CS 5007 – Group Project Abstract

Fall 2019

(1) Motion Planning For 2–DOF Planar Manipulators

The purpose of this project is to showcase and benchmark different robotic motion planning algorithms with a variety of different obstacles and states, all created by the user. The many varied approaches and options available when choosing a motion planner necessitate a convenient, intuitive method of comparison and benchmarking in order to select the most appropriate one. To that end, this project aims to provide an intuitive, graphical view of motion planners and their capabilities in a simplified, 2-dimensional case. Due to the large variety of motion planners and solvers, a highly modular approach has been taken. Ease of expansion and incorporation of additional elements is a must, and thus the project relies heavily on object-oriented principles including inheritance and object composition. By splitting motion planning into a set of independent, simple objects, complex behavior was achieved through their composition. In that regard, what has been achieved so far is a strong foundation and a large set of key features implemented. Most of the framework has been set and tested, and even 2 of the 4 proposed motion planners were implemented, showcasing the ease with which more can be added due to the present framework.

(2) A Survey of Methods Used to Detect Anomalies in Online Banking Data

Using three different methods, we attempt to determine which is the most accurate at detecting outliers in time-series data from a Bolivian bank. We utilize support vector regression, general filtering and statistical detection methods, and a k-means program as our algorithms. Once we successfully acquired models which can detect outliers, we will determine their efficacy. Successful outlier detection is the first step in identifying factors which can predict future anomalies before they occur. We have been able to implement support vector regression with a number of different kernels on a subset of the data. We have also been able to utilize a Kalman filter in order to identify anomalies. Additionally, we implemented a custom written naive k-means algorithm and implemented additional steps for anomaly detection through careful data manipulation and model analysis.

(3) Hate Speech Detection on Twitter

The project aims to solve a data science problem of hate speech detection for the social media platform of Twitter. We are building a hate speech detection algorithm that given a new tweet will classify it as hate tweet or a regular tweet. To train the algorithm, we are using the labelled twitter data from the Analytics Vidhya data science challenge. The project consists of 3 main phases. Data Cleaning, Vectorization & Classification. For data cleaning, python libraries like NLTK, Spacy are used. The vectorization of the tweets is done by the Bag of Words model, TF-IDF & Word2Vec. For the classification, Naïve Bayes, Support Vector Machine and Random Forest algorithms were used. We finished 12 Iterations & we obtained a steep improvement in the accuracy from 74.9% to 96%. We crossed our target of 85% accuracy.

(4) Prediction of Manufacturer's Suggested Retail Price (MSRP) of Cars

The objective of the project is to predict manufacturer's suggested retail price of cars using statistical and machine learning models such as multiple linear regression, regularization regression, gradient boosting machine, and random forest. The raw dataset contains 11,914 observations and 16variables about the cars including MSRP. First, we looked at the dataset to understand the types of features after that we detected missing values and replaced them with reasonable values. Second, we performed data preprocessing: visualizing correlations between variables, one-hot encoding, and normalizing data. Then we trained the models and checked that the models are not overfitting by validation set. Mean absolute error reflects model performance on the test set. Finally, we improved the predictions by ensemble learning techniques and found that they increased accuracy.

(5) Cave of Pythons

The purpose of this project is to create "Cave of Pythons," a text-based adventure game that challenges the user to find the 'treasure'. The player has to accomplish certain objectives/puzzles in order to reach the 'treasure' and win the game. This style of game is inspired by text-based adventure games like "Zork". The purpose of this project is to use the skills and programming tools we have learned in class to create an entertaining and engaging experience for players. We are creating this game using PYCharmIDE, the Python library Tkinter, and the Python library PIL. For planning, we are creating mock-ups for the game map and the GUI that the player will be using. We are choosing the number and purpose of the different classes and objects we are using. Based on these planning considerations we are developing pseudo code to help us write our code. The code comprises of four main sections. These sections are Define all Command Functions, Initialize State of Objects, Update Variables/Objects, and Create the GUI. We are also creating four classes, NPC, Room, Player, and Item, to complete our project using object-oriented programming. The game is now in a working state and ready for users to enjoy.

(6) CNNs on Prediction of Clinical Image

The Convolutional Neural Network (CNN) is a class of deep learning neural network, which is applied to analyze image information. Melanoma, also known as malignant melanoma, is a type of cancer that develops from the pigment-containing cells known as melanocytes. Deep convolutional neural network (CNN) have been successfully applied to detect cancer, diabetic retinopathy, and dermatologic lesions from images. This study built a self-design model to distinguish the differences between two labels in training set.

(7) Predictive Analysis using Machine Learning on Drug Review Dataset

In this project, we planned to come up with a predictive model of the drugs reviews that could be used by pharmaceutical companies during manufacturing and by medical practitioners and patients for treating several illness conditions, based on the drug ratings and reviews. We came up with the following strategies to achieve this: (1) Data Visualization (2) Sentiment Analysis (3) Predicting patient's condition based on the reviews (4) Predicting drug rating based on reviews & (5) Optimization. We have performed actions like data cleaning and preprocessing followed by some visualization using D3 & React, Vega Lite API and Python. We have also performed sentiment analysis using Spacy and NLTK and model prediction comparison using machine learning techniques like Linear SVC, Logistic Regression, SGD and so on. The results show that "Birth Control" and "Depression" are the most rated and reviewed illness condition and Logistic Regression gives us more accuracy among all the models. The visualizations also show that the drug ratings and reviews have gone up during the 10-year period, we are trying to assess. The accuracy of drug ratings (range 1-10) predicted by review has reach 79% by Logistic Regression, but polarized rating has higher accuracy with 85% by Decision Tree.

