Machine Learning

Lesson 2

Overfitting, Linear Regression & Decision Trees

Lecturer: Dr. Msagha J Mbogholi, PhD

Flashback from Lesson 1

- A simple program cant give meaning to all the data in the digisphere. There is need for ML techniques to extract the data and give meaning to it.
- A learning task can be divided into three distinct parameters: the task (T), performance measure (P) and the training experience (E).
- There are four steps that are involved in machine learning design as described by Mitchell (1997). These are: choosing the training experience, choosing the target function, choosing a representation of the target function, and choosing a function approximation algorithm.
- Machine learning systems can be classified according to the amount and type of supervision they get during training. There are four major categories: Supervised learning, unsupervised learning, semi supervised learning, and reinforcement learning

Content

- Introduction
- Definitions & Concepts
- Overfitting
- Linear Regression
- Decision Trees

Part 1

Introduction

Introduction

- In our previous lesson we discussed the different types of machine learning and how they are classified.
- In this lesson we dig deeper into various machine learning techniques that help to solve problems.
- Regression was introduced in lesson one as falling in the class of supervised learning techniques. In this lesson it is discussed further and two different types of linear regression are discussed.
- The lesson will also discuss the working of decision trees.
- However, before discussing all these some new terms are introduced first including the concept of overfitting.

Part 2

Definitions & Concepts

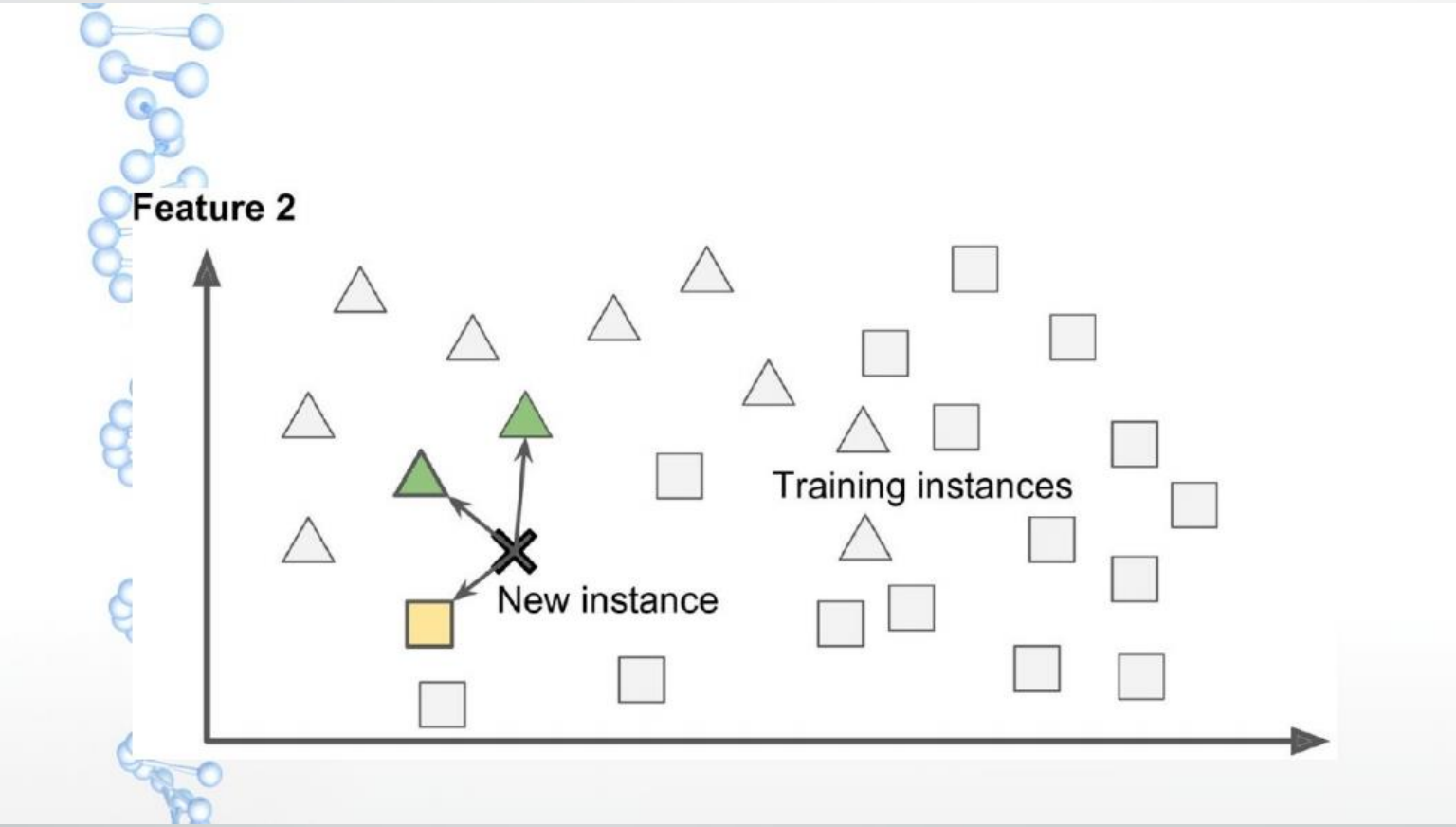
2.1 Definitions

- Node – a point where a choice must be made.
- Noise – when collecting data it is to be expected that some of the data may either be irrelevant or incorrect (can be due to something as simple as human error or error in the data collection instrument (s)). This kind of data when included in a dataset is referred to as noise.
- Predictor variable – it is an independent variable that is used to predict some other variable or outcome.
- Target (variable) – it is the one being investigated based on the input from the predictor variable (s).

