Machine Learning library in Haskell

Divyanshu Bhartiya (10250)

Deepak Pathak (10222)

Vineet Hingorani (Y9227654)

Rahul Ajmera (Y9227451)

Linear Regression

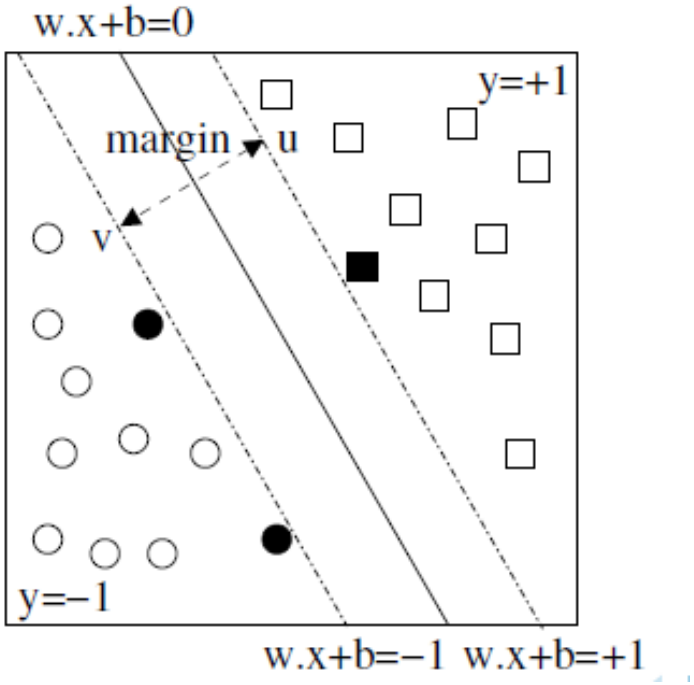
- Given a variable Y and p variables $X_1, \ldots, X_p$, we need to find the relationship between y and each X_j
- $Y = b_0 + b_1 X_1 + \ldots + b_p X_p$

- The coefficients are obtained using back substitution, which is a fast technique as compared to finding inverse.
- The implementation gave almost the same efficiency as compared to implementations in other languages(compared with MATLAB).

Logistic Regression

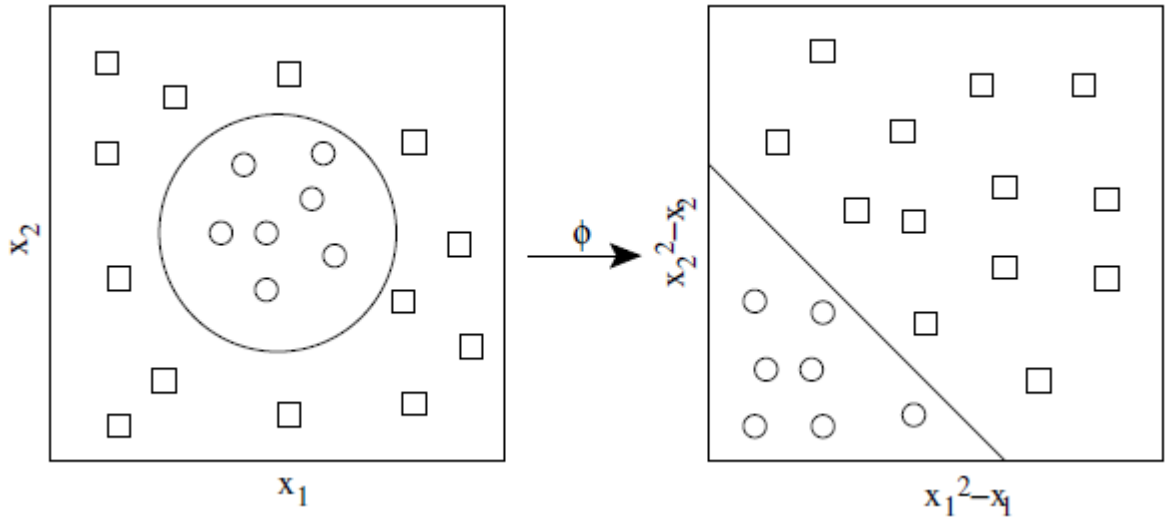
- Given a variable Y and dependencies with weights on class variables $C_1, \ldots, C_p$, we need to predict the dependent quotient on class variables.
- Based on recursive implementation, followed by iterations to reach a steady state.
- The implementation was efficient as recursion almost take same order of time in any case(compared with MATLAB).