Fall 2018

(1) Self-Righting Writing Utensil

The team sought to develop a prototype for a motion correcting writing utensil. Throughout the course of this project, the team developed the control algorithm for a one-dimensional reaction wheel inverted pendulum as well as the appropriate hardware-software interface to allow the device to communicate with the operator computer. An underlying key goal of this project was to is to learn and practice the fundamentals of hardware-software interfacing. The team composed and implemented the required functions for the reaction wheel to compensate for tilt; however, because the motor was damaged, the results are from simulation instead of prototype testing and experimentation.

(2) House Price Exploration and Modeling

We are working on an exploratory data analysis of house price data because we want to have a better understanding of data and construct features for further predictive modelling. This house price dataset contains 79 explanatory variables describing almost every aspect of residential homes in Ames, Iowa. First, we explore the shape and size of all dataset. Then we explore the correlation between explanatory variables and house price. After this, we further explore how the relationship of each variable against sale price looks like. We analyze numerical variables by scatterplots and categorical variables by boxplots. We have already explored the size and shape of this dataset, and analyzed the response variable, 'SalePrice'. Also, we have checked variables with large amount of missing value and had ideas about how to deal with them. Furthermore, we have categorized all the remaining variables into groups and plotted each variable against sale price.

(3) Predicting The Winning Team By Drafting Hero Using Statistical Model

DotA 2 or Defend of the Ancient 2 is a free-to-play MOBA game which consists of two teams, Team 1 and Team 2. Each team has five players and each player must select different playable characters. There are 113 playable characters and ten of them will be chosen to play. As technology is growing rapidly, Esport such as DotA 2, Overwatch, League of Legend, etc. becomes more popular and attach more people to play and watch. Our group consider the data from the machine learning repository of University of California Irvine to predict the probability of winning team by the composition of picking playable characters. We use 10-fold cross-validation to select the best model of our selected methods, e.g., logistic regression with feature selection, decision tree, random forest, and artificial neural network by calculating the accuracy of AUC. Logistic regression gives the best accuracy but we think the model is overfitting. However, the logistic regression with feature selection still give the best accuracy along with the model from neural network.

(4) Data Mining Competition On Kaggle: Digit Recognizer

The main purpose of our project is to recognize a handwritten number in an image. We try several methods, such as PCA (Principal Components Analysis) and CNN (Convolutional Neural Networks), and figure out the most accurate one. Now after trying PCA and CNN, we get the results: the accuracy of PCA is 0.9409632314862766 and the accuracy of CNN is 0.9923809523809524.

(5) Sensor Placement For Optimal Control In Parabolic PDE System

The parabolic Partial differential equations (PDEs) are often used to describe plenty of phenomena, such as thermal diffusion and wave motion [1][2]. They are also widely applied in many other fields such as financial mathematics, the famous Black-Scholes model represented by PDE governs dynamics of a financial market containing derivative investment instruments. The overall purpose of our project is to design a tool to place sensors in systems described by parabolic partial differential equations by exploiting existing sensor placement algorithms. In this project, our code is developed based on python together with libraries such as numpy, pandas dataframe and matplotlib. In the past months, we have figured out the mathematical mechanism behind CVT algorithm, greedy algorithm and

uniformly distribute algorithm. We also realized the algorithms via coding with Python. All the work listed in the proposal is completed.

(6) Visualization Of Domestic Flight Paths

The overall purpose of this project is to provide an interactive data visualization of air traffic along domestic flight paths and the populations of cities that they connect. Data pulled from TranStats, a product of the U.S. Department of Transportation, is used to inform the visualization. Panda Dataframes will be the primary method of delivering data to components of the program. The Matplotlib BaseMap toolkit will be used to project location data onto a map. This map will be displayed in a Tkinter based graphical user interface (GUI). The major milestone of this project is the integration of multiple packages to the development of an application built based on data. This framework can be further refined to produce visualizations specific to the needs of the intended audience. In addition to this milestone, results observations of the issues experienced when developing the backend of a platform-independent application using multiple operating systems.

(7) Data Visualization APP

This project aims to create a simple, straightforward and usable tool for non-programming users to create interactive diagrams for their data visualization tasks. In order to fulfill this goal, we decided to design and implement a desktop application called "DVAPP". This app can load data from csv file and visualize it with well-designed and interactive diagrams. With simple GUI, users will only need to click for a few times and get their wanted diagrams. In this final report, we will demonstrate our overall work for designing, implementing and testing DVAPP and also the details with design methods, implementations and validation of our work supported by screenshots. It can be said that we successfully fulfill the designed scope for the APP prototype. Right now, the APP already can function well and generate expected diagrams. In the future we will manage to add more features and supported diagram type in order to increase the functionality and usability.

(8) Breast Cancer Prediction Based On Logistic Regression Model, Decision Tree Model, Support Vector Machine Model And Adaboost Model

At present, the problem of breast cancer in the world is getting more and more serious, so it is very important to define whether breast tumor is benign or malignant. We plan to use logistic regression model, decision tree model, support vector machine model, adaboost model and dataset called Breast Cancer Wisconsin (Diagnostic) Data set obtained from Kaggle to implement the suitable statistical graphs to demonstrate the dataset more intuitionistic, namely, the data visualization of the dataset. We hope that through our analysis of the dataset, we can find a defined value for all aspects of benign tumors and malignant tumors, so that when patients provide their disease data, we can judge the severity of the disease. Once the prediction correction rate could reach to a comparable high level, it could be considered our model have a high effective rate for predicting the diagnosis of the tumor of breast cancer.