2.2 Concepts

- Instance-Based vs. Model-Based Learning:
- Machine learning can also be categorized as being either instance-based or model-based.
- In instance-based learning, a new instance which we wish to classify is compared to the instances in the training data.
- Fig 1 demonstrates how instance based learning works. The new instance takes the class of the instances that are most similar to it; in this case between the triangle and square instances.
- Another good example is with email: if a new email has many words in common with a spam email, the new email may be classified as spam.
- Instance-based learning is simple and will not work well in complex problems. Therefore it has limited application.

Instance – Based Learning

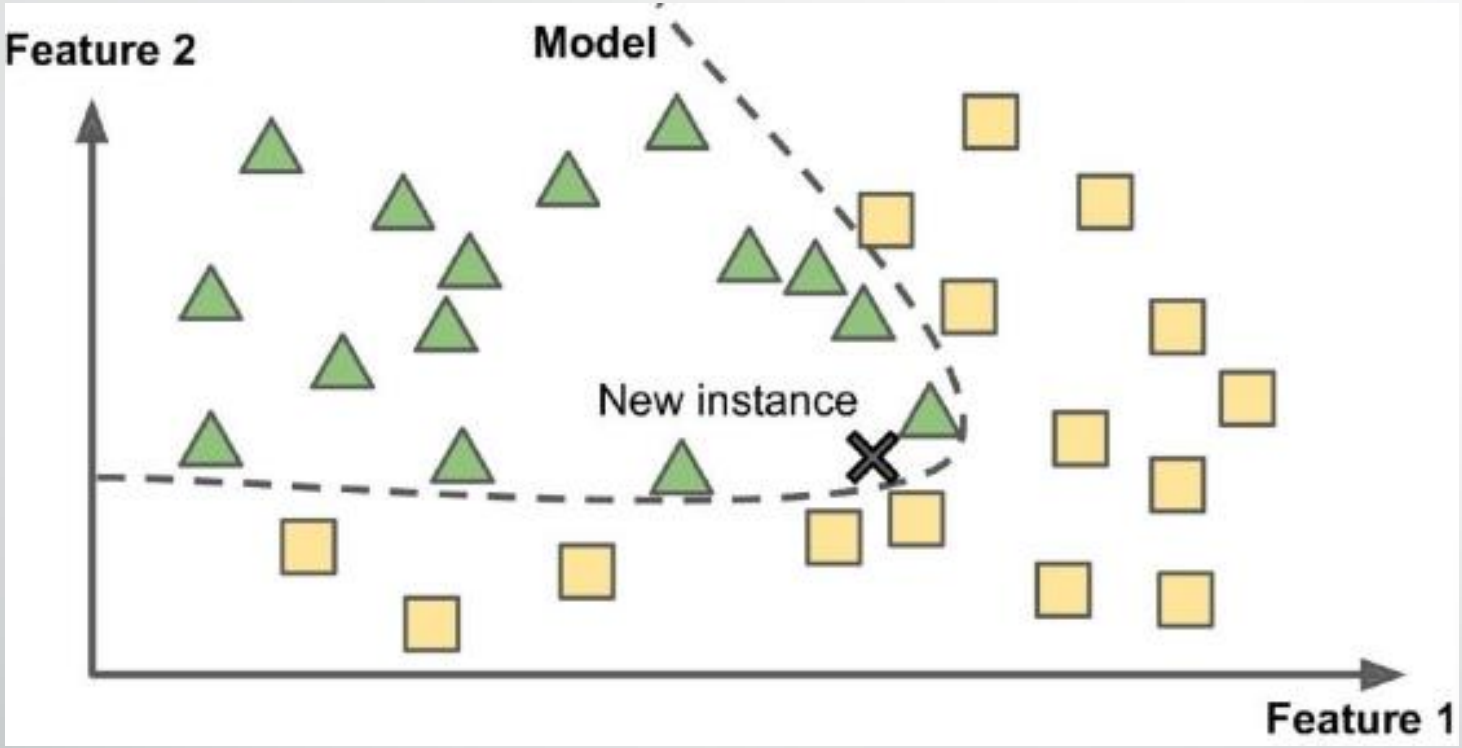


- Fig 1. Instance based learning (Mbogho, 2019)

2.2 Concepts (cont'd)

- In model-based learning, a model is built from the training examples.
- The model is then used to make predictions about new instances.
- Fig 2 demonstrates this concept well; based on the training instances the regions marked in green make up the model.
- The new instance marked 'X' can now be used by the model to make predictions.

Model Based Learning



- Fig 2. Model based learning (Mbogho, 2019)

2.2 Concepts (cont'd)

- Generalization - When a machine learning system can correctly predict the class or value of new instances, we say that it has generalised well.
- It is well and good for training examples to be correctly classified, or for correct values to be computed for them. However, the purpose of training is to generalize, i.e. to make correct predictions for new data that were not in the training set.
- To complete the new concepts learnt in this section bear in mind the triple trade off concept posited by Dietterich (2003).

