IST 707 - Data Analytics - Project Proposal

Drug Recommendation System: Classification of drug reviews into patient's conditions and rating prediction to recommend drugs

Group: Akshita Chandiramani, Ruchita Jadhav, Sanjeev Ramasamy Seenivasagamani

Overview:

The Drug Review Dataset from the UCI Machine Learning Repository provides patient reviews on specific drugs along with related conditions and a 10 star patient rating reflecting overall patient satisfaction. The data was obtained by crawling online pharmaceutical review sites. Our aim is to automatically classify drug reviews into patient's condition and predict drug ratings to recommend drugs to patients.

Goals:

Goal 1: Exploratory Data Analysis and Pre-processing

Tasks:
1) Perform descriptive analysis to answer questions about drug review data
2) Perform information retrieval and text mining (through tf-idf) to extract meaningful information from drug reviews

Expected results:
1) Understanding on what kind of drugs are there, what sorts of conditions do these patients have and how many drugs are present for each condition along with seasonality of drugs
2) Vectorized drug reviews based on term frequency-inverse document frequency

Goal 2: Predict patient's condition based on the review

Tasks: Apply classification algorithms on vectorized drug reviews to predict a patient's condition
Expected results: Accurate classification model that classifies drug reviews to match with a patient's medical condition
Algorithms to be used for Classification:
kNN - It is simple to implement, robust to noisy training data and effective if training data is large. Our data contains 884 distinct classes which will be used as k in the classification
SVM - It is effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient
Random Forest - It can handle large data sets with high dimensionality and handle missing data while maintaining accuracy. It reduces overfitting and is more accurate than decision trees

Goal 3: Sentiment Analysis of drug review to predict drug rating

Tasks:
1. Perform sentiment analysis on vectorized data to calculate sentiment rating of review and understand other aspects such as effectiveness and side effects
2. Apply linear regression to predict drug rating from sentiment analysis

Expected results:

1. Calculated sentiment ratings of various drugs based on drug reviews
2. Accurate regression model that predicts drug rating based on patient’s drug review
Algorithms to be used for Regression:
Linear regression: The prediction variable is continuous which requires a linear regression

Goal 4: Cluster Analysis on drugs

Tasks: Perform cluster analysis to find underlying grouping of drugs
Expected results: Distinct clusters of drugs based on side effects and effectiveness
Algorithms to be used for Clustering:
K-means: Easy to implement, computationally fast if k is small and produces tighter clusters
HAC: It outputs a structure that is informative which makes it easier to decide the number of clusters

Goal 5: Application Development

Tasks:
Develop an interactive R-shiny application to implement end-to-end analytical process of project
Expected results:
R-shiny application that performs the following:
1. Recommends drugs based on patient’s medical condition
2. Patient enters drug reviews for their condition

Dataset: The Drug Review Dataset from the UCI Machine Learning Repository
Tools: R-Studio, Rmarkdown, R-shiny
Algorithms:
Classification: SVM, Random Forest, kNN Algorithm
Clustering: k-Meanslgorithm a, HAC
Regression: Linear
Packages:
Data Wrangling and Data Munging: dplyr, tidyr
Data Visualization: ggplot2
Machine Learning: caret, arules, arulesViz, mlbench, rpart, C50, rattle, ElemStatLearn, klaR
Model Evaluation: pROC