Assignment 1: Regression

Innopolis University

Machine Learning Fall 2020 - Bachelors

1 General Instructions

In this assignment, you are going to solve two problems. You will try different regression models to find the best model that fits a function in the first one. In the second problem, you will use logistic regression to help a bank discover whether a customer will subscribe to the term deposit or not.

You are required to submit your solutions via Moodle as a single ipynb file. Do not forget to include your name in the submitted document.

The source code should contain adequate internal documentation in the form of comments. Internal documentation should explain why and how you apply the instructions.

Bonus points might be awarded for elegant solutions and great documentation. However, these bonus points will only be able to cancel the effect of penalties.

Plagiarism will not be tolerated, and a plagiarised solution will be heavily penalized for all parties involved. Remember that you learn nothing when you copy someone else's work, which defeats the exercise's purpose! You are allowed to collaborate on general ideas with other students as well as consult books and Internet resources. However, be sure to credit all the sources you use to make it clear what part of your solution comes from elsewhere.

2 Linear/Polynomial Regression (5 points)

First of all, you will create a synthetic dataset by sampling from data generated by adding some random Gaussian noise to a sinusoidal function. Here is the code that you should use for creating the data:

```
rng = np.random.RandomState(1)
x = 10 * rng.rand(100)
y = np.sin(x) + 0.1 * rng.randn(100)
```

Next, you must apply regression models on the previous data for predicting y. Choose the degree of your model (if it is polynomial) according to the model's performance. Evaluate the models' accuracy using cross-validation with k=10 (You should report MSE and std in each fold). You should also plot the prediction of your model on X_{test} where:

X_{test} = np.linspace(0, 100, 100)

Show if your model performs better with applying regularization (test both L2 and L1 regularization). Plot the the model against the X_{test} for the best model with and without regularization.

3 Logistic Regression (10 points)

Classification is a prevalent machine learning task and can be applied widely across various disciplines and problem statements. In this task, you will implement logistic regression on the Bank Marketing dataset from the UCI Machine Learning Repository. The dataset represents the direct marketing campaigns of a bank and whether the efforts led to a customer subscribing to a bank term deposit. The data is related to the direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

3.1 Data Reading and preprocessing (5 points)

First, we need to convert all non-numeric features into numeric ones. There are two popular ways to do this: **label encoding** and **one hot encoding**. For label encoding, a different number is assigned to each unique value in the feature column. A potential issue with this method would be the assumption that the label sizes represent ordinality (i.e., a label of 3 is greater than a label of 1). For one hot encoding, a new feature column is created for each unique value in the feature column. The value would be 1 if the value was present for that observation and 0 otherwise. However, this method could easily lead to an explosion in the number of features and lead to the curse of dimensionality.

Then, we need to impute all missing values. Removing the rows/cols with missing values could be an option when the number of such rows/cols is relatively low compared to what's left. However, in the usual case, they are not deleted, as deleting them could lead to information loss.

Another important step is data normalization or feature scaling. Since the range of raw data values varies widely, in some machine learning algorithms, objective functions will not work properly without normalization.

To summarize, here is a list of things that you need to do as part of data preparation:
Step 1: Import the libraries
Step 2: Import the dataset
Step 3: Encode the Categorical data
Step 4: Check out the missing values and impute them using the mean value strategy.
Step 5: Feature scaling
Step 6: Splitting the dataset into Training (80 %) and Test set (20 %)

3.2 Model Creation (5 points)

Finally, you will apply logistic regression to build a model that would predict if the bank's product (bank term deposit) would be ('yes') or not ('no') subscribed for an unseen instance.

To choose the best hyper-parameters for your model, you should use Grid-Search-CV. That is, you should try differnet variations of Logistic Regression using variations with respect to size of penalty ['l1', 'l2'], and type of solver used : ['liblinear', 'lbfgs'], $classifier\_\_C$: $np.logspace(-4, 4, 20)$.

Grid-Search: It is an exhaustive hyper-parameter tuning algorithm that checks all the possible hyper-parameter combinations and chooses the best one of them. For evaluation of every set of hyper-parameters, Grid-Search- CV uses cross-validation to evaluate the model's performance. You can refer to these links for more information: GridSearchCV or LogisticRegressionCV

After you have found the best set of hyper-parameters for the model, you will evaluate it on the separate test data-set and report: Accuracy, Precision, and Recall.