Machine Learning

Lesson 4

Logistic Regression & Support Vector Machine (SVM)

Lecturer: Dr. Msagha J Mbogholi, PhD

Flashback from Lesson 3

- Density estimation is done using histogram (histogram density estimation) and kernel density estimation (KDE).
- KNN is a supervised learning algorithm. The nearest neighbor of any new instance is the shortest distance in a straight line (called the Euclidean distance) from it to a classified instance.
- The curse of dimensionality may lead to incorrect classification due to irrelevant attributes being near to the instance due to many dimensions. As attributes increase so does the likelihood of incorrect classification.
- The search methods used for feature reduction include wrapper, filter and embedded methods.
- In collaborative filtering recommendations are made to a person based on other people who are similar to them.

Content

- Introduction
- Logistic Regression
- Support Vector Machine

Part 1

Introduction

Introduction

- In lesson 2 of this course the concept of regression was introduced.
- In that lesson the 6 types of regression were introduced. These were simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression, and discriminant analysis. (Mining, 2019).
- Each of these forms of regression was described with the exception of logistic regression.
- In this lesson we discuss logistic regression, its definition, how it differs from the classical linear regression form, its form, how it is used, where it is used, and some practical applications of logistic regression.
- In the second part of the lesson a discussion on support vector machines will be covered.

Part 2

Logistic Regression

2.1 What is Logistic Regression (LR)?

- In lesson 2 simple and multiple linear regression were discussed. With these forms of regression a predictor variable (in the case of simple linear regression) or predictor variables (in the case of multiple linear regression) were used to predict an outcome variable.
- This type of regression can be summarized in the form

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

where b_0 is the intercept and $b_1 \ldots b_n$ are slopes for the independent variables

$x_1 \ldots x_n$

- Such an equation is good when y is a continuous variable as was described in lesson 2. They can be used to determine diverse issues such as effect of mileage, number of engine services, and age on the value of a car.
- Car in this case is a continuous variable since the value of a car (outcome) will change based on its mileage, number of engine services, and age (predictor variables).
- However, how about in cases where the outcome variable is not a continuous variable?

  This is where logistic regression comes in.

2.1 What is Logistic Regression (LR)?

- Logistic regression like it's sibling linear regression, is applied in supervised learning environments.
- It is the type of regression used in scenarios where the outcome or dependent variable is binary (0 or 1) or discrete categorical (yes/no).
- Mining (2019) rightly describes it as "Logistic regression specifically is used to explain the relationship between one binary dependent variable and one or more nominal, ordinal, interval, or ratio-level independent variables"
- Since the learner is working with labelled data the outcomes in training are already known. However, in this instance the interest is in knowing what the probability is on the dependent variable in terms of a binary probability that will lie between 0 and 1.
- In other words the probability of the outcome occurring (1) or not (0), such as what is the probability of a person aged 30 years or less, with an income of more than $20000, and a college degree defaulting on a loan?
- 0 would mean they won't default (no) and 1 they will default (yes).

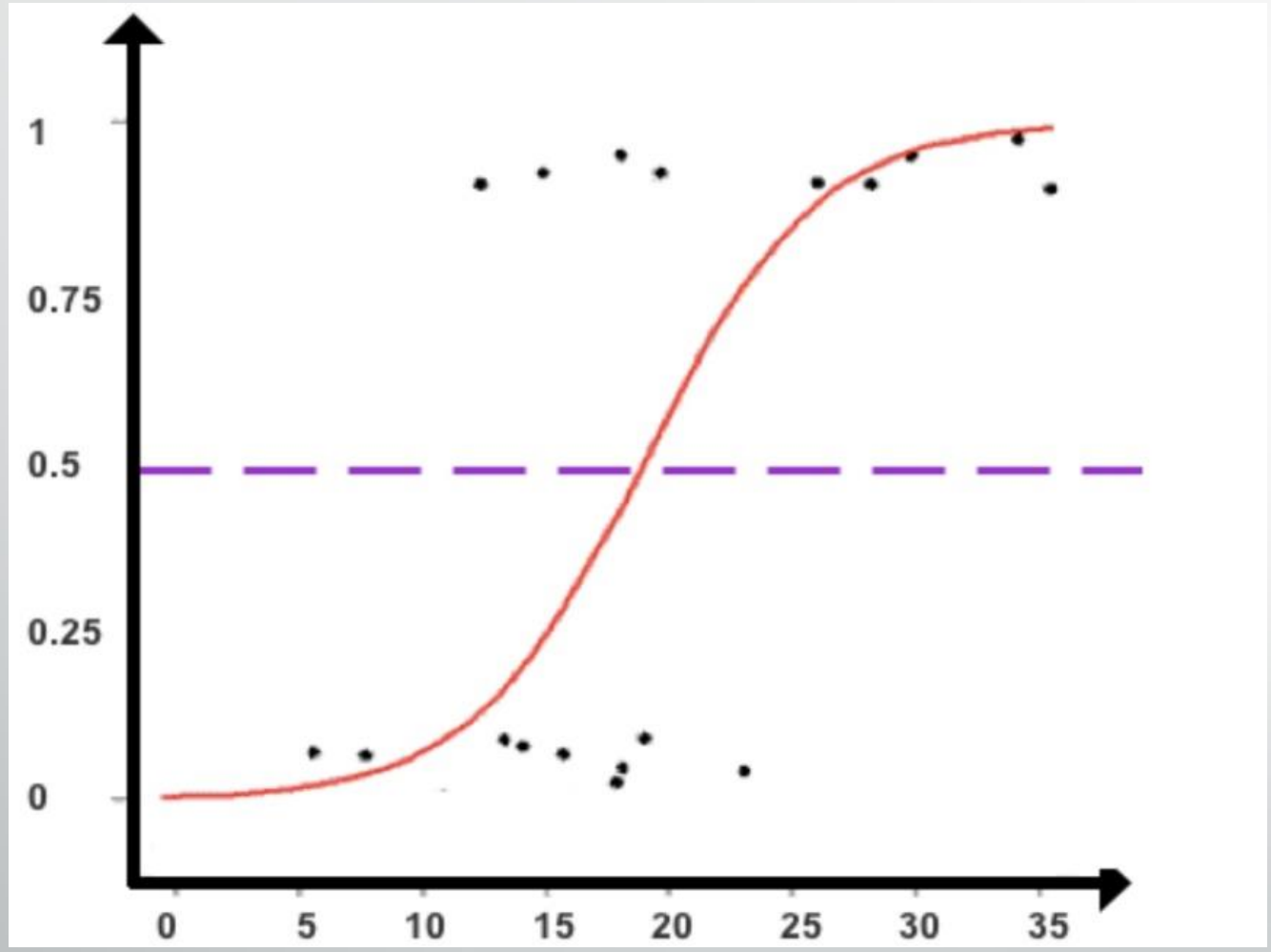
2.1 What is LR? (cont'd)

