# T/UDOM/2017/16518

CHAPTER 2

**BOOK: introduction to machine learning with python**

**Supervised learning**

In supervised learning, the user provides the algorithm with pairs of inputs and desired outputs, and the algorithm finds a way to produce the desired output given an input. In particular, the algorithm is able to create an output for an input it has never seen before without any help from a human.

**Classification vs regression**

**Classification**: predict a class label

-binary:distinguishing between exactly 2 cases

-multiclass:classification between more than 2 classes

Example:-yes/no questions

-spam/non spam

**Regression**: predict a continuous number or a floating number in programming term (or real number in math)

Example: -predict a person's annual income from education ,age,where they live (the predicted value is an amount and can have any number in a given range)

-predict the yield of a corn farm given attributes such as yields, weather, and number of employees working on the farm

It can also be an arbitrary number

**Generalization, overfitting, undefitting**

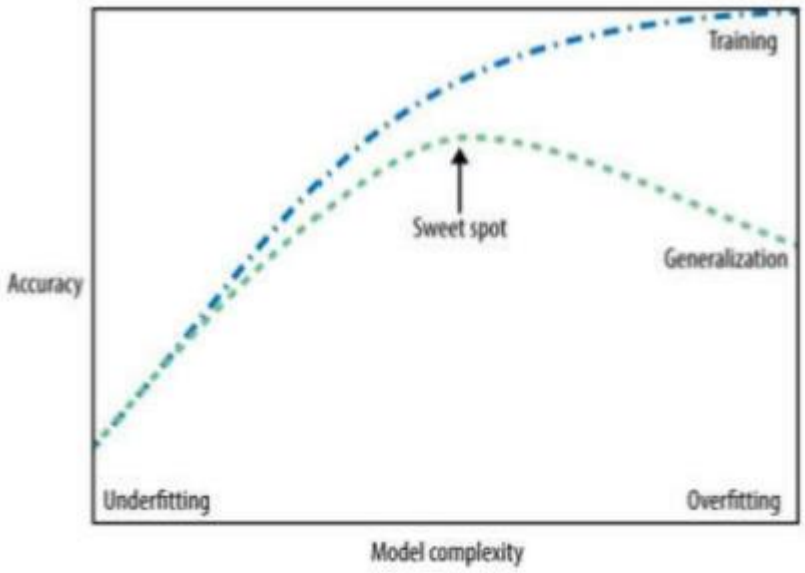
- if a model is able to make accurate predictions on unseen data, we say it is able to generalize from training set to the test set

**overfitting**: building the model that is too complex for the amount of data owned. Occurs when you fit a model too closely to the training set but is not able to generalize to new data choosing too simple model is called **underfitting**

-example: "everybody who owns a house buys a boat"

- ❖ Does not capture all the aspects of and variability in the data and does bad on the training set
- ❖ The more complex the model, the better will be able to pr  However, if the model becomes too complex, it fails to generalize well to new data  The model needed is the one in between the two

The tradeoff between overfitting is illustrated below




Model complexity vs dataset size

- The larger variety of data points your dataset contains, the more complex a model you can use without overfitting.
- Larger datasets allow building more complex models
- However, duplicating the same data point or collecting very similar data will not help
- Having more data and building more complex models can often work wonders for supervised learning tasks
- Never underestimate the power of more data

## Supervised learning algorithm

- **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