2.2 Concepts (cont'd)

- Alpydin (2010) describes these factors as follows:
- "In all learning algorithms that are trained from example data, there is a trade-off between three factors:
  - the complexity of the hypothesis we fit to data, namely, the capacity of the hypothesis class,
  - the amount of training data, and
  - the generalization error on new examples. "
- When the amount of training data is increased it is natural to expect the generalization error to decrease; thus increasing the training data leads to decreasing the generalization error.
- Increasing the complexity of the model also leads to an increase in the generalization error; this can be ameliorated by increasing the amount of training data but then again, up to a point.

Part 3

Overfitting

Overfitting

- An overfit model is one that has learnt the training data so well that it underperforms when applied to unseen data.
- This may be due to noise in the training data, which should not be captured by the model.
- Let us consider the design of a spam mail filter. Ideally spam emails will contain some normal language as well as the sort of language that is typical of spam.
- If our model captures every nuance of the spam training data, when it is put into use it will direct some legitimate messages into the spam folder. It will may also let through some spam messages that are not too strongly like the training data.
- Consider the working of such a filter:

Overfitting (cont'd)

- The training set of such a model will consist of several emails; some of these emails are spam while others are ham.
- The model will use the training set to learn how to classify an email as spam or ham (non-spam).
- There will be tests on the email that will be perfected in order to correctly differentiate spam from ham. For purposes of demonstration let us say that there are 3 tests that are performed on the email.
- By determining a threshold based on the training set a rule can be made to separate the spam from ham.
- This can be demonstrated in table 1.

Overfitting (cont'd):

Email	A1	A2	A3	Spam (0 – No, 1 – Yes)	Equation (threshold) $2(A1)^2 + 3A2 + A3$
1	0	0	0	0	0
2	1	1	1	1	6
3	1	0	1	1	3
4	1	1	0	1	5
5	0	1	0	0	3

- Table 1. Rules to determine spam

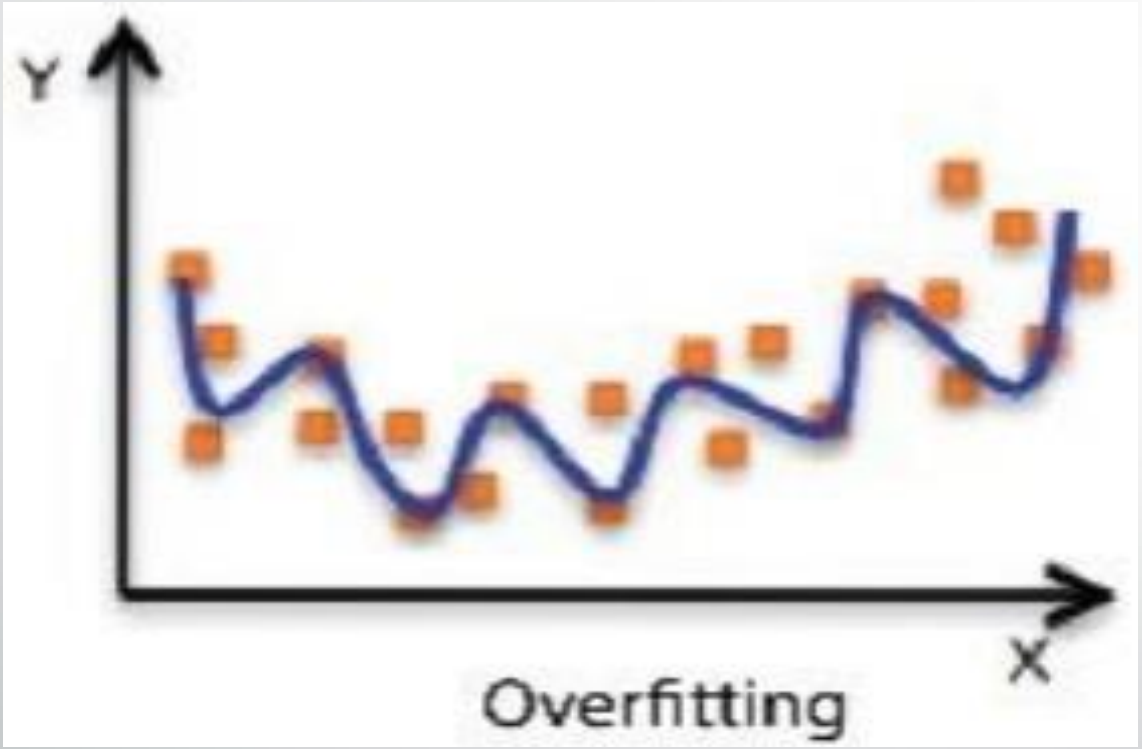
Overfitting (cont'd)

- Table 1 shows a training set consisting of 5 emails.
- There are 3 tests that are applied to the emails. The fifth column indicates whether the emails are spam (1) or ham (0).
- A thresholding function $2(A1)^2 + 3A2 + A3$ is used to determine whether future emails are to be classified as either ham or spam.
- For easier understanding supposing that all emails whose value for this function exceed 4 are to be treated as spam; this means that all emails whose value are 4 or less will be treated as ham will not be sent to the spam folder.
- If the model learns the training data too well then it will underperform as described earlier.