- LR uses the classical sigmoid function which is of the form:
- $y = \frac{1}{1+e\wedge(-x)}$ where x is the predictor variable (independent)
- And e is the Euler's constant (2.71)
- Using this expression will produce an S shaped curve in the graph.
- The significance of this is that the curve should always start at the bottom left side of the graph and curve up ultimately to the right side (hence the S sigmoid shape).

2.2 How does LR work?

- The procedure used is the same as that of simple linear regression, the only difference being that the dependent variable is now of a binary nature.
- The predictor variables will then be converted to a suitable nature (such as nominal numbers) so that they can input into the learning algorithm.
- The process then continues like in linear regression in terms of training the model as this is a supervised learning scenario.
- Let us demonstrate this using 18 data points.

2.2 How does LR work? (cont'd)



- Fig 1. Logistic regression (Theobald, 2021)

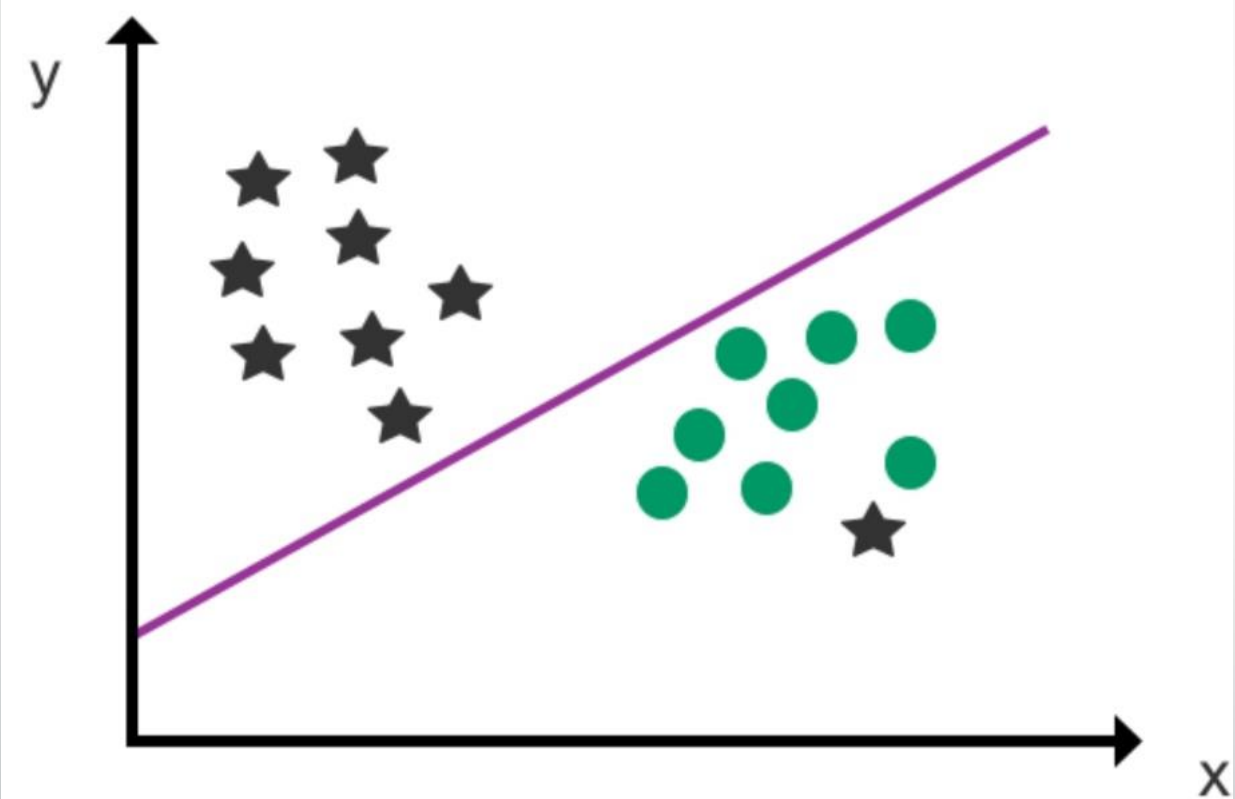
2.2 How does LR work? (cont'd)

- Fig 1 shows the output from a model with 18 data points.
- All the predictor variables are converted by the function into probabilities between 0 and 1 in relation to the outcome (dependent variable) and plotted on the graph.
- In this case the graph is interpreted as probabilities between 0 and 1; the 0 representing no/false and the 1 representing yes/true.
- As can be seen from Fig 1 a cut off point is declared at 0.5; this means that points above the dotted line will be classified as yes/true while points below 0.5 will be classified as no/false. Points that lie at the mid point (0.5) will lack classification (a rare occurrence due to the mathematics involved).

2.2 How does LR work? (cont'd)

- A key difference to note in Fig 1. is the difference in axis labels compared to the classical linear regression model.
- In linear regression the y – axis will have the outcome values based on different inputs (predictor variable); thus for a given value of the predictor variable with their coefficients the correct value of y can be determined (assuming no overfitting or underfitting).
- This is not the case with logistic regression since we are interested in a dichotomous outcome, hence the need to convert the data points based on outcome to probabilities between 0 and 1. The sigmoid function allows us to achieve this.
- Continuing from Fig 1 the outcomes based on the data points can now be classified.
- Fig 2 demonstrates the classification of our 18 data points based on the probabilities calculated in Fig 1.

