Capstone Project Possibilities

Tom Thorpe

March 13, 2018

Capstone Project Possibilities

Iris Dataset

This is a small data set used to classify 3 types of Iris flowers based on physical characteristics of the flower. It has 150 entries with 4 inputs of sepal and petal length and width in centimeters. The last column is the name of the iris and is the outcome that is to be predicted. The data can be found here:

https://archive.ics.uci.edu/ml/machine-learning-databases/iris/ in the iris.data file.

The questions to ask are:

-How can the four physical characteristics of the flower be used to predicate the type of flower?

-How should the data be scaled/adjusted for best use for machine learning?

Titanic Survival Dataset

This is a dataset of 1309 records, one per passenger on the Titanic with 12 columns of data per record. The outcome to predict is column 2: "Survived". There is missing data for age and sibling and parent child counts. Other input data include passenger class, price of ticket, sex, fare, cabin number and location of embarkation. The data can be found here:

https://www.kaggle.com/c/titanic/data

The train.csv dataset contains the expected outcome. The test.csv does not have the expected outcome. The expected outcome for the test.csv dataset can be found in the gender_submissions.csv dataset.

The questions to ask are:

-Which data can be used to predict survival rate?

-How should the data be scaled/adjusted for best use for machine learning?

Forest Coverage Dataset

This is a much larger dataset of 580,000 rows and 54 columns of data. The last column, Cover Type is the type of tree at the sampled location and is the value to be predicted. There are 7 types of trees in the outcome. Only one type of tree is listed for each record. The variables include elevation, slope of the terrain, distance to water and roads and one of 40 different soil types. The data can be found here:

https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/

The questions to ask are:

-Which data can be used to predict the type of tree at each sample location/record?

-How should the data be scaled/adjusted for best use for machine learning?

-Does breaking out the components of the soil type as inputs give better results than clustering soil components into a one of 40 selection. For example one soil type is described as consisting of, "Vanet ad Ratake families complex, very stony" in the Lower Montane climatic zone in the igneous and metamorphic geologic zone.

One investigation would be to see if splitting out the families, rock types, climatic and geologic zones into inputs would improve the reliability rate.