ML1010 - APPLIED MACHINE LEARNING

Project Proposal - IMDB Movie Sentiment

Prepared for: Annie Lee, PHD, MMATH
Prepared by: Jairo Melo, Ing. MSC
March 27, 2019
Proposal number: 001

Motivation

Sentiment analysis is a well-known method in the work of natural language processing. The objective is to determine the polarity of a text captured either by a survey, observations, descriptions of events, and opinions or reviews.

The IMDB dataset provides a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided.
http://ai.stanford.edu/~amaas/data/sentiment/

Project Plan

Through CRISP-DM, I will provide the steps to complete the Sentiment analysis for the movie review dataset.
Data understanding:
Data preparation
Modeling
Evaluation
Deployment

Machine Learning Approach

I will need to convert each review to a numeric representation, also knows vectorization. After, I'm planning to use Logistic Regression to build a classifier as it's easy to interpret and performance is good.
As an alternative, I could use VADER package to extract the sentiment of each document. VADER, (Valence Aware Dictionary and sEntiment Reasoning), is a lexicon and rule-based tool that is specifically tuned to social media which could fit this dataset quite well.