Overfitting (cont'd)

- Fig 3 exemplifies overfitting.
- The blue line represents a model like our spam filter model.
- As can be observed, the model has learnt the training data too well (i.e. has overfit the training data), and is therefore probably going to make wrong predictions for new data.

Overfitting (cont'd):



- Fig 3. Overfitting a model (Mbogho, 2019)

Overfitting (cont'd)

- A good example is how students often study a course. Many students attempt to cram the notes given in the lecture so that they can pass the exam.
- But when asked to apply what they know to a problem that was not presented in class, they don't know where to begin!
- They have overfit the training data (class notes).
- Another example is when students are preparing for exams. In certain institutions some lecturers may repeat past exam questions and students know this.
- Thus students spend a considerable amount of time going through past papers and comparing their answers with the provided solutions.
- However, when they get to the exam they find there is no past paper question (trouble!) and they panic and do not do well in the exam; generalization failed.

Underfitting

- Underfitting is the opposite of overfitting.
- In this case, the model does only shallow learning and fails to capture important aspects of the training data.
- When applied to new data, the underfit model will also do poorly. But it will do poorly even on the training data, despite being exposed to it.
- Continuing the student analogy, while the overfitting student learns the lecture material too well, the underfitting student pays little attention. He can neither apply what was taught to new situations nor remember much of it in the first place.

Underfitting (cont'd)

- Fig 4 shows underfitting in a model.
- The choice of a straight line is clearly not the correct one to use based on the training set.
- Fig 5 shows a curve chosen rather than a line, which is more representative, despite the fact there may be some noise in the data.

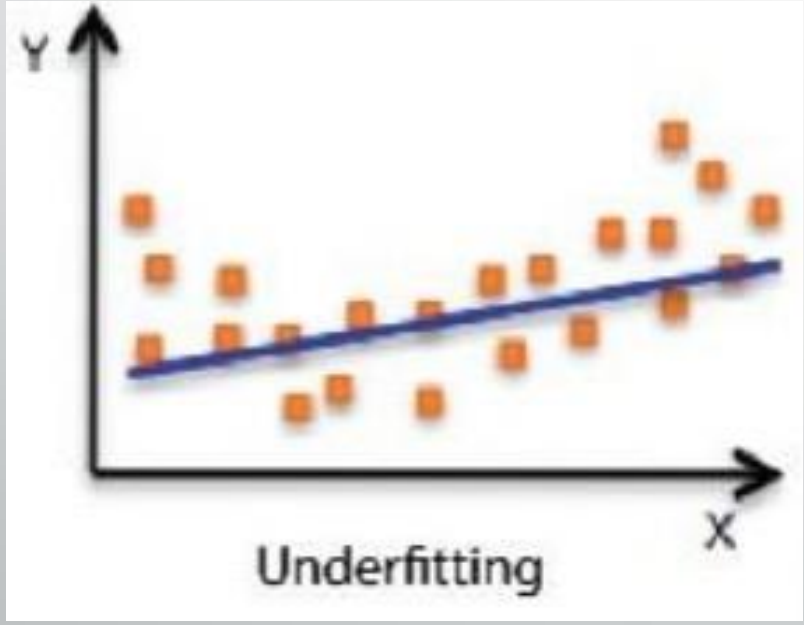


Fig 4. Underfitting (Mbogho, 2019)

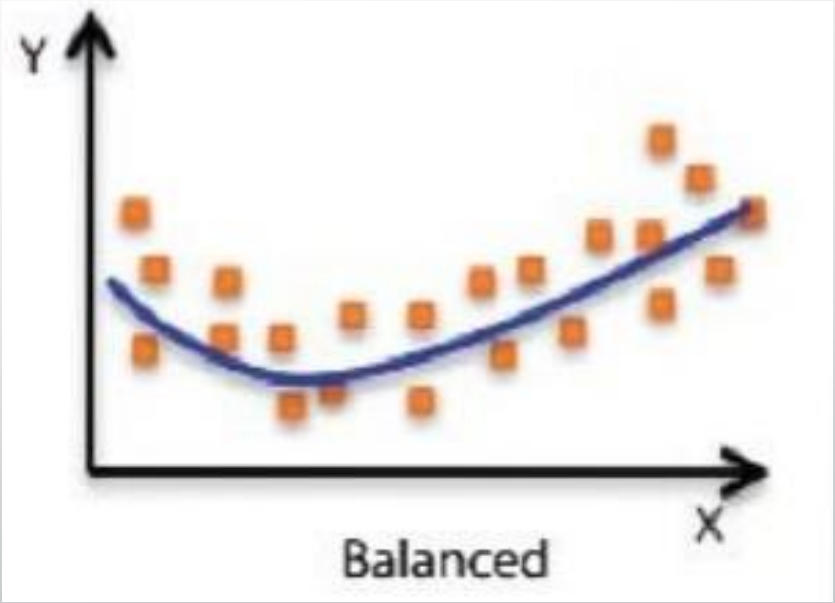


Fig 5. Balanced (Mbogho, 2019)

Validation & Testing

- A validation set is used to adjust certain training parameters as the training is going on, such as the number of epochs (number of passes of the training dataset that the algorithm has completed).
- For example, to avoid overfitting, we may want to check from time to time how the model performs on unseen data.
- We set aside some of the labelled data so that it is not used in training.
- From time to time we apply to the validation set the model trained so far
- When the performance starts to go down, we know that we have attained enough training.