2.2 How does LR work? (cont'd)

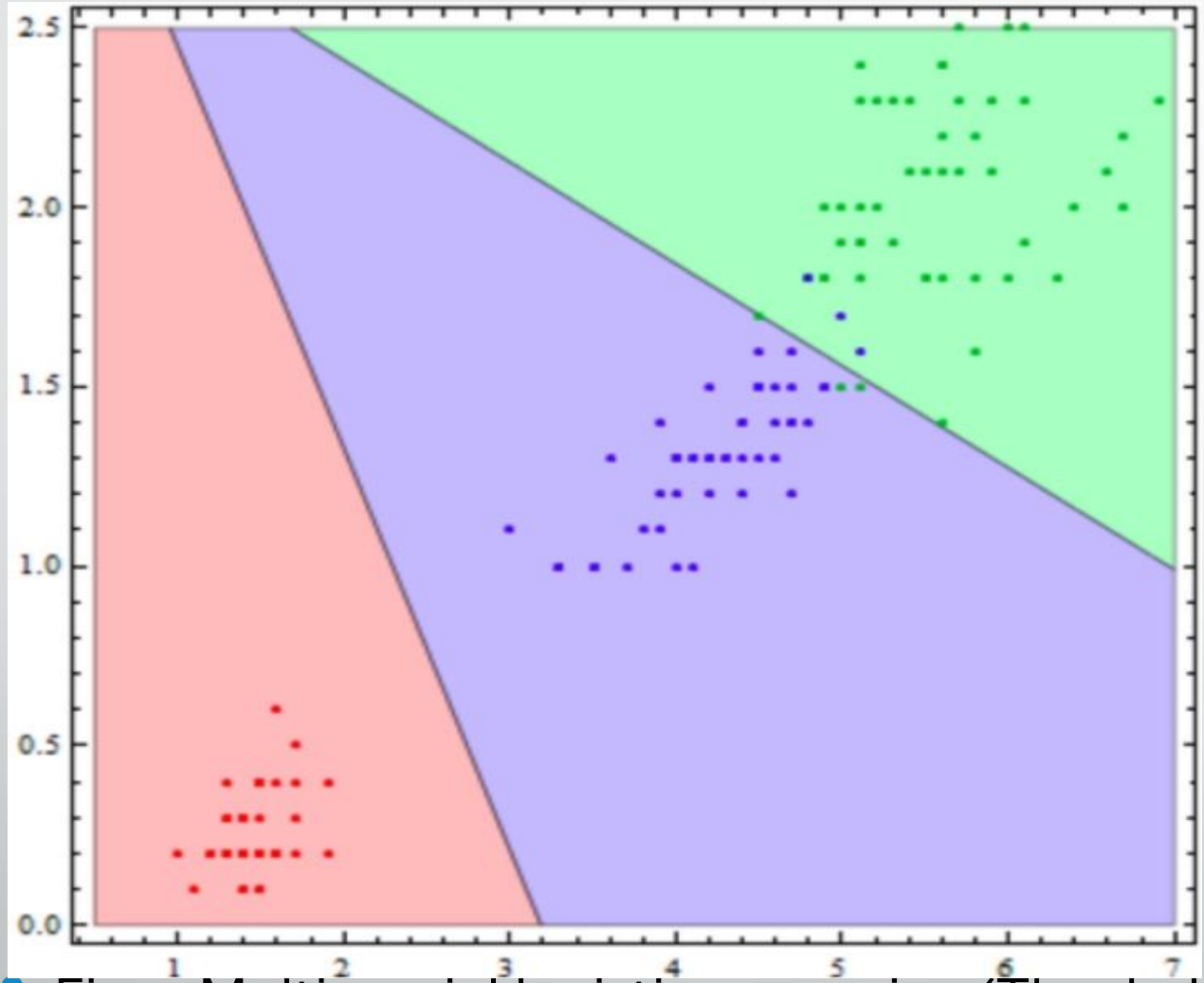


- Fig 2. Logistic regression classification (Theobald, 2021)

2.2 How does LR work? (cont'd)

- Fig 2 shows the different data points classified as either black stars or green circles based on where they were on the sigmoid function graph.
- This shows that an outcome can be classified as being one or the other (black star or green circle) but not both; however, in fig 2 we observe that one of the black stars lies on the side of the green circles? Why is this so?
- These scenarios are easy to interpret as long as one knows how to do the computations based on the mathematics.
- What happens in those cases where there is more than one outcome? This is depicted in multinomial logistic regression.
- Fig 3 shows an example of such a scenario.

2.2 How does LR work? (cont'd)



- Fig 3. Multinomial logistic regression (Theobald, 2021)

2.2 How does LR work? (cont'd)

- Fig 3 shows a scenario where multinomial logistic regression is applied.
- In this case there is more than one output (multiclass classification). The figure also shows that values in the y axis now change since this is no longer a dichotomous scenario.
- Multinomial logistic regression can be used in scenarios where classification is in multiclass, for example when a car can be classified as either a salon(sedan), pickup or station wagon.
- Of course this shows that multiclass classification isn't the ideal to use in logistic regression; it works best in scenarios where outcome is of a binary nature.

2.3 Assumptions of LR

- Training set (dataset) should not have NULL or missing values
- No correlation between independent variables.
- There should be many data points (at least 10) to ensure more accuracy.
- Avoid data that will contain outliers and also datasets that are too large.
- Predictor variables should be linearly related to the log odds (helps to attain a normal distribution).
- For a simple and well explained explanation on log odds please visit https://towardsdatascience.com/https-towardsdatascience-com-what-and-why-of-log-odds-64ba988bf704

2.4 Applications of LR

- Trauma Injury and Injury Severity Score (TRISS) is used to measure the chances of a patient surviving based on the trauma and injuries sustained, as well as their age. The formula is given by:
- $P(s) = \frac{1}{(1+e\wedge(-b))e}$
- Where P(s) is the probability of survival, and $b = b_0 + b_1(RTS) + b_2(ISS) + b_3(\text{Age index})$; while RTS is the Revised Trama Score.
- RTS is the trauma score and ISS is injury severity score.
- Notice the similarity of the formula with that of the sigmoid function?
- Thus using the P(s) it can be determined whether a patient will survive the trauma and injury based on the given parameters.

