Basharat Mubashir Ahmed

Machine Learning Assignment 1

312093

Tutorial Group 2

Task 1

1. Inferring whether a text-based review of a product is negative, neutral or positive demands that the algorithm needs to classify a review given the three discrete categories. Hence, this is an example of a classification problem.

2. When the computer network detects unusual traffic that indicates a possibility of attack against the network, this is the case of a behavior that gives rise to suspicions and hence is an example of anomaly detection or outlier detection.

3. Given the features of the house which describe the property and its location, we need to predict the selling price of the house which in this case is a continuous variable. Hence, this problem is an example of a regression problem

4. Learning to set the bitrate for a video stream viewed by a user at any point in time based on the given information current connection quality, state of video buffer and user location falls into the category of regression where we need to predict a continuous value based on certain input feature.

5. Here, we have two discrete choices, whether the advertisement will be clicked on or not given the information about user, ad and web page. Hence, this is an example of a classification problem where the output is binary either yes or no.

6. Learning to generate photorealistic images of human faces is a problem of generative modeling as here we are to generate new instances of human faces given a particular distribution.

7. We are to detect groups of messages that might possibly be spam messages. This is an example of clustering wherein we look for groups where objects have similar properties, in this case, emails.

8. Learning to play a real-time strategy computer game such as StarCraft is an example of reinforcement learning. Here, the optimal strategy for the game will be deduced by a sequence of decisions so as to yield a reward which is the optimal strategy to win the game.

Task 2

Task	Features
1. Infer whether a text-based review of a product is negative, neutral, or positive	Feature can be a vector of words and the associated frequencies of those words in spam and ham emails.
2. In a computer network, detect unusual traffic that may be indicative of an attack against the network	Features can be the data comprising the traffic that is present in normal situations and also of the situations where more traffic is expected to detect the anomaly.
3. Predict the selling price of an apartment or house based on features describing the property and its location	For such a problem, the features can be the number of rooms, area of the house, locality, presence of private garage space.
4. Learn to automatically set the bitrate for a video stream viewed by a user at any point in time based on information about the current connection quality, state of the video buffer, location of the user etc	For this problem, features can be the user location, the strength of the user's connection, the state of video buffer.
5. Predict whether or not an advertisement shown on a web page will be clicked by a user, based on information about the user, ad, and web page	Features for this would include the past trends of the user, interest of the user, interests of users who bear similarity to the target user, information on similar items in the past to which user has reacted to.
6. Learn to generate photorealistic images of human faces	Features for this situation can be a set of images of human faces on which the model will learn.
7. As an email service provider, detect groups of similar email messages that may constitute a spam campaign	Features can be a known set of emails containing words that are typically known to be spam along with the label for spam or not.
8. Learn to play a real-time strategy computer game such as StarCraft	Features can be the current game state in the form of a map which represents the environment state in which the model is to maximize the reward.

Task 3

1. Here, the situation is to build a filter which classifies an email as spam or not. In this scenario, the assumption that the test set and the real time system will receive examples of the same distribution is not fully valid. This is because, over the years with increase in technology, the spammers have improvised new techniques to creep in our inbox by analyzing the past trends and hence the emails which the filter will receive at application time can deviate significantly from the test data.

2. The model here has been trained by means of speakers that have recruited via ad posted in a local newspaper. Now, each locality has different accent and the way one communicates and pronounces words. Here, as the model was trained by speakers from the ad, the diversity will be very limited and thus, this model cannot be scaled to a model to be deployed worldwide due to the wide variation in the way people recognize spoken commands all over the world. Such a model would suit the particular locality from where the model was trained but not worldwide. Hence, the I.I.D assumption will not be suited for the model.

3. Stock prices of companies show a very dynamic changing pattern and the trends of stock price movements accumulated from the past data will deviate significantly when we try to predict the stock prices of the companies in future based on the data we collected. So, in this scenario the I.I.D assumption will not be valid due to the differences in the data arising at run time as compared to that used for training.

4. The dynamics of road movement involving the movement of cars and pedestrians do not follow the same trend every time and is also of a dynamic nature. It cannot be assumed that the trend in car and pedestrian movements we see for the last three months will be valid for the future also. Also, each state has its own set of such movements and gathering data from only two states will further lead to deviation from the I.I.D assumption that the distribution of data at application time will follow the test set distribution. However, if the distribution of cars and pedestrians with all other cities are a standard model that is followed throughout in US, then the I.I.D assumption can hold true with reasonable accuracy.