Validation & Testing (cont'd)

- A test set is similar to a validation set in that it will not have been used in training; it is set aside.
- It differs from the validation set in that it is used only once at the end to assess the performance of the final model.
- Applying a model to the test set and obtaining its performance can be useful for comparing it with other models and then choosing the best performing model.

Part 4

Linear Regression

4.1 Introduction

- Regression simply put, is a measure of how one variable relates to another one (the former variable might relate to more than one variable).
- This means that the variable under investigation, let's call it y, is being investigated to determine its relationship with one or more other variables say, x1, x2, x3 and so on.
- Y is referred to as a dependent variable while x is referred to as an independent variable.
- This means that the value of y depends on the value of x, whether x is a single variable or more than one variable like the case of x1, x2, x3 and so on.

4.1 Introduction (cont'd)

- The value y is called an outcome variable while the value x is referred to as a predictor variable.
- There are 6 different linear regression types that can be used and these are: simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression, and discriminant analysis. (Mining, 2019).
- These different types shall be discussed later on in this lesson. However, based on the description provided, below we shall focus on simple and multiple linear regression.
- For both types of linear regression we wish to use an equation to define the relationship between the predictor/independent variable(s) (x in our case) and the outcome/dependent variable (in our case y)

4.1 Introduction (cont'd)

- In all linear regression modelling there is only one outcome (y), while there may be one or more input/predictor/independent variables (x).
- Mathematically when there is only one independent variable, the relationship is represented using the formula for a linear equation, i.e.

$$y = a + bx ,$$

- Where a is the value of the intercept and b is the slope of the line (coefficient).

4.1 Introduction (cont'd)

- In the case where there is multiple linear regression, meaning there are more than one independent variables then the equation is of the form:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_n x_n$$

Where b_0 is a constant, and $b_1 \ldots b_n$ are slopes for the independent variables

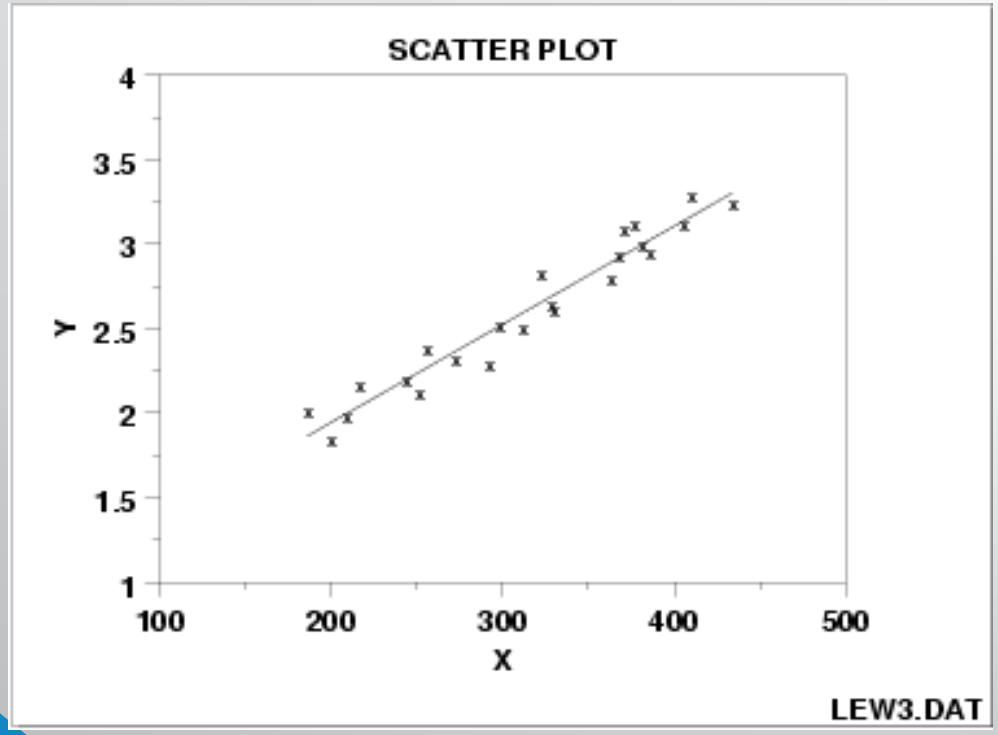
$x_1 \ldots x_n$

- The aim of the model is to find a best fit for all the values of b given x, such that the value of y will be as close to the actual outcome as possible. In other words to reduce the error between the predicted outcome and the actual outcome (residual).
- Just as a reminder this is a supervised learning model and there will be a training data set to train the model.

4.1 Introduction (cont'd)

- Just as with most mathematical theorem and formulae there are assumptions that are made when dealing with linear regression. These are:
- Linear relationship – it is assumed that there is a linear relationship between y and the independent variable(s). The rule of thumb is to use not less than 20 cases for each independent variable x.
- Fig 6 demonstrates in (a) a linear relationship an in (b) an non-linear relationship.

