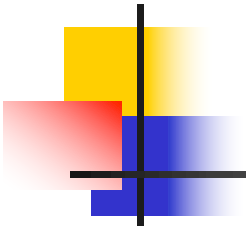
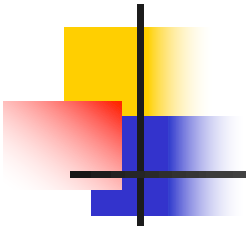
Machine Learning

ITAM



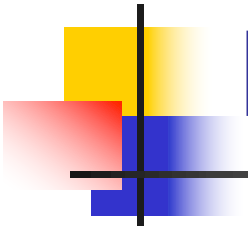
Decision Trees

- Supervised Learning
- Classification and regression algorithms
  - To approximate functions with a discreteresponse variable
  - To approximate functions with a continuous response variable



Decision Trees

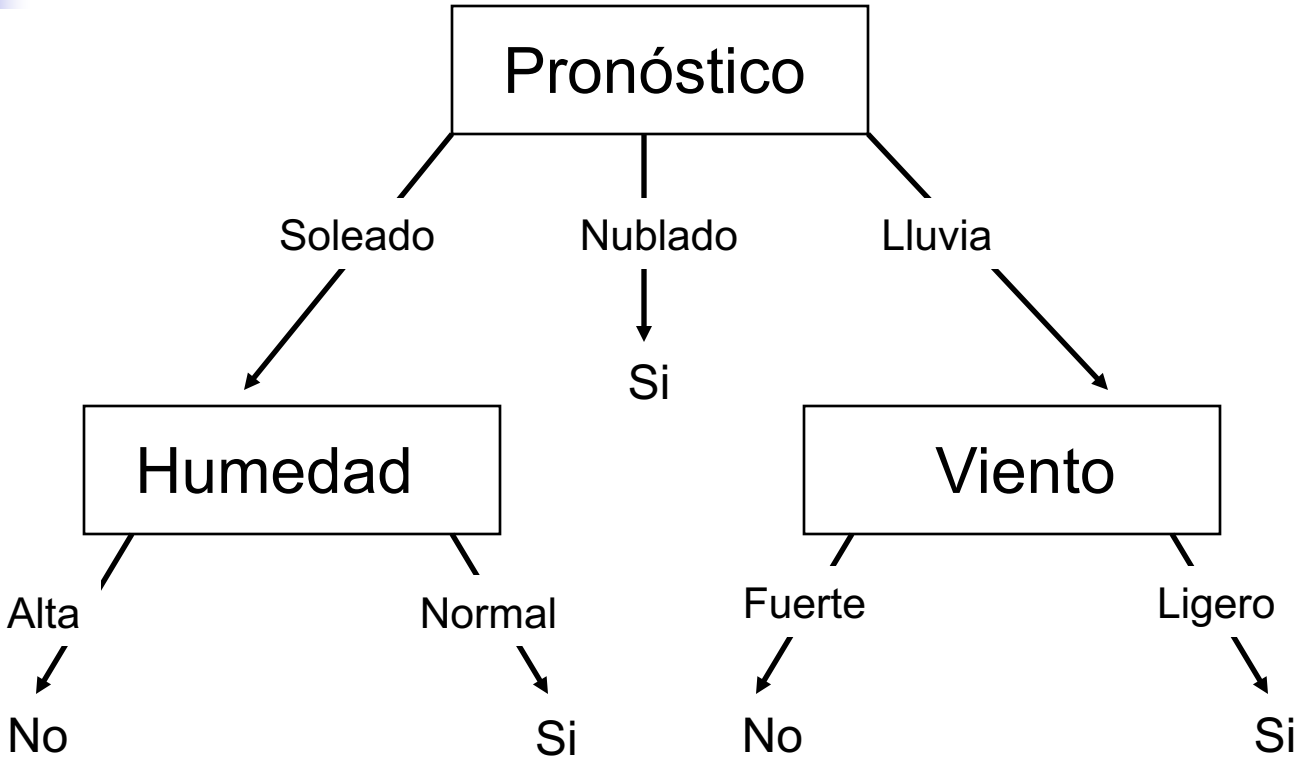
- These algorithms proceed by creating a tree in which each internal node corresponds to an attribute or feature of the problem instances
  - Terminal nodes or leafs indicate the classification of an instance
- Each node implements a test about the value of an attribute and each brach represents the resulting value
- An instance is classified by moving downwards int the tree, from the root to a leaf following the braches indicated by its attribute values