2.4 Applications of LR (cont'd)

- Similarly LR can be used in determining the probability of a patient developing diabetes or such like illnesses where there are independent variables that can be identified in pursuit of the calculation of such probabilities.
- In the financial sector LR can be used to determine suitability of a candidate to take a loan or mortgage. It can also be used to determine whether one is suitable to get a credit card and what loan limits to apply.
- In politics, LR is used to calculate the probability of a presidential candidate winning, including their odds.
- How about in betting, especially in football games? What are the odds of Chelsea beating Liverpool in the FA cup final later on this year?
- In education LR can be used to investigate the probability of a candidate scoring a first class honours degree based on perhaps high school grade, class attendance, area of residence, and so on (I would love to see someone do this).
- Can you think what the possible outcomes for the four examples above would be?

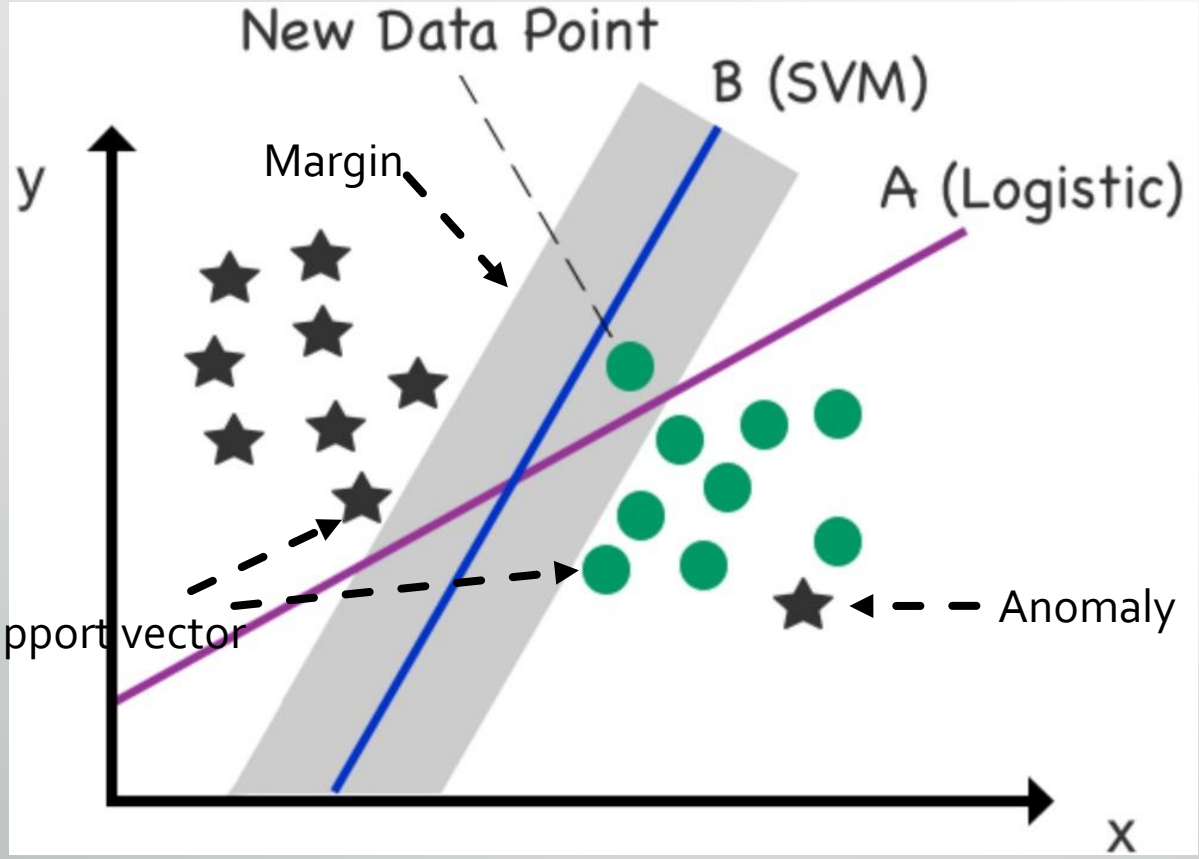
Part 3

Support Vector Machine (SVM)

3.1 Introduction

- In part 2 of this lesson we have discussed the use of logistic regression in determining outcomes based on a dichotomous probability value.
- By so doing the outcome can be classified as being in the form of either…or, for example a dependent variable is a square or a circle. Further by looking at the range of probabilities a cut off line may be drawn such that outcome probabilities exceeding 0.5 can be classified as true while those below 0.5 are classified as false.
- As will be seen in this part it is not enough to just use the boundary to classify the outcome based on probabilities, rather the boundary plays a more significant role in classification especially on borderline points.
- This is the purpose of the support vector machine (SVM); it is a supervised learning algorithm that is used for both classification and regression problems.

3.2 SVM Decision Boundary



- Fig 4. SVM and LR compared (adapted from Theobald, 2021)

3.2 SVM Boundary (cont'd)

- Fig 4 carries over from fig 3 and introduces the concept of how SVM works.
- In the figure there are 17 data points. Using logistic regression a boundary line is drawn that differentiates the two different classes of data.
- The logistic boundary is denoted by the line A.
- A second boundary is introduced called the SVM. The key difference between the logistic boundary and the SVM boundary is that whereas the logistic boundary minimizes its distance from all the data points, the SVM boundary maximizes its distance from the data points. How does it do this?
- It does this by introducing a margin between it and the nearest data point.
- The margin is calculated by taking the distance from the nearest data point and multiplying it by 2. This area is marked in gray and is referred to as the margin.

