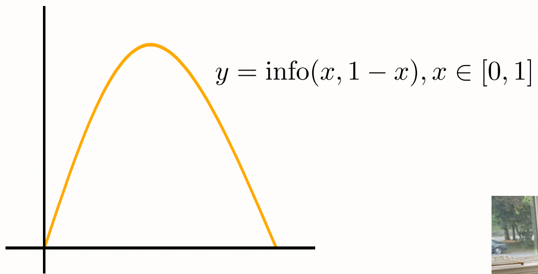
$$\text{info\_attribute}(E, X) = \sum_{i=1}^{n} \frac{|E_i|}{|E|} \cdot \text{info}(E_i)$$

$$\text{gain}(E, X) = \text{info}(E) - \text{info\_attribute}(E, X)$$

$$\text{ratio}(E, X) = \frac{\text{gain}(E, X)}{\text{intrinsic\_info}(E, X)}$$

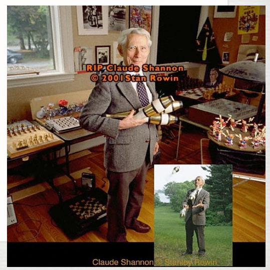
$$\text{intrinsic\_info}(E, X) = -\sum_{i=1}^{n} \frac{|E_i|}{|E|} \cdot \log_2\left(\frac{|E_i|}{|E|}\right)$$

C4.5: SELECTING THE BEST TEST



$$\text{info}(1, 0) = \text{info}(0, 1) = 0$$

$$\text{info}(1/n, \ldots, 1/n) = \log_2(n)$$

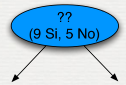


AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
3)	Nublado	Alta	Alta	Flojo	Si
4)	Lluvia	Moderada	Alta	Flojo	Si
5)	Lluvia	Baja	Normal	Flojo	Si
6)	Lluvia	Baja	Normal	Fuerte	No
7)	Nublado	Baja	Normal	Fuerte	Si
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
10)	Lluvia	Moderada	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si
12)	Nublado	Moderada	Alta	Fuerte	Si
13)	Nublado	Alta	Normal	Flojo	Si
14)	Lluvia	Moderada	Alta	Fuerte	No

AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
3)	Nublado	Alta	Alta	Flojo	Si
4)	Lluvia	Moderada	Alta	Flojo	Si
5)	Lluvia	Baja	Normal	Flojo	Si
6)	Lluvia	Baja	Normal	Fuerte	No
7)	Nublado	Baja	Normal	Fuerte	Si
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
10)	Lluvia	Moderada	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si
12)	Nublado	Moderada	Alta	Fuerte	Si
13)	Nublado	Alta	Normal	Flojo	Si
14)	Lluvia	Moderada	Alta	Fuerte	No

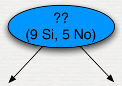


$$\mathrm{info}(E) = -\sum_{j=1}^{k} \frac{\mathrm{frec}(C_j, E)}{|E|} \cdot \log_2\left(\frac{\mathrm{frec}(C_j, E)}{|E|}\right)$$

$$\mathrm{info}(E) = -\left[\frac{9}{14} \cdot \log_2\left(\frac{9}{14}\right) + \frac{5}{14} \cdot \log_2\left(\frac{5}{14}\right)\right] = \boxed{0{,}940}$$

AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
3)	Nublado	Alta	Alta	Flojo	Si
4)	Lluvia	Moderada	Alta	Flojo	Si
5)	Lluvia	Baja	Normal	Flojo	Si
6)	Lluvia	Baja	Normal	Fuerte	No
7)	Nublado	Baja	Normal	Fuerte	Si
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
10)	Lluvia	Moderada	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si
12)	Nublado	Moderada	Alta	Fuerte	Si
13)	Nublado	Alta	Normal	Flojo	Si
14)	Lluvia	Moderada	Alta	Fuerte	No



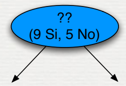
$$\mathrm{info}(E_{Pron.=Soleado}) = -\left[\frac{2}{5} \cdot \log_2\left(\frac{2}{5}\right) + \frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right)\right] = 0{,}971$$

$$\mathrm{info}(E_{Pron.=Nublado}) = -\left[\frac{4}{4} \cdot \log_2\left(\frac{4}{4}\right) + \frac{0}{4} \cdot \log_2\left(\frac{0}{4}\right)\right] = 0$$

$$\mathrm{info}(E_{Pron.=Lluvia}) = -\left[\frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) + \frac{2}{5} \cdot \log_2\left(\frac{2}{5}\right)\right] = 0{,}971$$

AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
3)	Nublado	Alta	Alta	Flojo	Si
4)	Lluvia	Moderada	Alta	Flojo	Si
5)	Lluvia	Baja	Normal	Flojo	Si
6)	Lluvia	Baja	Normal	Fuerte	No
7)	Nublado	Baja	Normal	Fuerte	Si
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
10)	Lluvia	Moderada	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si
12)	Nublado	Moderada	Alta	Fuerte	Si
13)	Nublado	Alta	Normal	Flojo	Si
14)	Lluvia	Moderada	Alta	Fuerte	No



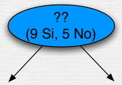
$$\mathrm{info}(E_{Pron.=Soleado}) = -\left[\frac{2}{5} \cdot \log_2\left(\frac{2}{5}\right) + \frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right)\right] = 0{,}971$$

$$\mathrm{info}(E_{Pron.=Nublado}) = -\left[\frac{4}{4} \cdot \log_2\left(\frac{4}{4}\right) + \frac{0}{4} \cdot \log_2\left(\frac{0}{4}\right)\right] = 0$$

$$\mathrm{info}(E_{Pron.=Lluvia}) = -\left[\frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) + \frac{2}{5} \cdot \log_2\left(\frac{2}{5}\right)\right] = 0{,}971$$

AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
3)	Nublado	Alta	Alta	Flojo	Si
4)	Lluvia	Moderada	Alta	Flojo	Si
5)	Lluvia	Baja	Normal	Flojo	Si
6)	Lluvia	Baja	Normal	Fuerte	No
7)	Nublado	Baja	Normal	Fuerte	Si
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
10)	Lluvia	Moderada	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si
12)	Nublado	Moderada	Alta	Fuerte	Si
13)	Nublado	Alta	Normal	Flojo	Si
14)	Lluvia	Moderada	Alta	Fuerte	No



$$\text{info}(E_{Pron.=Soleado}) = -\left[\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right)\right] = 0{,}971$$

$$\text{info}(E_{Pron.=Nublado}) = -\left[\frac{4}{4}\cdot\log_2\left(\frac{4}{4}\right)+\frac{0}{4}\cdot\log_2\left(\frac{0}{4}\right)\right] = 0$$

$$\text{info}(E_{Pron.=Lluvia}) = -\left[\frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right)+\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)\right] = 0{,}971$$

AN EXAMPLE ...

Therefore, the average info due to *Pronóstico* is:

$$\text{info\_atrib}(E, Pronóstico) = \frac{|E_{Pron.=Soleado}|}{|E|}\cdot\text{info}(E_{Pron.=Soleado}) +$$
$$+\frac{|E_{Pron.=Nublado}|}{|E|}\cdot\text{info}(E_{Pron.=Nublado}) + \frac{|E_{Pron.=Lluvia}|}{|E|}\cdot\text{info}(E_{Pron.=Lluvia}) =$$
$$= \frac{5}{14}\cdot 0{,}971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0{,}971 = 0{,}694$$

The *info gain* and the *ratio* are:

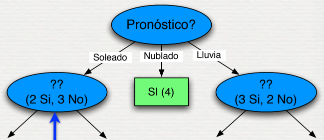
$$\text{ganancia}(E, Pronóstico) = \text{info}(E) - \text{info\_atrib}(E, Pronóstico) = 0{,}940 - 0{,}694 = 0{,}246$$

$$\text{info\_part}(E, Pronóstico) = -\left[\frac{5}{14}\log_2\left(\frac{5}{14}\right)+\frac{4}{14}\log_2\left(\frac{4}{14}\right)+\frac{5}{14}\log_2\left(\frac{5}{14}\right)\right] = 1{,}577$$

$$\text{ratio}(E, Pronóstico) = \frac{\text{ganancia}(E, Pronóstico)}{\text{info\_part}(E, Pronóstico)} = \frac{0{,}246}{1{,}577} = 0{,}156$$

AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si

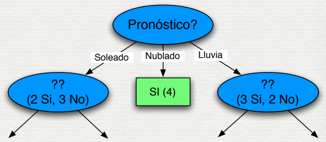


$$\text{info}(E') = -\left[\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right)\right] = 0{,}971$$

$$\text{info\_atrib}(E', Humedad) =$$
$$= \frac{|E'_{Hum.=Normal}|}{|E'|}\cdot\text{info}(E'_{Hum.=Normal}) + \frac{|E'_{Hum.=Alta}|}{|E'|}\cdot\text{info}(E'_{Hum.=Alta}) =$$
$$= \frac{2}{5}\left[\frac{2}{2}\log_2\left(\frac{2}{2}\right)+\frac{0}{2}\log_2\left(\frac{0}{2}\right)\right] + \frac{3}{5}\left[\frac{0}{3}\log_2\left(\frac{0}{3}\right)+\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] = 0$$

AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si

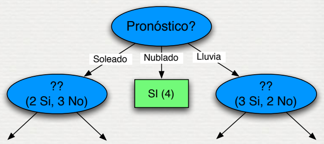


$$\text{info}(E') = -\left[\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right)\right] = 0{,}971$$

$$\text{info\_atrib}(E', Humedad) =$$
$$= \frac{|E'_{Hum.=Normal}|}{|E'|}\cdot\text{info}(E'_{Hum.=Normal}) + \frac{|E'_{Hum.=Alta}|}{|E'|}\cdot\text{info}(E'_{Hum.=Alta}) =$$
$$= \frac{2}{5}\left[\frac{2}{2}\log_2\left(\frac{2}{2}\right)+\frac{0}{2}\log_2\left(\frac{0}{2}\right)\right] + \frac{3}{5}\left[\frac{0}{3}\log_2\left(\frac{0}{3}\right)+\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] = 0$$

AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si



$$\text{info}(E') = -\left[\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right)\right] = 0{,}971$$

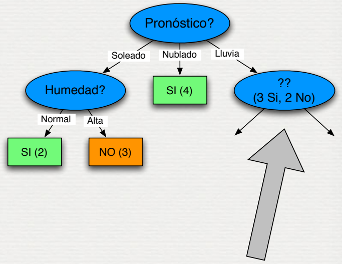
$$\text{ganancia}(E', \mathit{Humedad}) = 0{,}971 - 0 = 0{,}971$$

$$\text{info\_part}(E', \mathit{Humedad}) = -\left[\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right)\right] = 0{,}971$$

$$\text{ratio}(E', \mathit{Humedad}) = \frac{0{,}971}{0{,}971} = 1{,}0$$

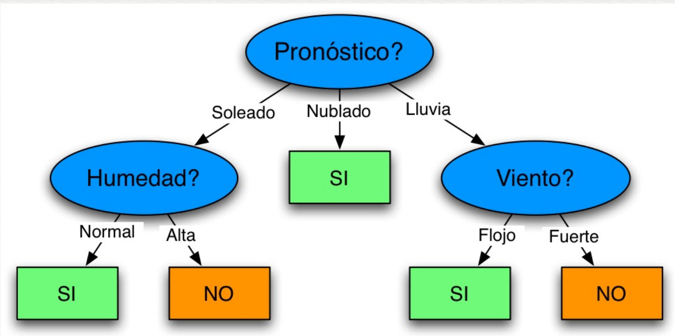
AN EXAMPLE ...

	Pronóstico	Temperatura	Humedad	Viento	Apto?
1)	Soleado	Alta	Alta	Flojo	No
2)	Soleado	Alta	Alta	Fuerte	No
3)	Nublado	Alta	Alta	Flojo	Si
4)	Lluvia	Moderada	Alta	Flojo	Si
5)	Lluvia	Baja	Normal	Flojo	Si
6)	Lluvia	Baja	Normal	Fuerte	No
7)	Nublado	Baja	Normal	Fuerte	Si
8)	Soleado	Moderada	Alta	Flojo	No
9)	Soleado	Baja	Normal	Flojo	Si
10)	Lluvia	Moderada	Normal	Flojo	Si
11)	Soleado	Moderada	Normal	Fuerte	Si
12)	Nublado	Moderada	Alta	Fuerte	Si
13)	Nublado	Alta	Normal	Flojo	Si
14)	Lluvia	Moderada	Alta	Fuerte	No



We have to repeat the process to build the tree below this node

AN EXAMPLE ...



C4.5 is a *bit* more

1. Numeric values
2. Pruning of trees
3. Missing values handling

ID3 → C4.5: Numeric attributes

- **Standard method: binary splits**
  - •E.g. temp < 45
- **Unlike nominal attributes, every attribute has many possible split points. Solution is straightforward extension:**
  - Evaluate info gain (or other measure) for every possible split point of attribute
  - Choose "best" split point
  - *Info gain* for best split point is info gain for attribute
- **Computationally more demanding**

Weather data

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…				

If outlook = sunny
If outlook = rainy
If outlook = overc
If humidity = norm
If none of the abo

