Introduction to Machine Learning

Ali Al Shammaa

July 1, 2021

1 Introduction

In this document, the focus is an intro to ML with a focus on underlying concepts using decision trees.

2 Definition of learning

Learning can be defined as the process of finding a function approximation to the function that maps the inputs to the outputs in a sample of the dataset called the training set without seeing the full dataset. This approximation should be able to generalise about future inputs from the training inputs.

The labels in a dataset can be continuous (Regression) or discrete (binary or multi classification).

3 Probabilistic model of learning

The probabilistic model of learning is based on a probability distribution function over all the possible input/output pairs, instead of functions that maps inputs to outputs. Let D be the data generating probability distribution over all the possible input/output pairs. For a given pair of input/output, D produces the probability that the pair is reasonable.

However, we don't have access to D, but only a sample of it. Our aim in this model is to approximate this distribution using the sample, the training set, by outputting for a given input chosen at random the label with the highest probability. D gives high probability for reasonable Input/Output pairs and low probability for unreasonable ones. A pair can be unreasonable in two ways. Assume we are considering the MNIST problem. One way is when the input is unlikely such as a very awkward looking four, and the other way is when the label is very unlikely for the image.

If we introduce a cost function, we can then capture the idea of doing well on the training sample. The cost function will be the average of the individual costs for each example. For the test sample drawn from D, we can just go with a measure of accuracy (perhaps, error rate).

The expected loss (i.e. average loss) for a discrete probability distribution D is the sum of all the probability-weighted loss across all pairs in D. In this case, more probable pairs are penalised more if the corresponding loss is high. The expected loss is

$$E_{(x,y)\ D}(C(f(x),y)) = \sum_{(x,y)\ D} [D(x,y)C(f(x),y)]$$

”The difficulty in minimizing our expected loss from Eq is that we don't know what D is! All we have access to is some training data sampled from it!”

4 What if we had access to D

The Bayes Classifier