4.1 Introduction (cont'd)



- Fig 6(a) Linear relationship (itl.nist.gov)

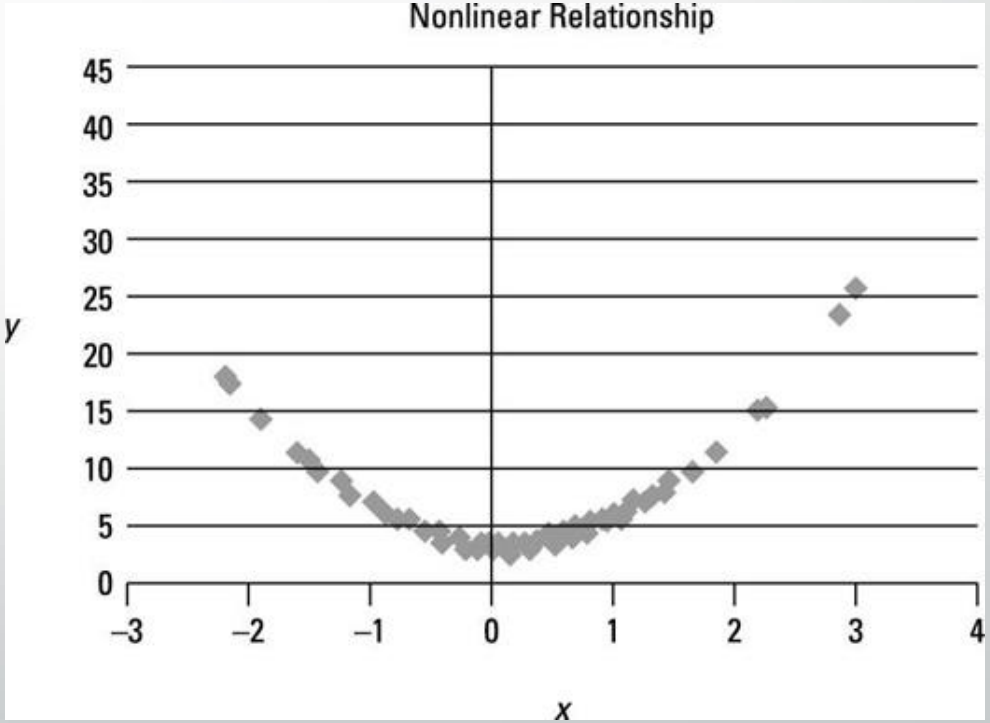


Fig 6(b) Non linear relationship (dummies.com)

4.1 Introduction (cont'd):

- Normality – both x and y should be normally distributed.
- Minimal or no multicollinearity – there is no exact relationship between the variables.
- No autocorrelation – the errors or residuals should be independent of each other.
- Homoscedasticity – the value of the residuals should be the same across the line for all values of the independent variable (x).

4.1 Introduction (cont'd):

- The next question that arises then is how to evaluate the performance of the model? There are several tests that can be done and these are mentioned here.
- The learner is urged to do some further reading on the tests themselves:
  - R squared test
  - Adjusted R squared test
  - Mean squared error (MSE)
  - Root MSE

4.2 Other Types of Linear Regression

- Ordinal regression – this form of regression involves a dependent variable and one or more independent variables. The difference is in the nature of the variables; the dependent variable should be ordinal, meaning it is formed on the basis of a hierarchical scale, e.g. low, medium, high. The independent variables should be either dichotomous (having only two possible outcomes, for example true/false) or nominal (same as ordinal but no order/hierarchy).
- Multinomial regression – similar to ordinal; the difference is that the dependent variable must be always nominal while the independent variables may be either dichotomous or interval (or ratio).
- Discriminant analysis – this form of regression has the dependent variable being nominal while the independent may be either ratio or interval.

4.3 Working of the model

- As described earlier regression is a form of supervised learning. This implies that there has to be some training data available consisting of features (inputs) and labels (outputs).
- The idea is that you need to have a theory as to how these features are related to the outputs.
- Using this theory you can train the model and try to get the best fit for your model. Once the model has been trained you can then introduce new data and see what kind of output you will get from the trained model.
- Based on the model one can input different values of the independent variable(s) and develop theories on the relationship between the independent variables and the dependent variable. Further one can also introduce new features and see their effect on the outcome.
- Regression models are very useful in predicting what will happen (y) based on a variety of features (x).

4.4 Application areas

- Linear regression can be used in many areas of every day life. A few areas include:
- Remember the height of the covid pandemic in 2020-1? Scientists came up with all different models predicting where it would spread to first, which populations were most vulnerable, etc?
- In supermarkets management can use regression analysis to determine the relationship between sales of different products, for example profitability goes up when certain items are placed next to each other, etc
- In the field of education factors can be determined to see how the performance of students can be improved.
- In agriculture farmers can use linear regression to determine the factors that can lead to a good yield in crops, and how much of such to use.

Part 5

Decision Trees

5.1 Introduction

- I am currently in Nairobi, Kenya. I would like to make a trip this holiday to the beautiful coastal city of Mombasa. What do I need to consider in making the decision as to what mode of transport to use?
- I can either fly, go by bus, or take a train. If I choose to fly I need to have at least KShs 20000 for a return ticket. If I travel by bus then it will only cost me KShs 4000 for a return ticket. However, should I use the train then it will cost me KShs 2000 for a return economy class ticket.
- Further should I choose to fly I have to make a decision between different airlines and which airport they fly from. Should I choose to travel by bus then I have to find a bus company that can pick me along the way since I live far from the central business district. If I choose to travel by train I will have to use a taxi to the train station.