Decision Trees

- Example (taken from Quinlan 1986)
  - Instances have the following form:
    - <Pronostico,Temperatura, Humedad, Viento>
    - The possible values are:
    - Pronóstico $\in$ {Soleado, Nublado, Lluvia}
    - Temperatura $\in$ {Caliente, Templado, Frío}
    - Humedad $\in$ {Alta, Normal}
    - Viento $\in$ {Fuerte, Débil}
  - We want to classify a particular day as good or not to play tennis

Decision Trees



- For example, the instance:
  - (Pronostico=Soleado,Temperatura=Caliente, Humedad=Normal, Viento=Fuerte) would be considered good for tennis

Decision Trees

- Note: This tree doesn't use all attributes. Its possible that a subset of attribute is enough for the current task
- In genera a decision tree represents a disjunction of conjunctions. For example, days good for tennis can be characterized as:

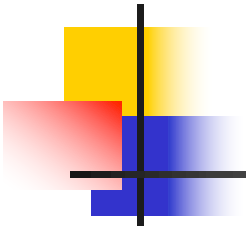
  ((Pronostico = Soleado) **and** (Humedad= Normal))

  or

  (Pronóstico= Nublado)

  or

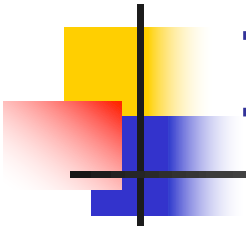
  ((Pronóstico=Lluvia) **and** (Viento=Débil))



Decision Trees

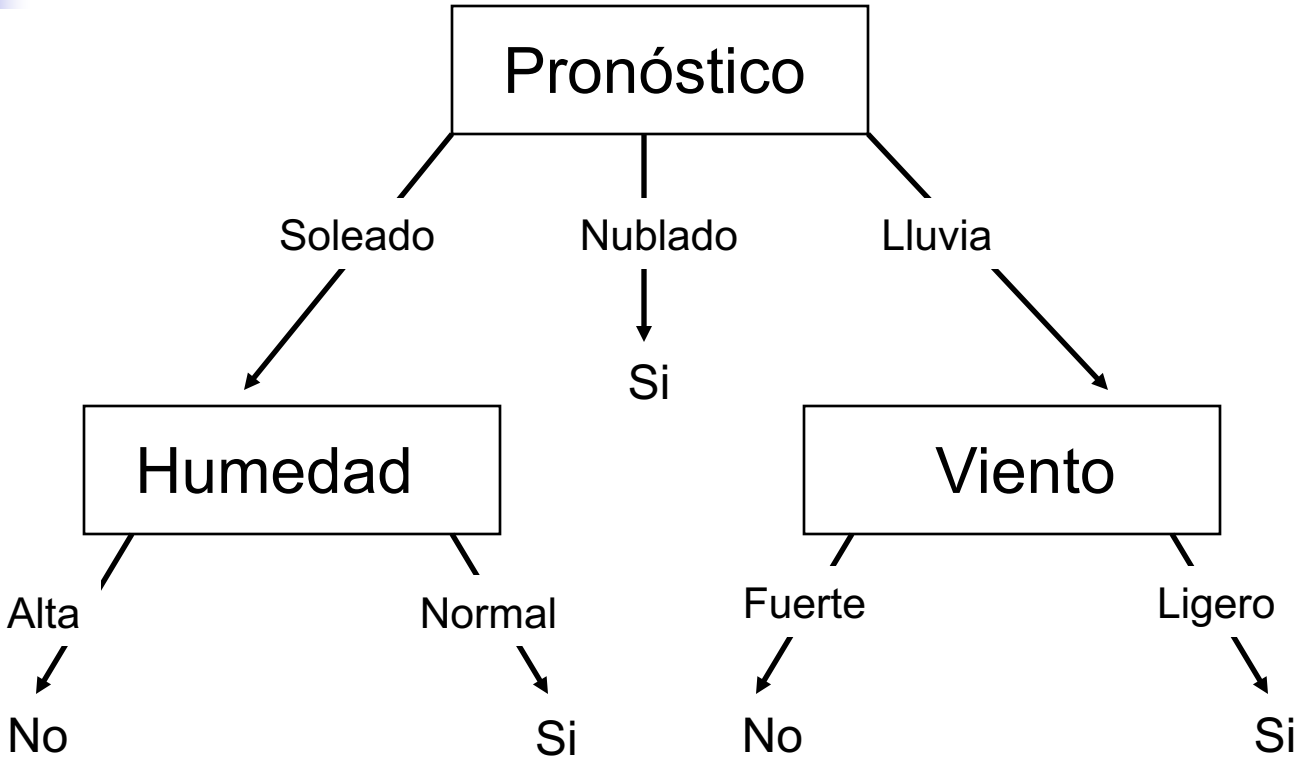
- We are going to focus on a particular kind of decision tree: ID3
- It only works for classification and discrete attributes
- It ilustrates the working of the algorithm
- The ideas are the same as those governing the trees that are actually used

