greatlearning
Learning for Life

AIML Project List

A comprehensive collection of the top projects done by our learners from various domains.

TEXAS
The University of Texas at Austin

GREAT LAKES
EXECUTIVE LEARNING

Graded Projects

1. Analyze health information to make decisions for insurance business

Context

This project uses Hypothesis Testing and Visualization to leverage customer's health information like smoking habits, BMI, age, and gender for checking statistical evidence to make valuable decisions of insurance business like charges for health insurance.

Dataset

The data at hand contains the medical costs of people characterized by certain attributes.

Domain

Healthcare

Tools & Concepts

Hypothesis Testing, Data Visualization, Statistical Inference, Python

2. Identifying potential customers for loans

Context

This case is about a bank (Thera Bank) whose management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with a minimal budget.

Dataset

The file contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Domain

Banking

Tools & Concepts

Logistic Regression, KNN, Classification, Python

3. Diagnosing Parkinson's disease using Random Forests

Context

Parkinson's Disease (PD) is a degenerative neurological disorder marked by decreased dopamine levels in the brain. It manifests itself through a deterioration of movement, including the presence of tremors and stiffness. There is commonly a marked effect on speech, including dysarthria (difficulty articulating sounds), hypophonia (lowered volume), and monotone (reduced pitch range). Additionally, cognitive impairments and changes in mood can occur, and the risk of dementia is increased.

Traditional diagnosis of Parkinson's Disease involves a clinician taking a neurological history of the patient and observing motor skills in various situations. Since there is no definitive laboratory test to diagnose PD, diagnosis is often difficult, particularly in the early stages when motor effects are not yet severe. Monitoring the progression of the disease over time requires repeated clinic visits by the patient. An effective screening process, particularly one that doesn't require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and non-invasive tool for diagnosis. If machine-learning algorithms could be applied to a voice recording dataset to accurately diagnosis PD, this would be an effective screening step prior to an appointment with a clinician.

Dataset

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to the "status" column which is set to 0 for healthy and 1 for PD.

Domain

Medicine

Tools & Concepts

EDA, Logistic regression, Decision Trees, Python

4. Classifying silhouettes of vehicles

Context

The purpose is to classify a given silhouette as one of three types of vehicles, using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different angles.

Dataset

The data contains features extracted from the silhouette of vehicles from different angles. Four “Corgie” model vehicles were used for the experiment: a double-decker bus, Chevrolet van, Saab 9000, and an Opel Manta 400 cars. This particular combination of vehicles was chosen with the expectation that the bus, van, and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.

Domain

Automobile

Tools & Concepts

Support Vector Machines, Principal Component Analysis, Classification, Python

5. Predicting the Strength of high-performance concrete

Context

Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate.

Dataset

The actual concrete compressive strength (MPa) for a given mixture under a specific age (days) was determined from the laboratory. Data is in raw form (not scaled). The data has 8 quantitative input variables, and 1 quantitative output variable, and 1030 instances (observations).

Domain

Cement manufacturing

Tools & Concepts

Regression, Decision Trees, Feature Engineering, Python

6. Product Recommendation System

Context

Online E-commerce websites like Amazon, Flipkart uses different recommendation models to provide different suggestions to different users. Amazon currently uses item-to-item collaborative filtering, which scales to massive data sets and produces high-quality recommendations in real-time.

Dataset

Amazon Reviews data (data source) The repository has several datasets. For this case study, we are using the Electronics dataset.

Domain

E-commerce

Tools & Concepts

Collaborative Filtering, Recommender Systems, Python

7. Implementing an Image classification neural network to classify Street House View Numbers

Context

Recognizing multi-digit numbers in photographs captured at street level is an important component of modern-day map making. A classic example of a corpus of such street-level photographs is Google's Street View imagery composed of hundreds of millions of geo-located 360-degree panoramic images. The ability to automatically transcribe an address number from a geolocated patch of pixels and associate the transcribed number with a known street address helps pinpoint, with a high degree of accuracy, the location of the building it represents.

More broadly, recognizing numbers in photographs is a problem of interest to the optical character recognition community. While OCR on constrained domains like document processing is well studied, arbitrary multi-character text recognition in photographs is still highly challenging. This difficulty arises due to the wide variability in the visual appearance of text in the wild on account of a large range of fonts, colors, styles, orientations, and character arrangements. The recognition problem is further complicated by environmental factors such as lighting, shadows, secularities, and occlusions as well as by image acquisition factors such as resolution, motion, and focus blur.

Dataset

In this project, we will use a dataset with images centered around a single digit (many of the images do contain some distractors at the sides). Although we are taking a sample of the data which is simpler, it is more complex than MNIST because of the distractors
SVHN is a real-world image dataset for developing machine learning and object recognition algorithms with the minimal requirement on data formatting but comes from a significantly harder, unsolved, real-world problem (recognizing digits and numbers in natural scene images). SVHN is obtained from house numbers in Google Street View images.