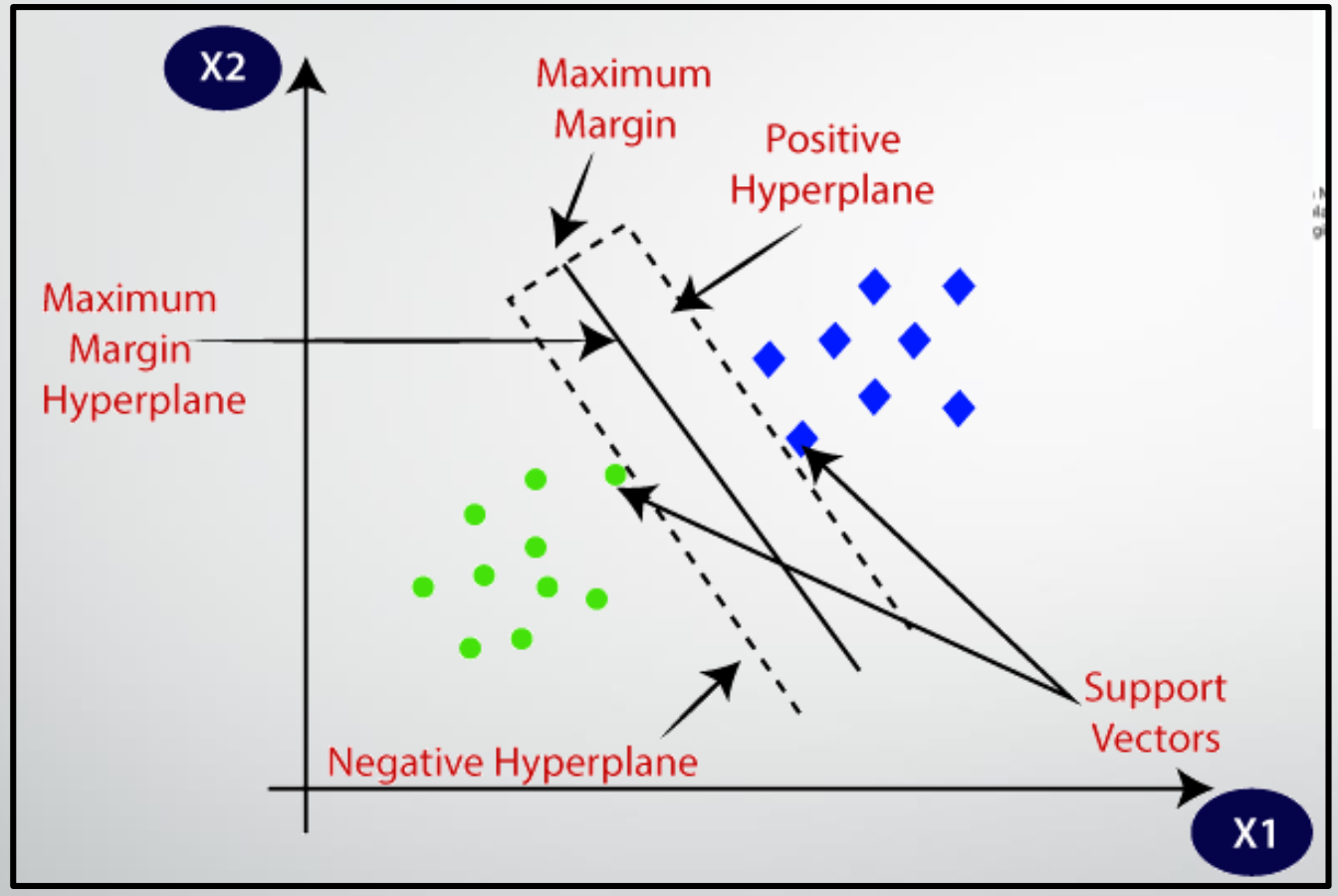
3.2 SVM Boundary (cont'd)

- The margin acts as a support to properly classify data points, especially those found in the boundary.
- Suppose a new data point is introduced as depicted in Fig 4.
- Using the logistic boundary it is clearly erroneously found on the side of the black stars (which it is not).
- However, using the SVM boundary the new point is correctly classified on the green circles side.
- This is due to the support offered by the SVM boundary and margin.
- Some new terms are also introduced in fig 4 which emphasize the power of the SVM boundary.

3.2 Strength of SVM

- The points referred to as support vectors are used to determine the margin used by SVM.
- They are the points nearest to the boundary, and it is from these points that the margin is computed.
- As was seen in fig 2. logistic regression goes out of its way to classify outliers as well as anomalies. This is not the case with SVM.
- In fig 4 the establishment of the SVM boundary does not take into account these anomalies and therefore they have minimal or zero impact on it. This gives SVM a clear distinct advantage over logistic regression.
- This also means that SVM is ideal to use in situations of complexity with small datasets, such as those mentioned above.
- Let us introduce some more terms associated with SVM

3.3 SVM Hyperplanes



- Fig 5. SVM and its hyperplanes (javatpoint.com)

3.3 SVM Hyperplane (cont'd)

- The hyperplane of SVM is simply the boundary that is defined as shown in fig 4 and fig 5; there are theoretically many lines that can be used as boundaries for the two classes. Recall that the hyperplane is defined using the support vectors and the margin is the distance between the support vectors and the boundary multiplied by 2. further the optimal hyperplane is the where the margin is maximum. Consequently the hyperplane is the optimal line for separating the data.
- Fig 5 depicts a scenario of an n-dimensional space where n = 2.
- The solution here is in the form of a linear regression which works well in this hyperplane. However, this is not always the case.
- In fig 5 the positive hyperplane refers to the points that are in the positive domain (in our case the blue squares) while the points in the negative domain (the green circles) are in the negative hyperplane.

3.4 SVM Types

- There are two types of SVM depending on how the data is separated:
- Linear SVM – this is used when the data can be separated clearly by a line, that is to say, it is linearly separable into two distinct groups (classes). Take for example fig 4 and fig 5
- Non-linear SVM – this is used in those cases when the data can not be linearly separated; this is the case in most real world scenarios.
- The choice of which to use lies in how well the data can be linearly separated.