Decision Trees
ID3



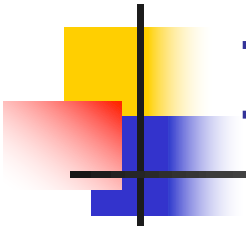
- Only categorical attributes
- This algorithm starts to construct the tree from the root
  - Which attribute should be tested first?
- Each attribute is statistically evaluated to determine how well it, by itself, is able to classify the training data
- Once an attribute is selected, there is a succesor node created for each of the possible values of the attribute
- To select the nodes of the next level, the previous process is repeated using ONLY the training examples corresponding to that branch---using the examples that have the proper values for the attributes for the nodes in that branch
- This process is repeated until all examples are classified or there are no more attributes to test

Decision Trees
ID3



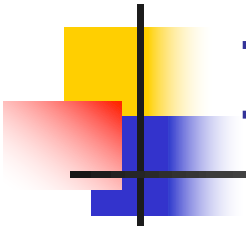
- Por ejemplo si se estima que Pronóstico es el atributo que por si solo clasifica mejor los ejemplos de entrenamiento, se pone en la raíz. Para cada nodo hijo se escoge un atributo de la misma forma, usando el subconjunto de los ejemplos de entrenamiento correspondientes

Decision Trees Information



- Which attribute is the best-stand alone classifier?
- Which attribute separates the training instances best?
- La medida que se utiliza es la **ganancia de información**
- What is information?
  - The amount of information contained in a message (or event) is the amount that it reduces our ignorance
    - The sun is coming out tomorrow
    - The lottery's winning number next week is : 23 45 21 34 91

Decision Trees Information



- The measure que are looking for must have high values for messages for whose content we have a low expectation and low for those we highly expect
- For example
  - A unbiased coin
    - P(aguila)=0.5 y P(sol)= 0.5
  - A bias coin
    - P(aguila)=0.9 y P(sol)=0.1
- Which event gives us more information: tossing the biased or the unbiased coin on average?

Decision Trees Information

- We use Entropy as the measure. The formula for a binary variable is :
  - $H(M) = -(P(\text{aguila})\log_2 P(\text{aguila}) + P(\text{sol})\log_2 P(\text{sol}))$
  - Maximizing H(M) we have
    - P(aguila)=P(sol)=0.5
- This means that if we have a binary classifier and of the 20 training examples 10 belong to one class and 10 the other with respect to a certain attribute, this attribute maximizes the entropy

Decision Trees
Information

- In general, the formula for *c* categories:
  - $H(M)=\Sigma_{i=1,\ c} -p_i \log_2 p_i$
  - The sum is from 1 hasta c, where p_i is lthe proportion of *M* that belongs to class *i*
  - *M* is the set of training examples
  - The base of the logarithm doesnt really matter as long as you keep the same one through all calculations

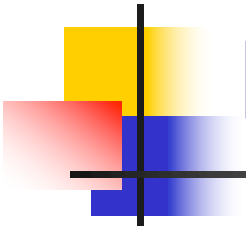
Decision Trees ID3
How to choose a node

- Coming back to how to choose a node to partition *M*
- Information gain is the expected reduction in entropy caused by partitioning *M* with respect to attribute *A (*How much does it reduce our uncertainty?)
- Gain(*M,A*)=

