**Homework 1**

Machine Learning Course

Oğuz Elçi

[oguz.elci@gmail.com](mailto:oguz.elci@gmail.com)

24/03/2021

1. How would you define Machine Learning?

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. In contrast to programming a system to perform a certain task, in machine learning practices, the inputs and outputs are fed to the model and the program which is supposed to maintain the relationship between inputs and outputs is tried to be got. The model is actually created after the determination of the relationship between inputs (features) and outputs (labels). Then this model is used to predict the outputs when a new input is fed to the model.

2) What are the differences between Supervised and Unsupervised Learning? Specify example 3 algorithms for each of these.

If we’re able to label the data and the data is either of continuous numerical nature or of discrete nature, we can use supervised learning but on the otherhand if the data is not labeled and we aim the model to discover the hidden patterns in the dataset and cluster them, we can use unsupervised learning technique.

Examples :

Differentiating e-mails as spams or not-spams : Classification : Discrete values : Supervised

The effect of daily alcohol consumption on blood pressure : Regression : Continuous values : Supervised

Predicting the weather : Multiple linear regression : Continuous values : Supervised

Finding customer segments from a dataset of customers : Clustering : Unsupervised

Bad or good comments on a product which is sold online : Clustering : Unsupervised

Weather forecast : K-means clustering : Unsupervised

3) What are the test and validation set, and why would you want to use them?

The model is a kind of mathematical model which finds out the relationship between inputs and outputs that fed to it. While doing so, the data in hand are divided in to some subsets for its convenience.

In simpler models approximately 70% of the data is set aside as training set and the rest as test set. The model learns the relationship with training set. Test set is a set which model has not encountered before. The test set is used to see if the model gives foreseen outcomes with this new dataset.

On the otherhand the data is divided to training, validation and test sets in more complex models. Model is trained with train dataset and validation dataset is used to tune hyperparameters. In this approach data is divided in to 3 so that the training set is 70% of all data whereas validation set and test set is 20% each.

4) What are the main preprocessing steps? Explain them in detail. Why we need to prepare our data?

Removal of duplicate data : The repetitive data is deleted if collected by mistake. The abundance of data forces the model show tendency to that data.

Missing data : In input, output relationships if some data are missing at some data points, these voids are filled with median or mean

In case of inbalanced data : If a group of data is less or more compared to the other group, undersampling or oversampling is done to make comparison reliable.

Outliers : Outliers are determined with 2 different methods. In one of which if the data is more than 3 sigmas away from the mean, it is considered as outlier and not included in the model training not to create bias.

Standardization/normalization : This allows the model converge to a better fit.

Bucketing : The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin.

5) How you can explore countionus and discrete variables?

Continuous Data can take any value (within a range). The examples are the following :

* A person's height: could be any value (within the range of human heights), not just certain fixed heights,
* Time in a race: you could even measure it to fractions of a second,
* A dog's weight,
* The length of a leaf

Discrete Data can only take certain values. For example; the number of students in a class. We can't have half a student! Or the results of rolling 2 dice can only have the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12.

6) Analyse the plot given below. (What is the plot and variable type, check the distribution and make comment about how you can preproccess it.)

The plot is a distribution plot. But the distribution does not show normal distribution. Since it is the distribution of petal width of numerous petals, we can easily say that the petals are either from two different plants or they were collected in two different times the latter group is probably more grown petals compared the other group.

The variable is continuous.

Pre processing : We may need to divide these petals in to two groups. Bucketing method can be applied.

<matplotlib.axes.\_subplots.AxesSubplot at 0x2b8ced01320>

