Определение сложности выборки с помощью универсальной аппроксимирующей модели

В данной работе исследуются свойства обучающей выборки в задачах классификации и регресии. Сложность обучающей выборки может быть определена с помощью различных принципов таких как принцип минимальной длины описания и байесовский подход. Однако сложность может быть также вычислена с помощью аппроксимирующих моделей. Данный подход является более прикладным. В данной работе предложено определение сложности обучающей выборки с помощью двухслойной полносвязной нейронной сети. Согласно теореме Цыбенко нейронная сеть прямой связи с одним скрытым слоем может аппроксимировать любую непрерывную функцию многих переменных с любой точностью. Соотвественно, в качестве определения сложности было использовано количество нейронов на внутреннем слое нейронной сети.

В работе была показана корректность данного определения, а также исследованы его свойства. Для демонстрации полезности определения сложности предложенным способом были проведены вычислительные эксперименты на синтетических данных с различными параметрами мультикоррелированности признаков и уровнем шума. Также были проведены эксперементы на реальных данных для задач регресии и бинарной классификации.

Determination of sampling complexity using a universal approximating model

This paper investigates the properties of the training dataset in classification and regression problems. The complexity of the training data can be determined using various principles such as the minimum description length principle and the Bayesian approach. However, the complexity can also be computed using approximating models. This approach is

more applied. In this paper, we propose to determine the complexity of the training dataset using a two-layer fully connected neural network. According to Tsybenko's theorem, a neural network of feed forward with one hidden layer can approximate any continuous function of many variables with any accuracy. Accordingly, the number of neurons on the inner layer of the neural network was used as the definition of complexity.

The correctness of this definition was shown, and its properties were investigated. To demonstrate the usefulness of determining the complexity of the proposed method, computational experiments were carried out on synthetic data with different parameters of multicorrelation of features and noise level. Experiments were also carried out on real data for regression and binary classification problems.