CS 229 Machine Learning Project Proposal

Topic:

Predicting which recommended content users click

Team Member:
Stanley Jacob
Lingjie Kong

Background

Several internet and news companies struggle with providing users with relevant advertisements on their websites. One such company is Outbrain, which provides advertisements that are usually related news articles located below the article the user is currently reading. Users generally have some motivation for reading the current article on new website like CNN, and they are more likely to visit other websites that provide them with relevant information. However, companies like Outbrain often have scarce information about the user and therefore struggle to determine optimal related articles to show for the current user. Incorrect predictions of user's desires result in irrelevant recommended sites that users most likely would not click. Thus, Outbrain needs to find a way to harness data from users and the advertisement.

Goal

The goal of our project is to select features related to users and advertisements and build a model that classifies whether a user will click a given ad. An underlying objective is to select features that can be used reliably for prediction. A long-term goal would be to provide users with more relevant advertisements that may also match their individual preferences.

Data

Outbrain provide information about users and advertisements on websites in a large dataset available on Kaggle. The dataset contains information related to 2 billion page views and about 17 million clicks done by 700 million users, so part of the challenge of our project will be dealing with this large dataset. Two examples of features they provide are user location and advertiser's id. They also provide other textual features related to the content of the advertisement and of the webpage itself.

Methodology

We want to explore different learning algorithms to accurately predict whether a user will click an advertisement. We will start by using logistic regression. Basically, we will model the loss function given the (x,y) pairs with feature extractors to minimize the total cost function. Then, we will try boosting through different weak learners. In addition, given the large number of features, we will try using support vector machines (SVMs). Lastly, we would like to implement a neural network.