3.5 Hard vs Soft margins

- Suppose you wish to classify two animals whose features are very similar, say a jaguar and a leopard.
- We have information that though at face value they look the same there are some features that separate the two:
- Jaguars are more compact, stockier, have a broader head and more powerful jaws. Further though both have rosette patterns on their skin, the jaguar's has spots in the patterns. Lastly the jaguar's tail is generally shorter than a leopard's. (wildcatsanctuary.org)
- So we have some features that can classify the two big cats; but some are hard to measure. For example how do we measure compactness or stockiness? Do we use their weight, diameter around the waist, or paw for that matter? It's a bit of a challenge.
- When it comes down to it whichever features we choose in classification the difference between these two beautiful creatures is very small.
- Suppose we now classify them using these features and plot them in a 2 dimensional space using SVM

  Fig 6 shows an example of how the data might look like

3.5 Hard vs Soft Margins (cont'd)

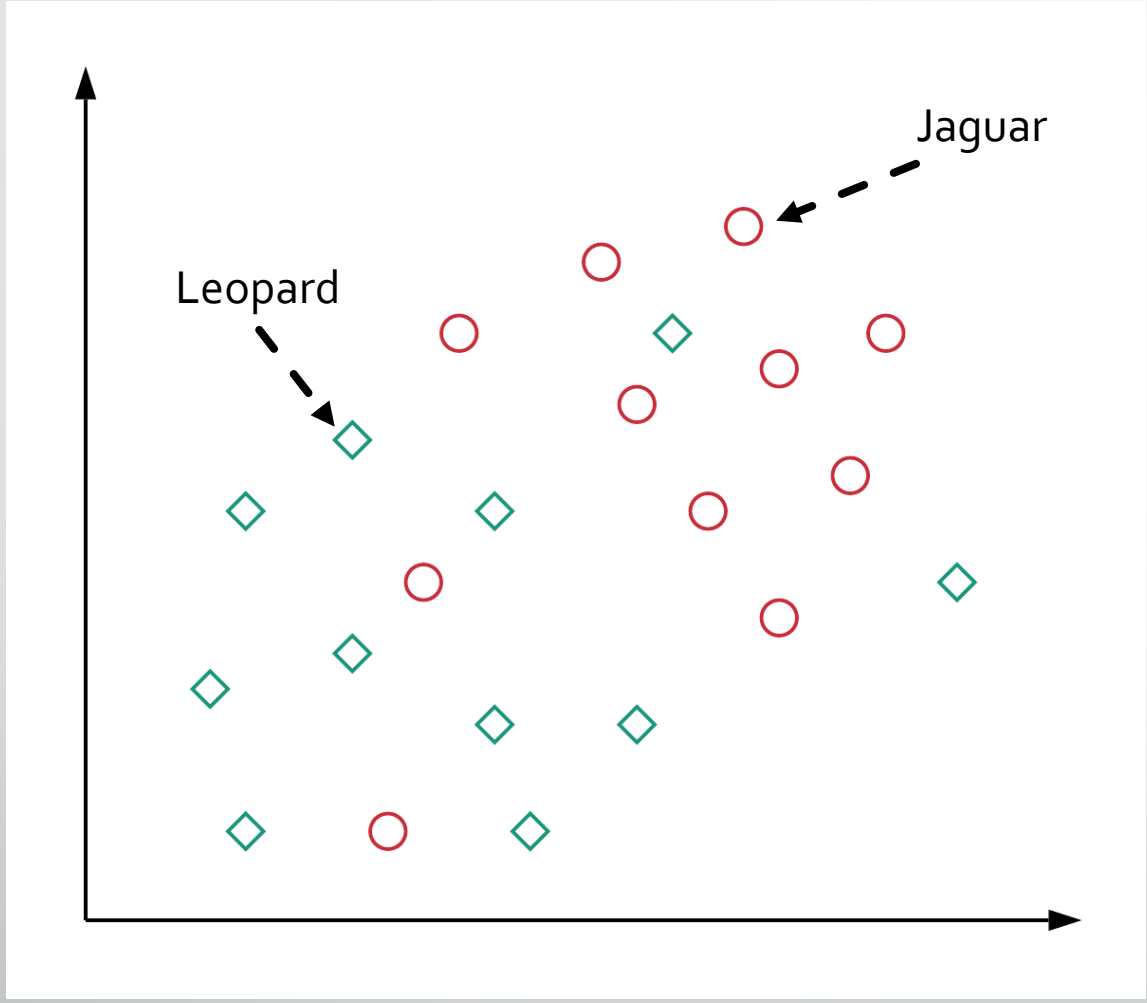
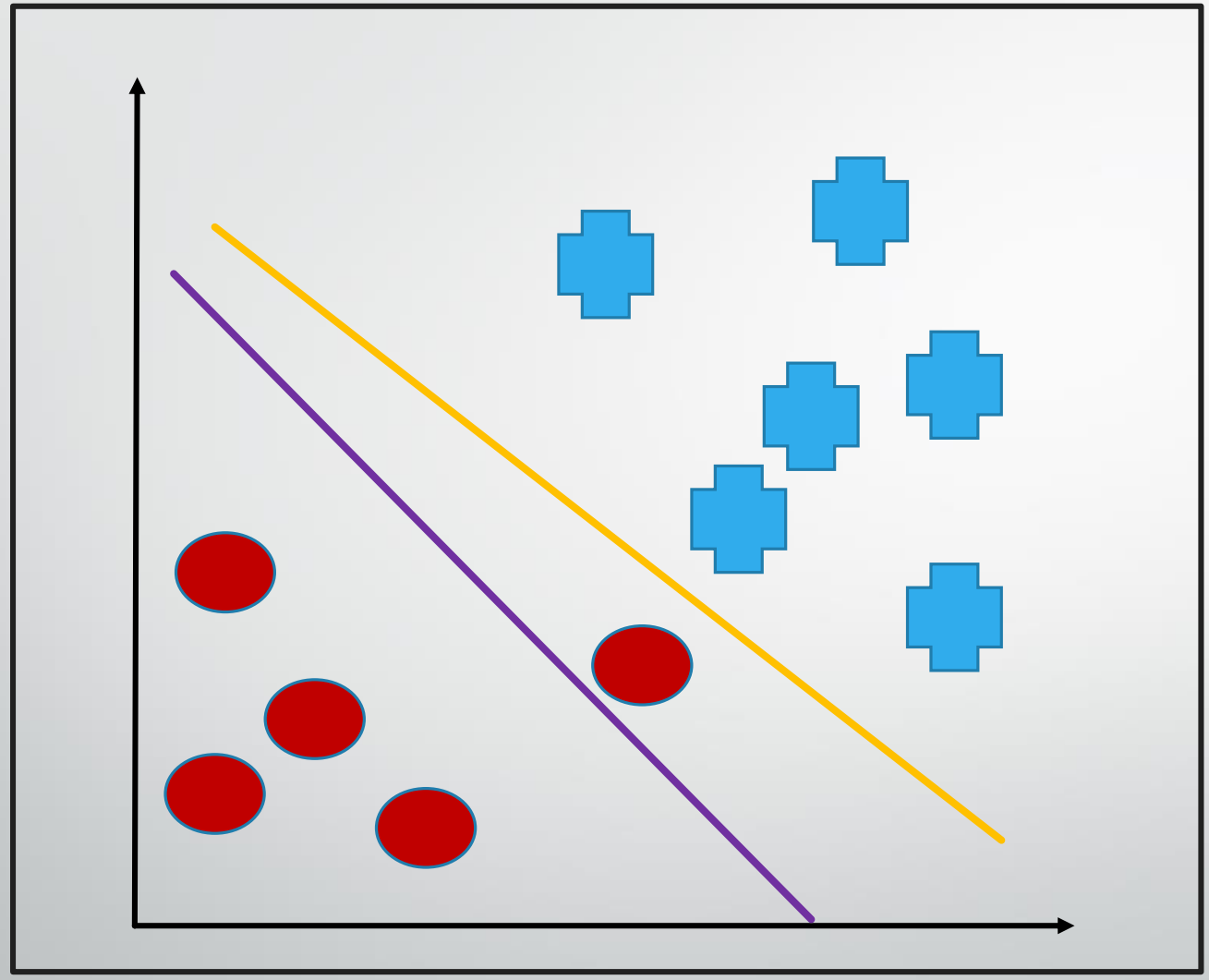