5.1 Introduction (cont'd)

- So which mode of transport should I use?
- Depending on the individual some factors will be more important than others, for example, budget might be a key consideration; for others, convenience would be the major consideration; while yet for others they may have acrophobia or aerophobia, the fear of heights and therefore flying is out.
- This is an example of a decision tree that is running through your mind as you think about the final choice you will make.

5.2 What are Decision Trees?

- A decision tree is a tree like structure like a flowchart that starts with a desired outcome, and flows down considering each option (decision) and what it takes to fulfill that option, until all the options have been exhausted and an outcome determined.
- The advantage of a decision tree is that it is highly adaptable meaning that decisions can be added or removed without affecting the overall structure of the tree.
- Due to its adaptability and easy to understand structure decision trees are used everywhere from mathematics to data science and to machine learning.
- In machine learning the decision tree is designed using different available algorithms. Let us examine this in greater detail.

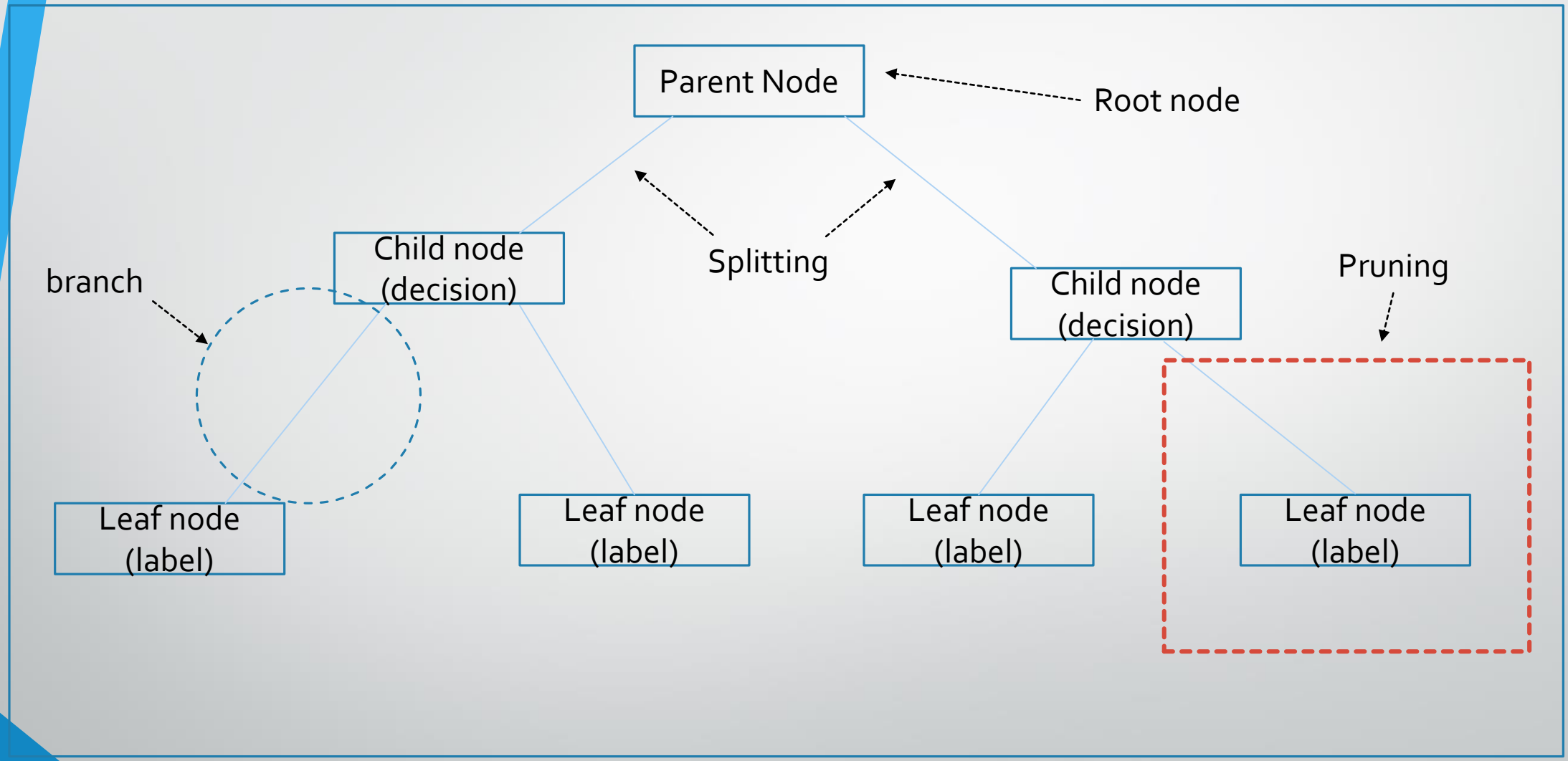
5.3 Decision Trees in Machine Learning

- Decision trees show us most likely what will happen given a particular set of circumstances. Looking at the example in our introduction, if I am on a budget then the choice of flying is out. If I don't like the option of going to the CBD then the train is my only option. You see?
- Each branch of the tree represents a specific outcome. In order to reach this outcome a series of decisions have to be made from the initial decision to the final one (outcome).
- The final outcome is a label and is referred to as a leaf node. All the other decisions in between are also referred to as nodes but it is only the final one in the branch that is referred to as a leaf node.
- The process of moving through the branch(es) is done using different classifiers and these will take us all the way down the branch to the final outcome. Consider a factor like budget described in our hypothetical situation; what would refer to it as in machine learning?