Support Vector Machine (SVM)



- Supervised learning model
- Binary non-probabilistic classifier
- Maximal margin classifier
- Non-linear classification , mapping inputs to higher dimensional space (kernel trick)
- Kernel function to find similarity between two classifiers

SVM Function Non-Linear classification

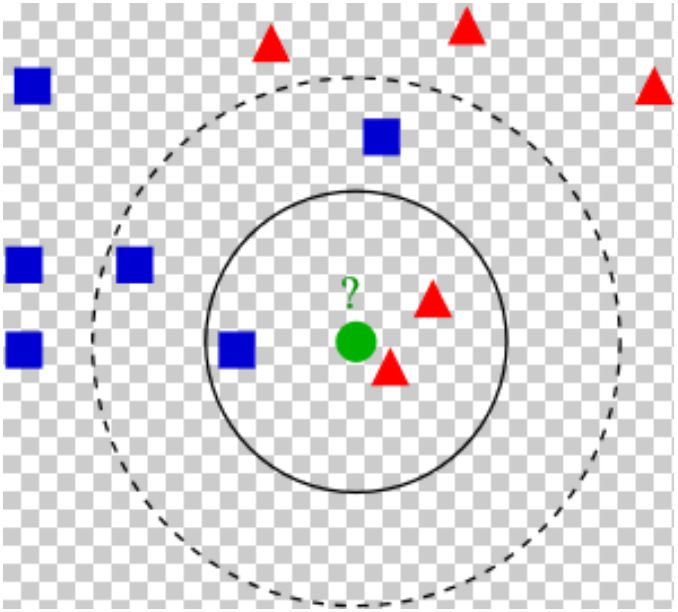


- Non –linearly separable data but in (x1^2 - x1; x2^2 - x2) space, data becomes linearly separable
- Implemented Iterative Conjugate Gradient Algorithm to solve Ax=b (since kernel matrix will be symmetric-positive)

HMM

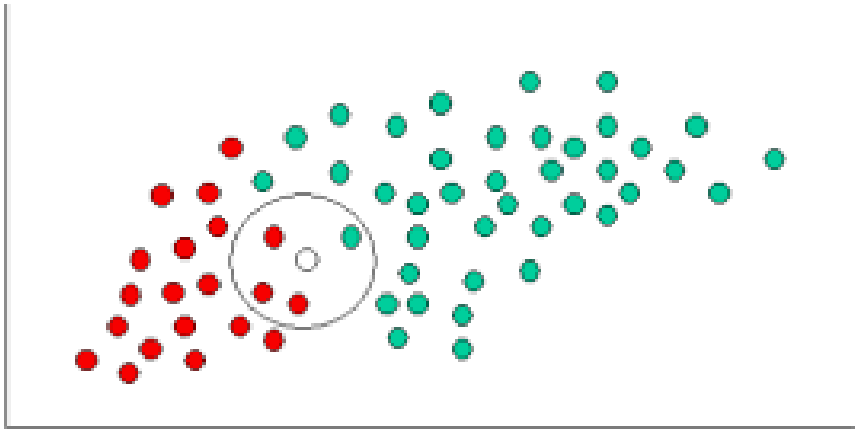
- Used Memoization, Recursive function also implemented
- Forward Algorithm:
  - Given n observation(O) and an HMM, outputs the probability of occurrance of state S_n
- Backward Algorithm:
  - Given nth state and an HMM computes the probability of occurrence of k observations after it
- Viterbi Algo:
  - Given a sequence of observation, computes the most likely sequence of states that resulted in these observations
- Baum Welch:
  - Given an observation sequence, it learns the HMM parameters using EM algo

KNN Classifier



- K-Nearest Neighbor classification
- a method for classifying objects based on closest training examples in the feature space
- the function is only approximated locally
- distance metric used is “Euclidean distance”
- number of neighbors can be altered

Naive Bayes Classifier

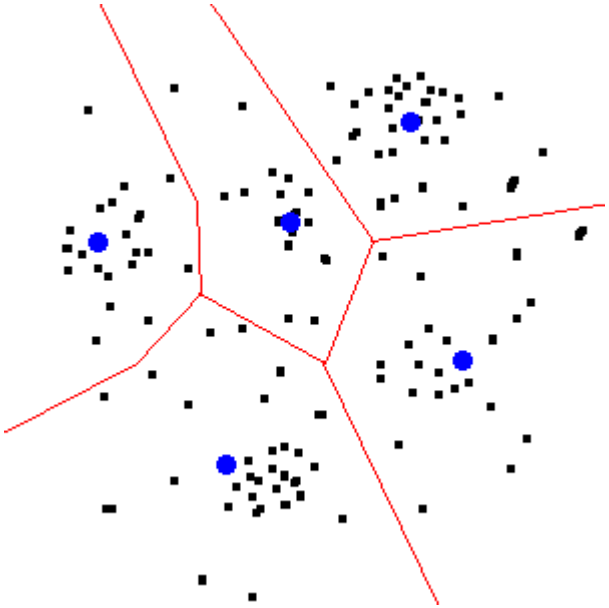


- Probabilistic classifier based on Bayes' theorem
- Independent feature model
- prob of a category * prob of feature to be in category
- maximum prob for a category gives the resultant category
- Feature space can be multi-dimensional
- Distribution can be guassian/poisson

Latent Semantic Analysis

- Technique of Natural Language Procesing
- Vectorial semantics and dimensionality reduction
- Analyzes relationships betweens terms and documents
- Create a term by document matrix with tfidf scores
- SVD decomposition of matrix
- Reduce it to ‘n’ topics
- Get respective vector of terms and documents

K-Means clustering



- Distance metric is “Euclidean Distance”
- Divides observations into k-sets
- Centroids are used to identify cluster
- Associate all points with nearest centroid out of k points

Packages used :

- Data.CSV
- Hmatrix
- Blas
- Lapack
- Gsl
- Log-float
- MemoCombinators

THANK YOU