Regression Models

- Regression means the approximation of a real-valued function $f : D \times C \rightarrow [-1,+1]$ by means of a function $h : D \times C \rightarrow [-1,+1]$ given a set of points d_i and their corresponding $f(d_i)$, the training data
- Various TC systems used regression models
- We describe the Linear Least Squares Fit (LLSF) approach of [Yang&Chute94]

Linear Least Squares Fit

- Classification can be seen as the task of attributing to test documents d_j an output vector $O_j=\langle c_{1j},..,c_{mj}\rangle$, given its input vector $I_j=\langle w_{1j},..,w_{nj}\rangle$
- Hence, building a classifier comes down to computing a $n \times m$ matrix M such that $I_j M = Oj$

Linear Least Squares Fit

- LLSF computes M from the training data by computing a linear least-squares fit that minimizes the error on the training set according to the formula

$$M = \arg\min_M ||IM - O||_F$$

where

  - $||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{i=1}^{n} a_{ij}^2}$ is the Frobenius norm of the matrix A
  - I is the |Tr| $\times$ n matrix of input vectors of the training docs
  - O is the |Tr| $\times$ m matrix of output vectors of the training docs
- The M matrix can be built by performing a singular value decomposition
- LLSF has been shown in [Yang99, Yang&Liu99] to be one of the most effective classifier. Its drawback is the high computational cost involved in computing M

LLSF by SVD

- The theory: you want to find a vector v such that Av is as close as possible to a vector b, i.e. to minimize the euclidean norm of the residuals ||Av-b||
- The derivative of $(Av-b)^t (Av-b)$ is $2A^tAv-2A^tb$
- Then the solution is at $A^tA\ v = A^t\ v$, i.e. $v = (A^tA)^{-1}A^t\ b$
- Let $A = USV^t$ the singular value decomposition of A and S is a diagonal matrix
- Then the pseudoinverse $(A^tA)^{-1}A^t = VS^+U^t$ and S^+ is S where each non zero entry is substituted by its reciprocal

Neural Networks

- A Neural Network (NN) TC system is a network of units:
  - Input units represent terms appearing in the document
  - Output units represent categories to be assigned
  - Hidden units are detectors that 'discover' correlations among terms present in the input
  - Adjustable weights are associated to connections between units
- NN are trained by the backpropagation algorithm: the activation of each pattern is propagated through the network, and the error produced is back propagated and the parameter changed to reduce the error
- Non linear NN components (hidden and output units) provide no advantage
- The use of NN for TC is declined in recent years