5.3 Decision Trees in Machine Learning (cont'd)

- The classifiers are algorithms that are used to assign labels to the different inputs given to the learner.
- Decision trees as can be inferred use supervised learning in training the learner.
- They use different algorithms such as random forest, ID3, CART (classification and regression tree) and gradient boosting. The detailed working of these algorithms are not discussed here due to time and scope; however, the reader is encouraged to do further reading from the references provided at the end of the lesson.
- Let us examine the parts of the decision tree and their roles.

5.4 Decision Tree Parts and Terms



- Fig 6. Parts and terms of a decision tree

5.4 Decision Tree Parts and Terms

- Fig 6 shows all the parts and terms associated with decision trees. Let us describe them more in detail.
- Root node - this represents the primary decision that needs to be made for this decision tree; for example in our case we can call it 'transport mode' since ultimately the decision to be made is what transport mode to use.
- Splitting - this refers to the act of splitting a node into further sub-nodes. This is demonstrated in our figure where the root node splits into two sub nodes which also split further down the tree.
- Decision node - this refers to a decision that needs to be made in pursuit of a final outcome. In fig 6. the first decision that needs to be made is 'transport mode'. In order to achieve this decision further sub decisions need to be made, down the tree until a final outcome is achieved at the leaf node (which is also a decision node).

5.4 Decision Tree Parts and Terms (cont'd)

- Leaf node – this is a final node, i.e., it is a node that does not branch out to other nodes. It is a final outcome (decision) as demonstrated in fig 6.
- Pruning – just as an ordinary tree would need pruning to make it healthier the same applies to decision trees. This act involves removing a decision from a decision tree to make it more practical or effective. This could be due to a variety of reasons.
- Branch – this is a subsection of the tree as depicted in fig 6. Every branch comes from the decision node above it.
- Parent/child node – every decision node that has a branch below it is the parent to the node below it in the tree. Every decision node from the branch is referred to as a child node.

5.4 Decision Tree Working

- How do you actually implement the algorithm in machine learning?
- Once you have determined your primary decision the algorithm is used by inputting the labeled data into it. All the nodes and branches will now be filled out automatically based on the provided labels.
- A good practical example of how this is done using the ID3 algorithm can be found in Mitchell (1997).

5.5 Critical Success Factors

- Mitchell (1997) points out some very important issues to do with decision tree learning:
- Avoid overfitting data – overfitting has been described earlier in this lesson.
- Allow for continuous valued attributes by for example allowing the algorithm to create/ assign a Boolean attribute that will split them along a certain threshold determined by you.
- Find a way to select attributes based on a measure other than information gain. This is because it is not desirable to separate training examples into too many subsets. Information gain "measures how well a given attribute separates the training examples according to their target classification."
- When some attribute values are missing it is appropriate to estimate them from other training examples.

5.6 Where to use Decision Tree Learning

- Mitchell (1997) also describes the ideal situation where decision tree learning should be used:
- Where instances are represented in pairs, e.g. black and white, road/air/rail (like our early example), and so on.
- When the target function has distinct value of output.
- Where contrasting descriptions may be required.
- When training data may have mistakes
- When training data contains missing feature values.

Summary

- In instance-based learning, a new instance which we wish to classify is compared to the instances in the training data.
- In model-based learning, a model is built from the training examples.
- An overfit model is one that has learnt the training data so well that it underperforms when applied to unseen data.
- In all linear regression modelling there is only one outcome (y), while there may be one or more input/predictor/independent variables (x).
- A decision tree is a tree like structure like a flowchart that starts with a desired outcome, and flows down considering each option (decision) and what it takes to fulfill that option, until all the options have been exhausted and an outcome determined.

References

- Alpaydin, E. (2010). *Introduction to machine learning*. MIT Press.
- Anderson, A., By: Alan Anderson and, and, A. A., 03-26-2016, U., From The Book: Business Statistics For Dummies, About the book author: Alan Anderson, author:, A. the book, & Anderson, A. (n.d.). *Use scatter plots to identify a linear relationship in simple regression analysis*. dummies. Retrieved April 13, 2022, from https://www.dummies.com/article/business-careers-money/business/accounting/calculation-analysis/use-scatter-plots-to-identify-a-linear-relationship-in-simple-regression-analysis-145935/
- Deepanshi, D. (2021, May 25). *Linear regression: Introduction to linear regression for data science*. Analytics Vidhya. Retrieved April 13, 2022, from https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/#:~:text=In%20the%20most%20simple%20words,the%20dependent%20and%20independent%20variable.
- *Engineering Statistics Handbook*. 1.3.3.26.2. scatter plot: Strong linear (positive correlation) relationship. (n.d.). Retrieved April 13, 2022, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda33q2.htm
- Mbogho, A. W. (2019). *Introduction 2. Machine Learning*.
- Mining, E. (2019). *Machine Learning for Beginners: A Complete and Phased Beginner's Guide to Learning and Understanding Machine Learning and Artificial Intelligence* . Independent.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.