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes

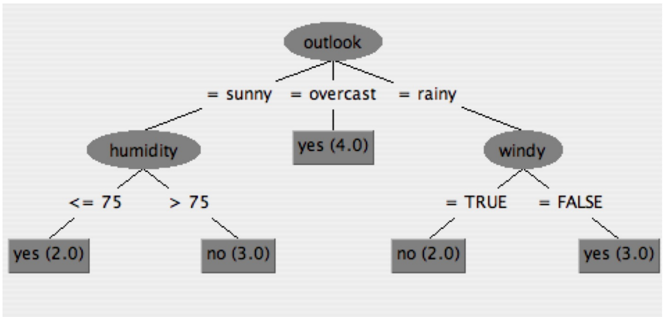
Example

- **Split on temperature attribute:**

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

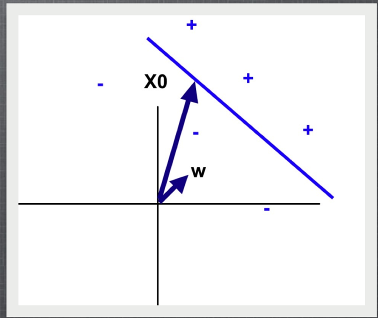
  - E.g. temperature < 71.5: yes/4, no/2
    temperature ≥ 71.5: yes/5, no/3
  - Info([4,2],[5,3])
    = 6/14 info([4,2]) + 8/14 info([5,3])
    = 0.939 bits
- **Place split points halfway between values** **(J48 does not proceed like this)**
- **Can evaluate all split points in one pass!**

Example



SVM: SUPPORT VECTOR MACHINES

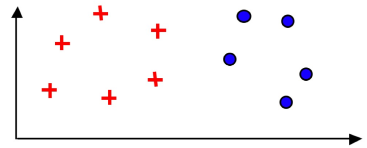
Scalar Products
Optimization



Support Vector Machines

- Classifying with scalar products
- Maximizing margin

S= {($\mathbf{x_i}$,y_i): $\boldsymbol{x_i} \in \mathbf{H}$, $y_i \in \{+1, -1\}$; i = 1, …,m}



Scalar product

- **Definition.-**

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = (x_1 \quad x_2 \quad \dots \quad x_n) \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \mathbf{x}^T \mathbf{y}$$

- **Properties.-**
  (1) **<x, y > = <y,x>**
  (2) <**x, (y+z)**> = <**x, y**> + <**x, z**>
  (3) <a **x, y** > = <**x,** a **y** > = a <**x, y** >

Scalar product (2)

Gives rise to a norm

$$\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$$

and then a metric (distance)

$$\begin{aligned}(d_H(\boldsymbol{x}, \boldsymbol{x}'))^2 = \|\boldsymbol{x} - \boldsymbol{x}'\|^2 &= \langle \boldsymbol{x} - \boldsymbol{x}', \boldsymbol{x} - \boldsymbol{x}' \rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle - 2\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + \langle \boldsymbol{x}', \boldsymbol{x}' \rangle\end{aligned}$$

Scalar product (3)

Geometric meaning

$$\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n \qquad \langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \|\boldsymbol{x}\| \cdot \|\boldsymbol{x}'\| \cdot \cos(\boldsymbol{x}, \boldsymbol{x}')$$

if two vectors have norm 1, their scalar product ranges from:

- +1 (same) to
- -1 (opposite).
- 0 (perpendicular)

Scalar product (4)

Geometric meaning

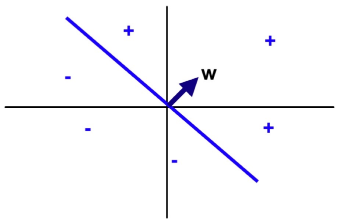
If we do not consider their lengths, the scalar product of 2 vectors measures their **similarity**

$$\langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \|\boldsymbol{x}\| \cdot \|\boldsymbol{x}'\| \cdot \cos(\boldsymbol{x}, \boldsymbol{x}')$$

Hyperplanes and scalar product

In a n-dimensional space, an **hyperplane** is a subspace of dimension n-1.

They are determined by one vector (**director vector**) perpendicular to the hyperplane



Equation:

<x,w> = 0

Splits the space in 2 regions

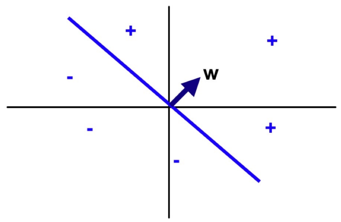
Pos= {**x**: <**x**,**w**> ≥ **0**}

Neg={**x**: <**x**,**w**> **< 0**}

Hyperplanes and scalar product

In a n-dimensional space, an **hyperplane** is a subspace of dimension n-1.

