University of Essex

Department of Mathematical Sciences

Part 1: Pilot-Study Proposal

Subject: CE802 Machine Learning

Word count: 727

INTRODUCTION:

A pilot-research proposal is a preliminary inquiry designed to gather data in preparation for a bigger, more comprehensive study. A pilot study's purpose is to examine the feasibility and acceptability of a method that will be used in a larger research, rather than to test hypotheses regarding the effects of an intervention. The research topic, aims, methods, and expected outcomes of the pilot study should all be clearly stated in the proposal. It should also include a thorough explanation of the research population, sample size, data collecting and analysis procedures, and ethical concerns.

Nowadays, many people travel by bus or airline. A travel insurance company can help travelers secure their holiday or trip from unplanned damages. Since Customers who are less likely to submit a claim in the future should pay a lesser premium, according to a travel insurance firm. We need to predict if a customer will submit a claim in the future.

The four primary components presented here are predictive tasks, informative features, learning procedures, and evaluation.

PREDICTIVE TASK

In this challenge, we must categorize on the basis of a machine learning model. We have a feature variable made up of labeled preparation information and an ideal target variable, much like in supervised learning. When data is used to predict categorical variables, controlled learning is often referred to as classification. When there are just two labels, this is referred to as binary categorization. When there are numerous classes, the problem is referred to as multi-class classification. A regression predictive model is used to predict continuous values. As a result, forecasting whether a client would claim insurance in the future comes under the category predictive task (True or False).

SPECIFIC INFORMATIVE FEATURES

Clinical and biochemical characteristics are strong indicators of diabetes risk. The strongest clinical predictor of diabetes is obesity, while the best biological predictor is baseline glucose. Other clinical factors that might indicate diabetes include hypertension, male smoking, and female triglycerides. Being overweight, having a family history of diabetes, and being physically inactive are all risk factors for type 2 diabetes. Machine learning algorithms for predicting the probability of getting type 2 diabetes have also been created. High triglycerides and low HDL "good" cholesterol can raise the risk of developing type 2 diabetes and cardiovascular disease.

LEARNING PROCEDURE

The method or approach used to train a machine learning model is referred to as the learning procedure. The learning process to be used is determined by the type of issue to be solved, the available data, and the performance indicators to be optimized.

Decision Trees (DTs): Decision trees are often easy to understand and comprehend. It is also non-parametric, and no distribution is required. Having said that, we don't have to worry about outliers or if the data set can be divided linearly when we utilize decision trees. Decision trees, as a heuristic approach, do not suffer from multicollinearity and are useful for just a few types of variables.

k-Nearest Neighbours (k-NN): A supervised learning method used for classification and regression, k-NN is a supervised learning technique. It works by locating the k-nearest data points in the training set to the input data point and predicting the output of the input data based on their output values.

SVMs (Support Vector Machines): The benefit of SVMs is their excellent accuracy and performance. It gives solid theoretical assurances against overfitting and features a flexible selection of kernels for data that isn't linearly separable. It is highly common in text categorization issues, especially when very high-dimensional spaces are involved.

EVALUATION

Many binary classification algorithms produce a prediction score as their final result. The model's confidence that a particular observation belongs to a certain class is indicated by the score. As a consumer of this score, we will interpret the score by selecting a classification threshold (cutoff) and comparing the probability score to it to determine whether the observation should be classed as positive or negative. Any observations with scores greater than the threshold are predicted to be in the positive class, while any observations with scores less than the threshold are predicted to be in the negative class.

REFERENCE

1. Browne, R.H. (1995). On the use of a pilot sample for sample size determination. Statistics in Medicine, 14, 1933–1940.