Machine Learning Challenge: Day 2

Welcome to the second day of our 30 days Machine learning Challenge.

Here is the Project One: Classification Walkthrough using Titanic Dataset:

The data has been split into two groups:

- training set (train.csv)
- test set(test.csv)

The training set includes passengers' survival status (also known as the ground truth from the titanic tragedy), which along with other features like gender, class, fare, and pclass is used to create a machine learning model. The test set should be used to see how well my model performs on unseen data. The test set does not provide passengers with survival status. We are going to use our model to predict passenger survival status.

Let's describe what the meaning of the features has given in both train & test datasets.

Variable Definition Key

- Survival
  - 0= No
  - 1= Yes
- pclass (Ticket class)
  - 1=1st
  - 2=2nd
  - 3=3rd
- sex
- age
- sibsp (# of siblings / spouses aboard the Titanic)
- parch (# of parents / children aboard the Titanic)
- tickets
- fare
- cabin
- Embarked Port of Embarkation.
  - C = Cherbourg,
  - Q = Queenstown,
  - S = Southampton
- Pclass: A proxy for socio-economic status (SES) This is important to remember and will come in handy for later analysis.
  - 1st = Upper
  - 2nd = Middle
  - 3rd = Lower

Supplementary contents:

Supervised Learning is a machine learning and artificial intelligence subcategory. It is distinguished by using labeled datasets to train algorithms that accurately classify data or predict outcomes. As input data is fed into a certain model, the weights are adjusted specifically until the model is properly fitted. This occurs as a part of the cross-validation process.

Supervised learning examples

There are some very practical applications of supervised learning algorithms in real life, including:

- Text categorization
- Face Detection
- Signature recognition
- Customer discovery
- Spam detection
- Weather forecasting
- Predicting housing prices based on the prevailing market price
- Stock price predictions, among others

In supervised Learning, a training set is used to teach models how to make the right output. This training dataset has both right and wrong answers, which helps the model learn over time. Using the loss function, the algorithm figures out how accurate it is and changes until the error is as small as possible.

Challenges of supervised Learning

Although supervised Learning can provide businesses with benefits such as deep data insights and improved automation, building long-term supervised learning models can be difficult. Some of these challenges are as follows:

1. Specific levels of expertise may be required for structured, supervised learning models.
2. Training supervised learning models takes time.
3. Datasets may contain more human error, causing algorithms to learn incorrectly.
4. Unlike unsupervised learning models, supervised Learning cannot independently cluster or classify data

Supervised Learning can be divided into two types of problems:

1. **Classification** and
2. **Regression**

Classification in Machine Learning

"Classification" refers to identifying, understanding, and placing things or concepts into distinct groups. With the help of these pre-classified training datasets, ML programs use a variety of algorithms to classify certain future datasets into respective and relevant categories.

Classification can be considered a specific subset of "pattern recognition." When applied to training data, classification algorithms can recognize recurring patterns (consecutive numbers, words, emotions, etc.) in new data. It entails categorizing the data. If you consider extending credit to someone, classification can help determine a borrower's likelihood of default. When the supervised learning algorithm labels input data into two classes, binary classification occurs. Multiple classifications imply categorizing data into more than two groups.

What is Classification Algorithm?

To classify raw data points in the wild, the Classification algorithm employs supervised Learning. By analysing the provided dataset, a computer program can "learn" to classify new data into the categories that have previously been established. To function properly, the Classification algorithm needs input data that has already been labeled, as it is a supervised learning technique with both input and output data. Each input variable in a classification problem is associated with a discrete output function (y) (x).

Classification is a pattern recognition technique in which training data and a classification algorithm are used to find instances of the same pattern in data that has not been subjected to the training process.

Naive Bayesian Model

For large finite datasets, the Bayesian classification model is used. It is a method for assigning class labels that uses a direct acyclic graph. The graph has one parent node and several child nodes. Furthermore, each child node is assumed to be independent and distinct from the parent.

Because the supervised learning model in ML assists in straightforwardly constructing classifiers, it works well with very small data sets. This model is based on common data assumptions, such as the assumption that each attribute is independent. Despite its simplification, this algorithm can be easily applied to complex problems.

Decision Trees

It is a flowchart-like model that includes conditional control statements for decisions and their likely outcomes. The output is concerned with labeling unexpected data.

The internal nodes in the tree represent attributes, while the leaf nodes represent class labels. A decision tree can solve problems that have both discrete and Boolean attributes. Two well-known decision tree algorithms are ID3 and CART.

The Random Forest model

The ensemble method is what the random forest model is. It works by making a lot of decision trees and then putting them into different groups. Say you want to predict which undergraduates will do well on the GMAT, a test needed to get into graduate programs in management. Given the demographic and educational information about a group of students who have already taken the test, a random forest model would be able to finish the job.

Support Vector Machines

It is a supervised learning algorithm developed in 1990. Vap Nick's statistical learning theory inspires it.

The Kernel function's main goal is to find a high-margin hyperplane that helps divide the observations into the maximum distance between the hyperplane and the nearest point from each class. Finding this constraint is important because it makes it less likely that the resulting hyperplane will be too good. SVM changes a space with fewer dimensions into a space with more dimensions using kernel functions like similarity functions. The Kernel functions can be different based on the type of data set. Most of the time, SVM uses more than one Kernel function because it's not always clear which Kernel function is best for making the data more complex.