Overview of Artificial Neural Networks Example

Dataset Link:

<https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling>

Dataset Content:

This dataset contains thirteen columns. Most of them provide information about customers (surname, credit score, balance, estimated salary etc.).The last column contains values of 0 and 1 which indicate whether the customer still has a bank account (0) or they left the bank (1)

Problem:

The bank hired us in order to inform them which customers are more prone to leave the bank in the next few months.

Solution:

We want to develop a classification model using Artificial Neural Networks which allows us to predict whether a customer is more probable to leave the bank or not based on some features we were provided.

Variables:

Independent variable 🡪 Credit score, Geography, Gender, Age, Tenure, Balance, Num of Products, Has Card, Is Active Member, Estimated Salary.

Dependent variable 🡪 Still a bank account holder (0) or left the bank (1).

**\*IMPORTANT NOTE**

Some of features (independent variables) we are going to use in this ANN example are categorical and therefore we need to encode them before moving forward. More specifically we have the Geography variable (3 categories: France, Spain, Germany) and Gender (2 categories: Male, Female). For Geography we are going to use *LabelEncoder* to convert the strings (names of countries) into integers 0,1,2 with each one representing a specific country.For Gender we are going to use *LabelEncoder* to convert the strings (gender types) into integers 0 (for Females) and 1 (for Males).

Since the Geography variable is not an ordinal variable (France is not greater than Germany, Spain is not greater than France etc.) we need to create a dummy variable. Simply put, machine learning models are based on equations and it is crucial to replace categorical data with numerical BUT since 1 (e.g. Spain) is greater than 0 (e.g. Germany) and 2 (e.g. France) is greater than 1 (e.g. Spain) our model will “think” that Spain (1) has a higher value than Germany (0). However this is not the case because these numbers just represent categories and there is no relational order among them. This means that instead of having one column for the Geography variable we are going to have three columns (one column for each category) as shown below:

|  |  |  |
| --- | --- | --- |
| France | Germany | Spain |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 0 | 1 |

|  |
| --- |
| Geography |
| France |
| Germany |
| Spain |
| Germany |
| France |
| France |
| Spain |

To do that we are going to use the *OneHotEncoder* class. After that, to avoid the so called dummy variable trap we are going to delete one of these columns.

We don’t need to create a dummy variable for the Gender variable as it only has 2 values (0 and 1).