Machine Learning Engineer Nanodegree

Capstone Proposal

Rakan Alkheliwi

May 25, 2018

Proposal

Domain Background

I work as an IT professional who receives email complaints and tickets from customers. I belong to a team with a different but similar set of skills varying depending on expertise.

A common problem in IT is having tickets bouncing from one department to another. Also, we face an issue with different segments of consumers; I can hypothesize that there are difficult consumers to deal with and those who are regular. In addition, we measure KPIs based on time optimization and we don't take into effect developing new employees in new fields.

Not much research has been conducted in this area, but for text classification a lot of papers has been done in this field. Using ngrams, bag-of-words, and ConvNets we can classify texts to see if ticket belongs to a certain department.

We have done a project where we segmented customers and I am planning on developing a similar strategy for customers and employees.

Problem Statement

In this proposal, I am trying to solve following problems toward the ultimate problem:

- Automated way of assigning tickets to responsible department

- Segmenting consumers based on their level of difficulty and grumpiness

- Segmenting employees based on knowledge and expertise

- Automated way to assign a ticket to some employee based on an intelligent formula or agent

This will end up with an agent that can replicate what a helpdesk can do.

Datasets and Inputs

In this project, I will be also applying what I learned in data analyst nanodegree. I will be analyzing following data sets:

- Newsgroup text from UCI

http://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

The data contains text data from twenty newsgroups where each folder contain same number of files. The purpose of analyzing this data is to build a machine that can analyze text and assign an email to the right department. Assume we are in the business of reviewing content produced by a random email to see where it belongs. The idea is to later build it in real data but due to limitations of using my company's data I'm not able to do so.

- IT Tickets from IBM Watson Analytics

https://www.ibm.com/communities/analytics/watson-analytics-blog/it-help-desk/

The data contains 100,000 tickets with employee who handled the ticket which we don't care about as we are building a machine to imitate him or do better. It contains resolution time, employee who solved problem, employee or customer who sent complaint, and customer satisfaction.

Further, I will analyze IT tickets dataset and produce from it another data set with average resolution time and count of satisfaction and use it to segment customers.

I will be also analyzing IT tickets dataset to come up with a dataset that shows average resolution time and satisfaction for employee who solved ticket and his expertise for different domains.

Solution Statement

We will need a set of models:

- One that will analyze text and assign it to correct newsgroup.

- One that segments customers

- One that segments employees

- A model that assign tickets to employees based on results of above models.

Benchmark Model

My system need to do better than what is shown in the IT tickets model and better than a simple round robin model. A round robin model will have employees clustered into groups but give the ticket in a greedy manner; give it to the most expert guy in the group, if he is busy give it to the next one with total disregard to developing novice employees.

Evaluation Metrics

The metrics are resolution time, satisfaction, and cost. I want to minimize resolution and have satisfied employees, but I also want to encourage knowledge sharing and development of new

employees. If I keep giving tickets to experts, these experts will retire and I will have non-developed employees.

The way to calculate resolution time and satisfaction will be based on a probability function based on the distribution of tickets resolution for each employee. Cost will be determined based on the cluster the employee belongs to.

Project Design

Build data sets

Here I will build a data set based on IBM Watson data where I get average resolution time and satisfaction count for each customer. Also, I will build a profile for each employee who handled tickets based on his average resolution and satisfaction for each app or domain.

Unsupervised Learning: K-Means

I want to segment my consumers based on rigid groups. I will be using Silhouette coefficient to determine best K for the model.

Plug segment into original IT IBM data

Unsupervised Learning: Gaussian clustering

Employees can't be identified rigidly into groups, instead they have a certain percentage of knowledge in each cluster. So I need a soft algorithm to distinguish employees.

Plug data into original data set

Use deep learning to analyze text

I will be using deep learning to analyze trends in data set where I will split data into training and testing segments. I will be using ConvNets, ngrams, and bag of words as ways to analyze text and come with best model.

Use Reinforcement *learning to build agent*

I will be using this algorithm to derive a formula and a model that performs better than what is shown in IT department. The reason I am using reinforcement learning it that I want to punish the model for excessively using experts and putting new employees aside. Your consultation is appreciated for this.

Build Monte Carlo Simulation

A model that will get emails at a particular rate, assign ticket to department and assign customer to segment, and then assign ticket directly to a person in the team.