Decision Tree Classifiers

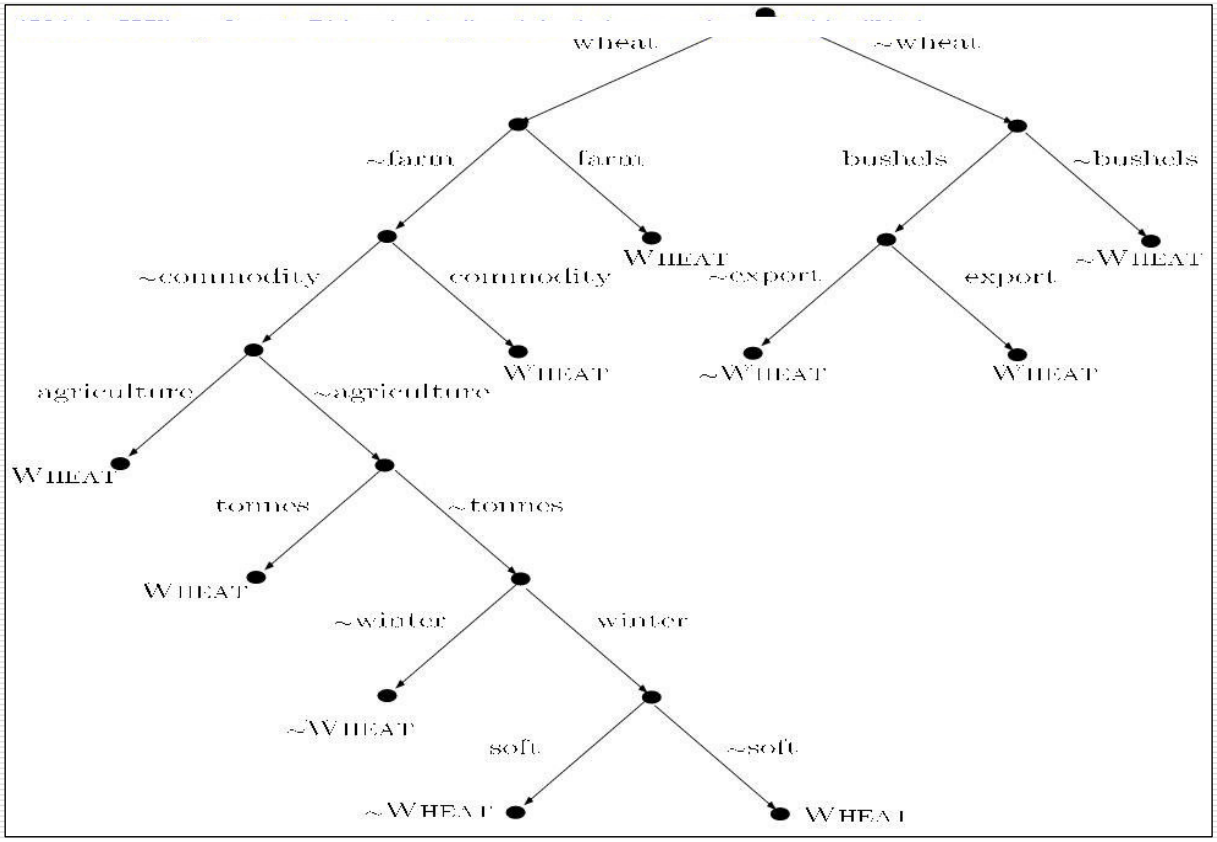
- A decision tree (DT) binary classifier for c_i consists of a tree in which
  - Internal nodes are labeled with terms
  - Branches departing fromm them correspond to the value that the term t_k has in the representation of test document d_j
  - Leaf nodes are labeled by categories
- The DT classifies d_j by recursively testing for the values of the term labeling, the internal nodes nodes have in representation of d_k=j
- In binary TC, leaves are labeled by either c_i, or $\neg c_i$ while in SL TC they are labeled by one among the classes in C
- When the term weights w_{kj} are in {0,1}, DT are binary trees; when they are in [0,1], DTs have variable ariety, and the branches are labeled by intervals $[i_s, i_{s+1})$, where each interval results from an entropy based quantization of the [0,1] interval

Learning a binary DT

- A general procedure to build a binary DT for a category c_i consists in a divide and conquer strategy:
  1. Assign all the docs in Tr to the root, and make it the current node
  2. If the docs in the current node have all the same value for c_i, stop and label the node with this value; otherwise select the term t_k which maximally 'discriminates' between c_i and $\neg c_i$
  3. Partition the set of docs in the current node into sets of docs that have the same value for t_k
  4. Create a child node for each such possible value and recursively repeat the process from Step 2
- In Step 2, selecting the term t_k that maximally discriminates among the classes (by e.g. information gain (C4.5) or Gini index (CART)) tends to maximize the homogeneity (in terms of attached labels) of the sets produced in Step 3, and hence to minimize the depth of the tree

Learning a binary DT

- In order to avoid overfitting, the growth of the tree may be interrupted before excessively specific branches are produced
- Nowadays DT text classifiers are usually outperformed by more sophisticated learning methods


wheat
~wheat
~farm
farm
bushels
~bushels
WHEAT
~commodity
commodity
~export
export
~WHEAT
agriculture
~agriculture
WHEAT
~WHEAT
WHEAT
WHEAT
tonnes
~tonnes
WHEAT
~winter
winter
~WHEAT
soft
~soft
~WHEAT
WHEAT