They are determined by one vector (**director vector**) perpendicular to the hyperplane



Equation:

<x,w> = 0

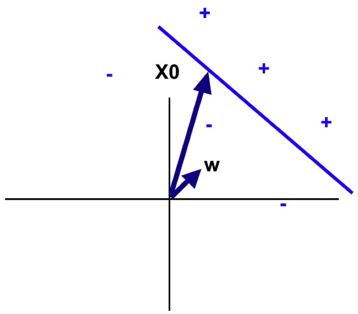
Splits the space in 2 regions

Pos= {**x**: <**x**,**w**> ≥ **0**}

Neg={**x**: <**x**,**w**> **< 0**}

Affine hyperplane

Def.- **hyperplanes** translated to a point $\mathbf{x_0}$ in the space



Equation:
<**x-x$_0$,w**> **=** <**x, w** > + b **= 0**

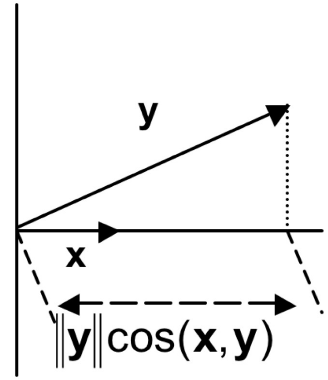
Splits the space in 2 regions
Pos= {**x**: <**x-x$_0$,w**> ≥ **0**}
Neg= {**x**: <**x-x$_0$,w**> **< 0**}

Distance to hyperplanes

$$if \quad \|x\| = 1, \quad then \quad \langle x, y \rangle = \|y\| cos(x, y)$$

The scalar product is

- the length of the projection of **y** in the direction of **x**
- the distance of ***y*** to the hyperplane determined by(perpendicular to) ***x***



Distance to affine hyperplanes

The **distance** from **x** ∈ H to an hyperplane
<w,x> + b = 0

$$d(x, Hyperplane) = \frac{\langle w, x \rangle + b}{\|w\|}$$

The **split** in two regions can be obtained by

y = sign (<***w, x***> + b) = sign(d(***x***, Hyperplane))

Distance to affine hyperplanes

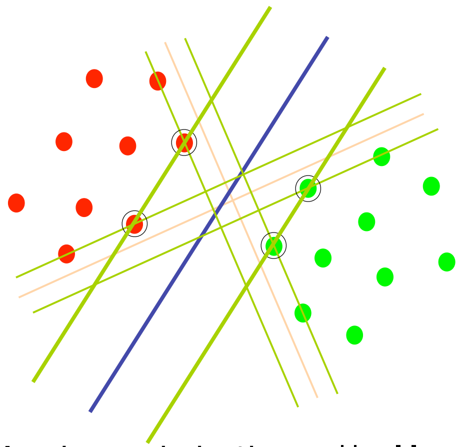
The **distance** from **x** ∈ H to an hyperplane
<w,x> + b = 0

$$d(x, Hyperplane) = \frac{\langle w, x \rangle + b}{\|w\|}$$

The **split** in two regions can be obtained by

y = sign (<***w, x***> + b) = sign(d(***x***, Hyperplane))

The optimization of SVM

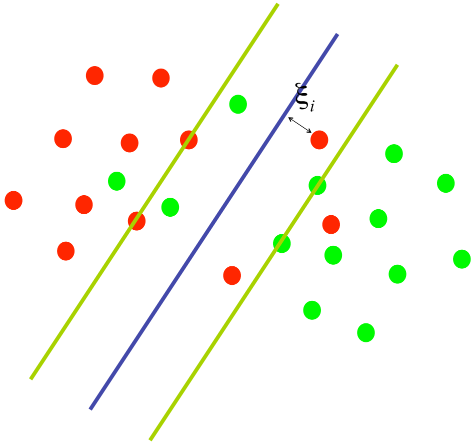


$$\text{margin} = \frac{1}{\|\mathbf{w}\|}$$

Margin maximization = ||**w**|| minimization

SVM in non separable datasets

Soft margin



SVM: convex optimization

To find the optimal soft-margin hyperplane =
Solving a quadratic convex optimization problem subject to some constraints

$$\begin{aligned} \min_{\mathbf{w},\xi,b} \quad & \frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle + C\sum_{i=1}^{m}\xi_i, \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \qquad i = 1, \ldots, m \end{aligned}$$

SVM: convex optimization

Solving a quadratic convex optimization:

- There are *very* fast optimizers
- Convex means:
  - there exist one global optimum and
  - no local optima (this is not the case in ANN)