T/UDOM/2017/11329
TUMAINI STEVEN

REPORT 2: CHAPTER 2

SUPERVISED LEARNING

supervised learning is used whenever we want to predict a certain outcome from a give input, and we have examples of input/output pairs. In supervised learning, we want to build a model on the training data and then be able to make accurate predictions on new, unseen data that has the same characteristics as the training set that we used.

Difference between Generalization, overfitting and underfitting

When we speak about **Generalization** If a model is able to make accurate predictions on unseen data, we say it is able to generalize from the training set to the test set. We want to build a model that is able to generalize as accurately as possible.

Overfitting means Building a model that is too complex for the amount of information we have. Overfitting occurs when you fit a model too closely to the particularities of the training set and obtain a model that works well on the training set but is not able to generalize to new data

Underfitting On the other hand, if your model is too simple say, "Everybody who owns a house buys a boat"—then you might not be able to capture all the aspects of and variability in the data, and your model will do badly even on the training set. Choosing too simple a model is called underfitting. The more complex we allow our model to be, the better we will be able to predict on the training data. However, if our model becomes too complex, we start focusing too much on each individual data point in our training set, and the model will not generalize well to new data. There is a sweet spot in between that will yield the best generalization performance.

Difference between classification and regression.

An easy way to distinguish between classification and regression tasks is to ask whether there is some kind of continuity in the output. If there is continuity between possible outcomes, then the problem is a regression problem. Think about predicting annual income. There is a clear continuity in the output. By contrast, for the task of recognizing the language of a website (which is a classification problem), there is no matter of degree. A website is in one language, or it is in another. There is no continuity between languages, and there is no language that is between English and French.

ALGORITHMS

- **KNN Algorithm:** is the simple algorithm that stores all the available cases and classify the new data based on a similarity measure. here we have the value ‘k’ which is th total number of neighbours’ chosen and how do we choose the value of ‘k’ is something we should ask ourselves

- **Decision tree algorithm:** is the graphical representation of all possible solutions to a decision. This decision is based on some conditions. Decision made can be easily explained Why is it called decision tree anyway? This is because I start from the root and then branches off to branches (various decision and various conditions)

- **Logistic regression algorithm:** is the most famous machine learning algorithm after linear regression. In a lot of ways logistic regression and linear regression are similar. But the biggest difference lies in what they are used for linear regression is used in predicting values while logistic regression is used in classifying values