Domain

Object recognition

Tools & Concepts

Neural Networks, Deep Learning, TensorFlow, Image Recognition

8. Face mask segmentation

Context

Predict and apply masks over the faces within images using CNN and image recognition algorithms. In this hands-on project, the goal is to build a system, which includes building a face detector to locate the position of a face in an image and apply a segmentation mask on the face.

Dataset

WIDER FACE dataset is a face mask segmentation benchmark dataset, of which images are selected from the publicly available WIDER dataset. This data have 32,203 images and 393,703 faces are labeled with a high degree of variability in scale, pose, and occlusion as depicted in the sample images. In this project, we are using 409 images, and around 1000 faces for ease of computation.

We will be using transfer learning on an already trained model to build our segmenter. We will perform transfer learning on the MobileNet model which is already trained to perform image segmentation. We will need to train the last 6-7 layers and freeze the remaining layers to train the model for face mask segmentation. To be able to train the MobileNet model for face mask segmentation, we will be using the WIDER FACE dataset for various images with a single face and multiple faces. The output of the model is the face mask segmented data which masks the face in an image. We learn to build a face mask segmentation model using Tensorflow.

Domain
Object detection

Tools & Concepts
Computer Vision, CNN, Transfer Learning, Object detection, Segmentation, TensorFlow

9. Face mask segmentation

Context
Recognize, identify, and classify faces within images using CNN and image recognition algorithms. In this hands-on project, the goal is to build a face recognition system, which includes building a face detector to locate the position of a face in an image and a face identification model to recognize whose face it is by matching it to the existing database of faces.

Dataset
Aligned Face Dataset. With images taken from Pinterest, this dataset includes over 10,000 images of 100 different celebrities. There is an average of 100 images included of each celebrity.

Domain
Computer Vision

Tools & Concepts
Computer Vision, TensorFlow, CNN, Siamese Networks, Triplet loss, TensorFlow

10. Sentiment Analysis using LSTM

Context
Word embedding is a type of word representation that allows words with similar meaning to have a similar representation. It is a distributed representation for the text that is perhaps one of the key breakthroughs for the impressive performance of deep learning methods on challenging natural language processing problems. We will use the IMDb dataset to learn word embedding as we train our dataset. This dataset contains 25,000 movie reviews from IMDB, labeled with a sentiment (positive or negative).

The objective of this project is to build a text classification model that analyses the customer's sentiments based on their reviews in the IMDB database. The model uses a complex deep learning model to build an embedding layer followed by a classification algorithm to analyze the sentiment of the customers.

Dataset

The Dataset of 50,000 movie reviews from IMDB, labeled by sentiment (positive/negative). Reviews have been preprocessed, and each review is encoded as a sequence of word indexes (integers).

Domain

Entertainment

Tools & Concepts

RNN, Word embedding, LSTM, Classification, TensorFlow

11. Sarcasm Detection

Context

Past studies in Sarcasm Detection mostly make use of Twitter datasets collected using hashtag based supervision but such datasets are noisy in terms of labels and language. Furthermore, many tweets are replies to other tweets and detecting sarcasm in these requires the availability of contextual tweets. In this hands-on project, the goal is to build a model to detect whether a sentence is sarcastic or not, using Bidirectional LSTMs.

Dataset

News Headlines dataset for Sarcasm Detection. The dataset is collected from two news websites, theonion.com and huffingtonpost.com. This new dataset has the following advantages over the existing Twitter datasets: Since news headlines are written by professionals in a formal manner, there are no spelling mistakes and informal usage. This reduces the sparsity and also increases the chance of finding pre-trained embedding.

Domain

News

Tools & Concepts

LSTM, Classification, GloVe, TensorFlow

Non-Graded Project

12. Data Analysis with Python

Context

The GroupLens Research Project is a research group in the Department of Computer Science and Engineering at the University of Minnesota. The data is widely used for collaborative filtering and other filtering solutions. However, we will be using this data to act as a means to demonstrate our skill in using Python to “play” with data.

Dataset

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set.

Domain

Entertainment

Tools & Concepts

Python, EDA, pandas, matplotlib

Capstone Project

1. Deep learning (CV/NLP) - Pneumonia Detection & Automatic Ticket Classification

Context

The capstone project is a focused approach to attempt a real-life challenge with the learnings from the program. The AIML capstone problems are classified under the themes of Computer Vision and Natural Language Processing. The Goals of the Projects achieved are tagged here.

- Computer Vision: Pneumonia Detection - Locate the position of inflammation in an image.
- Natural Language Processing: Automatic Ticket Allocation - Build a classifier that can classify the tickets by analysing text

Dataset

CV - Images of chest radiographs. NLP - Support ticket data.

Domain
Healthcare, IT Support

Tools & Concepts
CNN, Text Preprocessing, LSTM, Computer Vision, NLP, TensorFlow

2. Machine learning - Prediction of the house prices

Context
This project involves using various features variables available in the Innercity house price dataset. Different Machine learning models like Regression, Ensemble techniques were used to analyse and predict the house prices. Also, the grid search algorithm was used to tune different parameters associated with models.

Dataset
Dataset having features about the house.

Domain
Real estate

Tools & Concepts
Feature Selection, Model Selection, Regression, Cross Validation, Python