FAQ - Intro to Supervised Learning - Linear Regression

1. How to import a dataset in Jupyter Notebook?

Loading data files in Jupyter Notebook slightly differs from Google Colab. Please refer to **this link** and follow the steps mentioned to load a dataset.

Note: To avoid errors, please ensure that the name of the dataset is correct - check for lower case and upper case, check for spaces, name of the data set is like Automobile(1).csv, etc

2. What is Machine Learning?

Machine learning is the branch of science that deals with the machine's ability to learn from existing data, then perform some task based on the learning and use some performance indicators to measure how accurately the task is being performed.

For example: playing chess. The experience of playing many games of chess is the learning part. The task of playing chess and the probability that the program will win the next game is the performance indicator.

In general, any machine learning problem can be assigned to one of two broad classifications: Supervised learning and Unsupervised learning.

3. What is Supervised Learning?

Supervised learning is when you already know the label of the target variable. It is of two types: regression and classification. Regression is when the target variable is continuous and classification is when the target variable is in the form of categories or discrete values.

For Example:

Regression: House Price prediction based on area, number of rooms, lawn, pool, etc.

Classification: Predicting whether a person will be diabetic in the future or not based on bp, glucose, insulin, etc.

As you can see here, it is already known the target variable is the price in case of regression and whether the person is diabetic or not in case of classification. So we know what we want to predict.

4. What is a Model?

A machine learning model can be a mathematical representation of a real-world process. An ML model takes the training data as input and with the help of an ML algorithm, forms a mathematical expression that gives us an output.

For example, the **house price prediction model** will take the area, several rooms, lawn, pool, etc. as input and give the price of the house as output.

5. What do you mean by an algorithm?

An ML algorithm takes the training dataset to form a mathematical relationship between independent variables and dependent variables. The algorithm will give the best fitted mathematical expression based on available training data.

6. What is the intercept in linear regression?

The intercept in linear regression is the value at which the regression line crosses the y-axis. It is the estimated value of the target when all the predictors are 0. However, the intercept is not always interpretable.

7. Do I need to install any new libraries for this course?

The *sklearn* library is available by default with Anaconda, so you do not need to install it separately.
The *nb_black* extension can be installed by running the following command in a Jupyter notebook:

```
!pip install nb-black
```

8. Why should we use drop_first=True to drop one column while creating dummies?

Consider the feature 'Gender' which has two categories 'Male' and 'Female'. The two columns 'Gender_Female' and 'Gender_Male' will be formed because of *get_dummies()* and the Gender column will be dropped. Now, because *drop_first* = *True*, one column of each category will be dropped. Suppose 'Gender_Female' is dropped

Dropping a column is logical because if the gender is not male, it will be female. That extra column (Gender_Female) adds no value.

Even if gender had three categories 'female', 'male', and 'other', then also dropping one makes sense, because if it's not 'male' or 'other', then it has to be 'female'.

The machine only understands numbers, and the feature name makes no sense to it. Whenever a nominal category column is there, the feature will have values 1 or 0, i.e., whether that particular row belongs to that particular feature or not.

Let us again consider 'Gender'. If the feature is 'Gender_Female' and the value is 1, then the machine only understands that the value for this particular row is 1. If we include the 'Gender_Male' feature as well, the value there will be zero. So adding that column is not significant.

9. Why we use .fit() and .predict() ? What does .fit() and .predict() do?

.fit(): Forms a mathematical equation with the help of a training dataset.
.predict(): Will use the mathematical expression obtained from *fit()* and give the output based on it.

Consider the linear regression equation $y = a1x1 + a2x2 + a3x3$, where *x1*, *x2,* and *x3* are three different features. When we use *.fit()*, the values of all the coefficients are calculated, and when we use *.predict()*, the feature values for every particular row are used to calculate *y* for that row. So *.predict()* gives a Y value as output for each row.

10. Why do we split the data into test and train data while building a supervised learning model?

The goal of machine learning is to predict well on new data drawn from a (hidden) true probability distribution. Unfortunately, the model we are building at present can't see the whole truth; the model can only sample from an available dataset. If a model fits the current examples well, how can we trust the model to make good predictions on never-before-seen examples?

One way is to divide your data set into two subsets:

. Training set: a subset to train a model.
. Test set: a subset to test the model.

Separating the data enables you to evaluate your model generalization capabilities and have an idea of how it would perform on unseen data. Good performance on the test set is a useful indicator of good performance on new data in general, assuming that:

. The samples were drawn independently and at random from the distribution to create the test set.
. The test set is large enough.

11. When and how do you bring in test data?

Initially, the dataset that is provided for analysis is split into train and test sets. The test set should be such that it is representative of the population on which the model is going to make predictions.

12. Why do we need to study the mathematics behind the algorithms when we can implement the algorithms using simple codes?

The supervised learning course is the first step in the introduction to modeling in your AI-ML journey. The mathematical content in the videos is covered to give a perception of how algorithms work at the backend and the hands-on videos are covered to demonstrate the implementation of the algorithms. The mathematical intuition behind the algorithms is helpful when you need to improve the performance of your model, in such situations knowing about hyperparameters and what parameters are affecting your model will give you an advantage. A deeper understanding of the algorithms increases the interpretability power of your model.

Previous

Next