Fig 6. Example classification of leopard and jaguar (adapted from towardsdatascience.com, 2019)

3.5 Hard vs Soft Margins (cont'd)

- Fig 6 shows how the data may appear in the 2 dimensional space. The term we use to refer to this type of data is that it is linearly inseparable.
- The leopard vs jaguar is just an example of a real world example where two entities can not be clearly distinguished due to the features selected to separate the two.
- Now how do we deal with such data? This is where soft margins and hard margins come in.
- It is preferred to have a situation where we don't look for the perfect boundary as this may lead to overfitting; meaning that some of the unseen data will not be properly classified.
- Consider the example in fig 7 on the next slide.

3.5 Hard vs Soft Margins (cont'd)

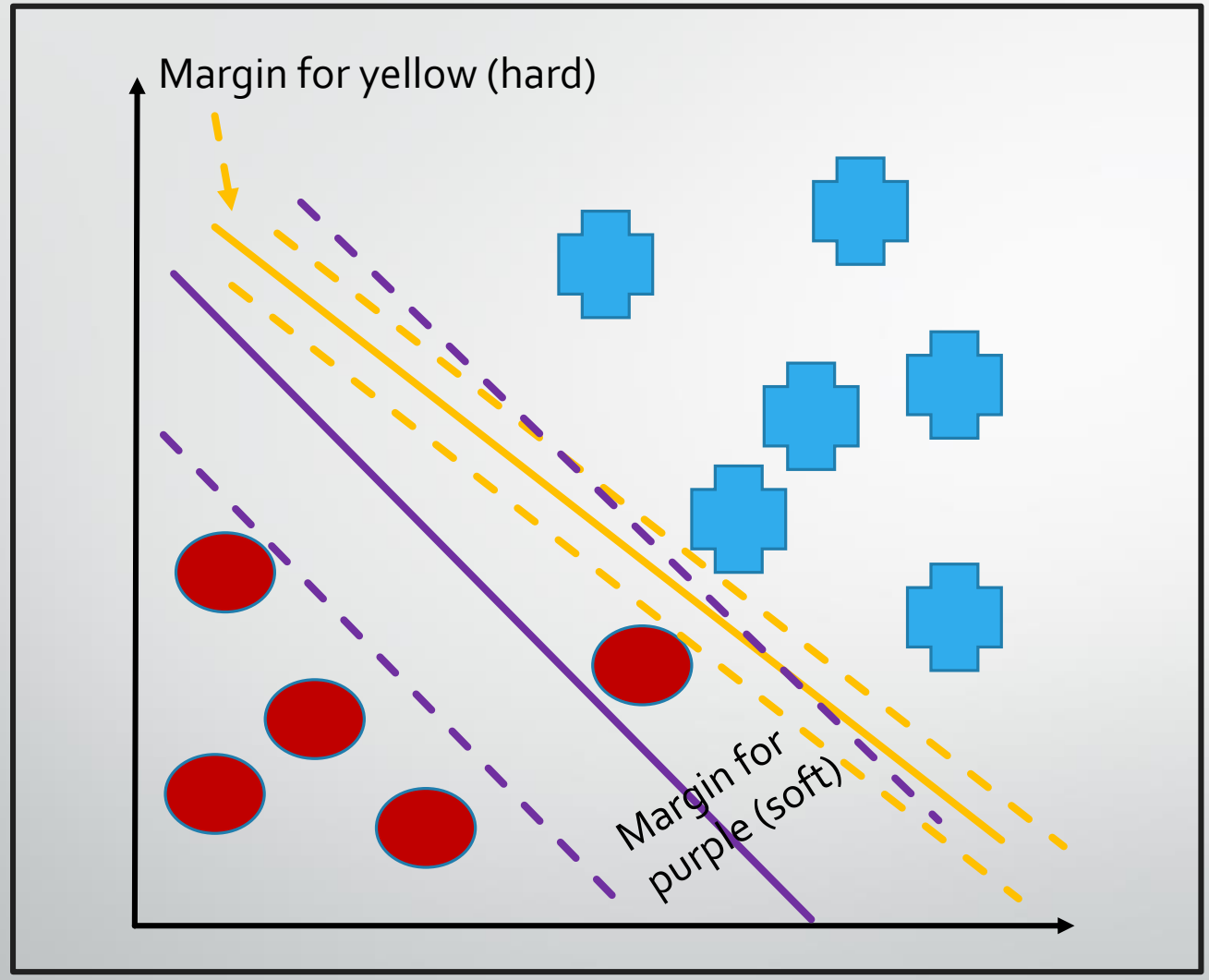


- Fig 7. Decision boundary choices.