  $H(M)-\Sigma_{v\ \varepsilon\ Values(A)}\ (|M_v|/|M|)*H(M_v)$

  - Where Values(A) are the possible values of attribute A
  - The first term is the entropy of *M,* the second is the average entropy of the partition of M with respect to A
  - The Gain (*M,A*) is the reduction of entropy given A; the information that A provides about the data classification; how much does it reduce the uncertainty about an average instance class

Decision Trees ID3 Example



- Lets focus attention on only two attributs from the tennis example
  - (Viento, Humedad,….)
    - Each attribute has two possible values: (Fuerte y Débil) y (Alto y Normal) respectively
- Suppose we have 10 instances

Decision Trees ID3 Example

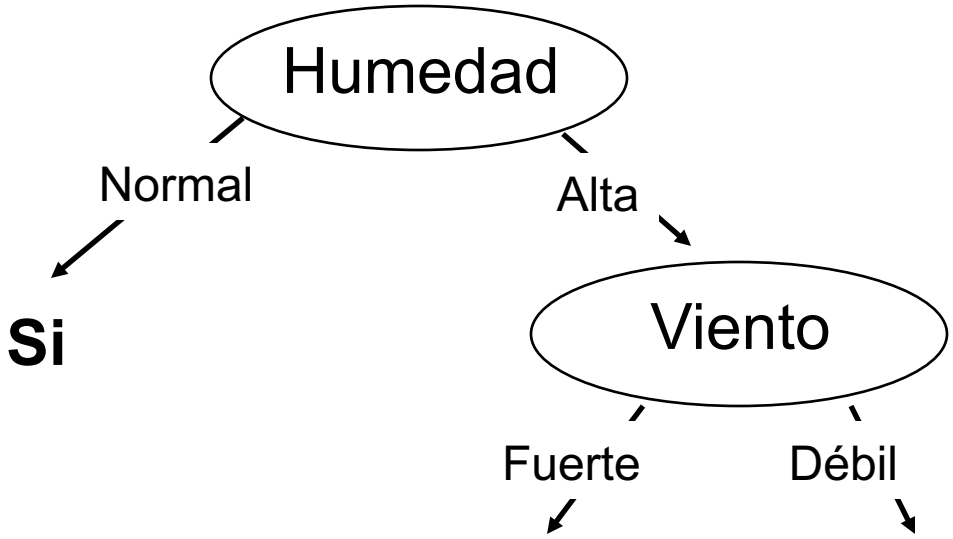
Viento	Humedad	Otros	Clase
Fuerte	Alta	...	no
Fuerte	Alta	...	no
Fuerte	Alta	...	no
Fuerte	Alta	...	no
Fuerte	Normal	...	si
Fuerte	Normal	...	si
Débil	Alta	...	no
Débil	Alta	...	no
Débil	Alta	...	si
Débil	Normal	...	si

H(M)=-(6/10)*$\log_2$(6/10) - (4/10)*$\log_2$(4/10)=0.97

Con respecto al Viento
H(M_F) =-(4/6)*$\log_2$(4/6) -(2/6)*$\log_2$(2/6)=0.92
H(M_D)= -(2/4)*$\log_2$(2/4) -(2/4)*$\log_2$(2/4)=1
Ganancia(M,Viento)= 0.97-6/10*0.92-4/10*1=0.018

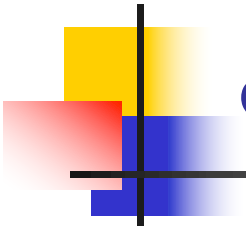
Con respecto a Humedad
H(M_A)= -(6/7)*$\log_2$(6/7) -(1/7)*$\log_2$(1/7)=0.59
H(M_N)= -(3/3)*$\log_2$(3/3) -(0/3)*$\log_2$(0/3)=0
Ganancia(M,Hume)= 0.97 -7/10*0.59=0.55

Decision Trees ID3 Example



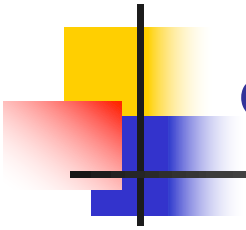
Si con 1/3
No con 2/3
Classified as **No**

Generalization to continuous attributes

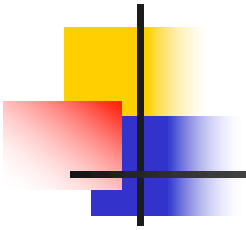


- So far only categorical attributes
- We can generalize to continuous attributes
- One strategy is simply to discretize the continuous attributes

Generalization to continuous attributes

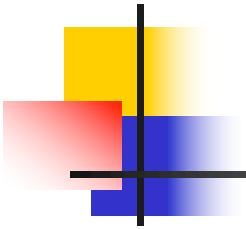


- One possibility is to divide the values of an attribute in ranges
  - Usually its divided only in two ranges. Empirically dividing in more ranges slows down the algorithm with no apparent range in predictive performance
  - Strategy: find the partition point that minimizes the error or that maximizes the information gain



Regression Trees

- The trees discussed so far can only be used for classification (intrinsic in information gain measure)
- We can extend the general principles of the method for regression problems
- The most comon instantiations of these algorithms are called CART (classification and regression trees)



Regression Trees

- We have to change the criterion for choosing nodes while creating the tree
- Strategy
  - The value of the model at a given node is the average value of the examples that it represents
  - We define the error as the mean squared error of the corresponding examples
  - We choose as root for the current subtree the node which leads to the least training error

Regression Trees

- Estrategia: Nodos Intermedios
  - Determinar los grupos en los que se divide un atributo
    - Si es categórico: las categorías
    - Si es continuo: los rangos
  - Para cada grupo calcular el error cuadrático medio de los ejemplos que le corresponden (como la diferencia de cada valor de entrenamiento con la media de los valores de entrenamiento para ese grupo (varianza))
  - Calcular el error del atributo como el promedio de los errores del grupo
  - Escoger el atributo con el mínimo error

$$\arg\min_{A}\left(\sum_{V\in Valores(A)}\frac{|M_V|}{|M|}Error(M_V)\right)$$

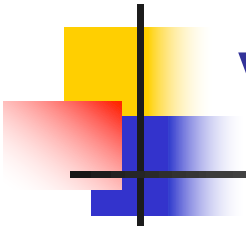
Other generalizations, variations and fine points

- The algorithms discussed so far look at one variable at a time. Other proposals look at more that one. There are no clear gains from these methods
  - What does this remind you of?
- Other measures to chose a node for categorical attributes in clasification trees
  - Gini impurity: is a measure of how many clases are represented in a node

$$I_G = \sum_{i \in Categorias} f_i(1 - f_i) = 1 - \sum_{i \in Categorias} f_i^2$$

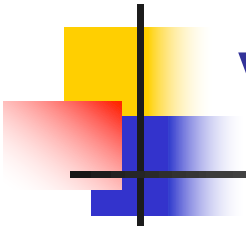
  - Where f_i is the proportion of data with category i represented in a node
  - We want to minimize or maximize impurity?

Other generalizations, variations and fine points



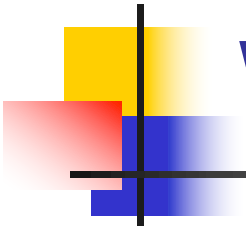
- The algorithms we discussed finish when all examples are classified or when we run out of attributes
  - This can lead to bad generlizatio, to having an overly complex model
- To mitigat this there are techniques to control the depth of the tree
  - Stop its growth at a given depth
  - Let it grow all the way and then trim it down using a validation set

Other generalizations, variations and fine points

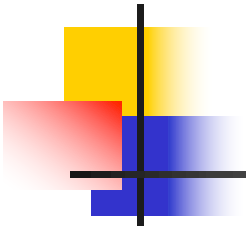


- Generalizations of Id3
  - Manages continous attributes
  - Uses other techniques for attribute selection
  - Mange missing values
  - Use of techinques to prevent over-fitting
- This results in the C4.5 and C5 (see5) algorithm designed by Quinlan

Decision Trees
When to use them



- Has to be supervised learning
- Useful when instances have some missing values
- Useful when we need an explanation for the model's output
  - Each path in the tree can be converted into an intelligible explanation



Exercise

- Get some data from the internet
- Use the decision tree in Sklearn
- Obtain the relevant performance metrics
- Plot the tree if possible