3.5 Hard vs Soft Margins (cont'd)

- Consider the two suggested boundaries in fig 7 (the yellow and the purple one).
- Which of the two would you consider as the better boundary decision line to use?
- Choosing the yellow boundary creates a very small margin of error and may lead to overfitting the data.
- Choosing the purple boundary allows for a wider margin and this may allow for a margin of error and therefore fewer mistakes in classification; in proper terms the purple boundary will generalize well on the unseen data.
- Let us introduce the margin for both boundaries in fig 8.

3.5 Hard vs Soft Margins (cont'd)



- Fig 8. Margins for yellow and purple boundaries

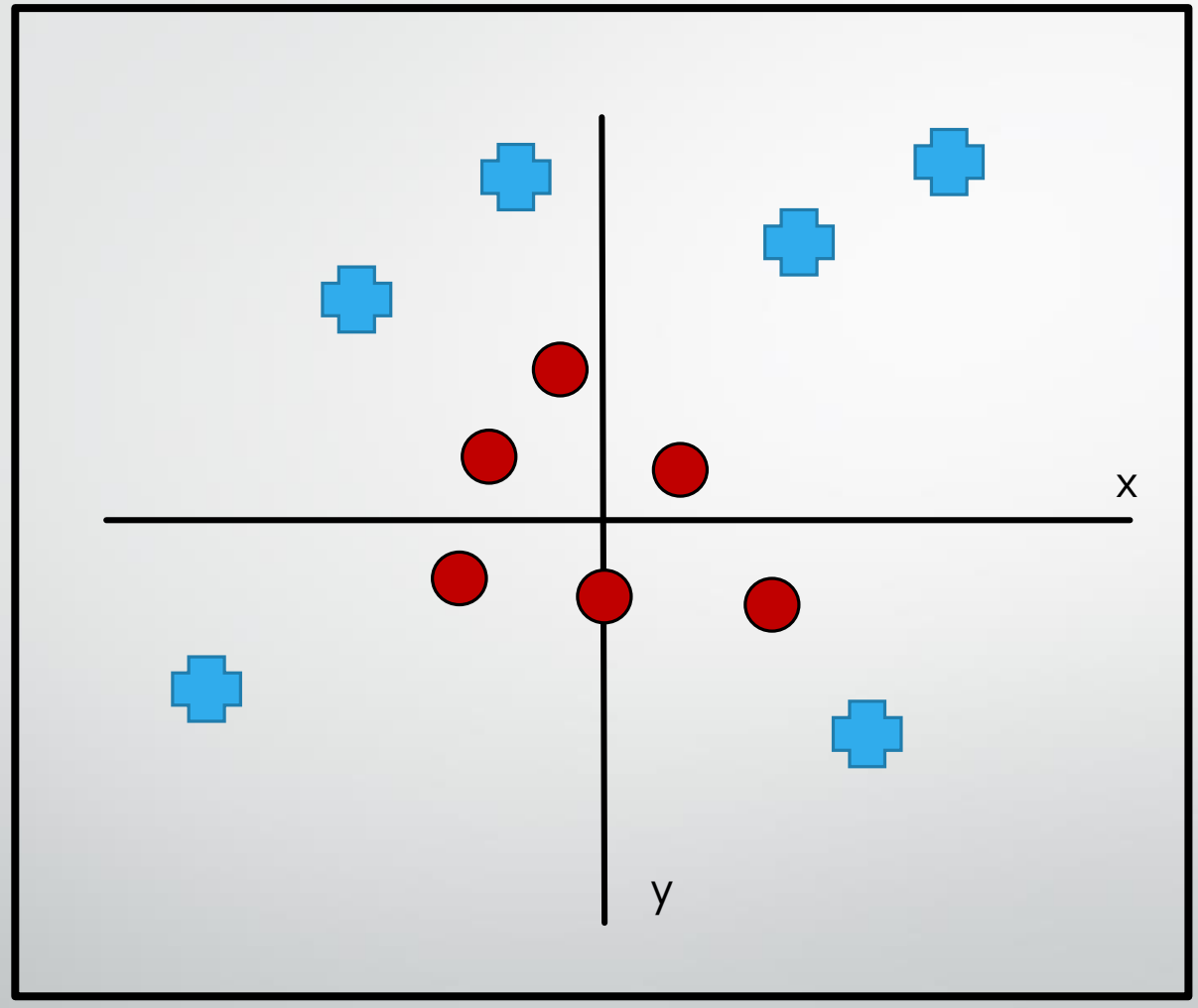
3.5 Hard vs Soft Margins (cont'd)

- From fig 8 it can be clearly seen that the margin for the purple boundary is much wider than the one for the yellow boundary. This is due the positioning of the support vectors in both cases.
- This modification of the boundary is done using a hyperparameter called C.
- As explained earlier it is better to allow for more errors in the training data in order to accommodate unseen data (better generalization) than it is to have small margins which won't accommodate unseen data well (overfitting).
- When we have a wide margin the value of C is low like in the case of the purple margin; conversely when the value of C is high it leads to a narrower margin like in the yellow boundary. The effect on both has been discussed.
- Consequently it is desirable to strike a balance in the value of C; if it's too high then there is the danger of overfitting; if it is too low then it will generalize too much which will lead to too many errors in unseen data.

3.6 Non-linear SVM Example

- In the examples examined so far the data has been linearly fitted in a two dimensional space.
- What happens in the case whereby the data is non-linear?
- In such a case there will be need to introduce a third dimension in order to accommodate such data. Remember your high school mathematics? This will be very useful in understanding this example.
- Consider the example in fig 9 on the next slide.

3.6 Non-linear SVM Example

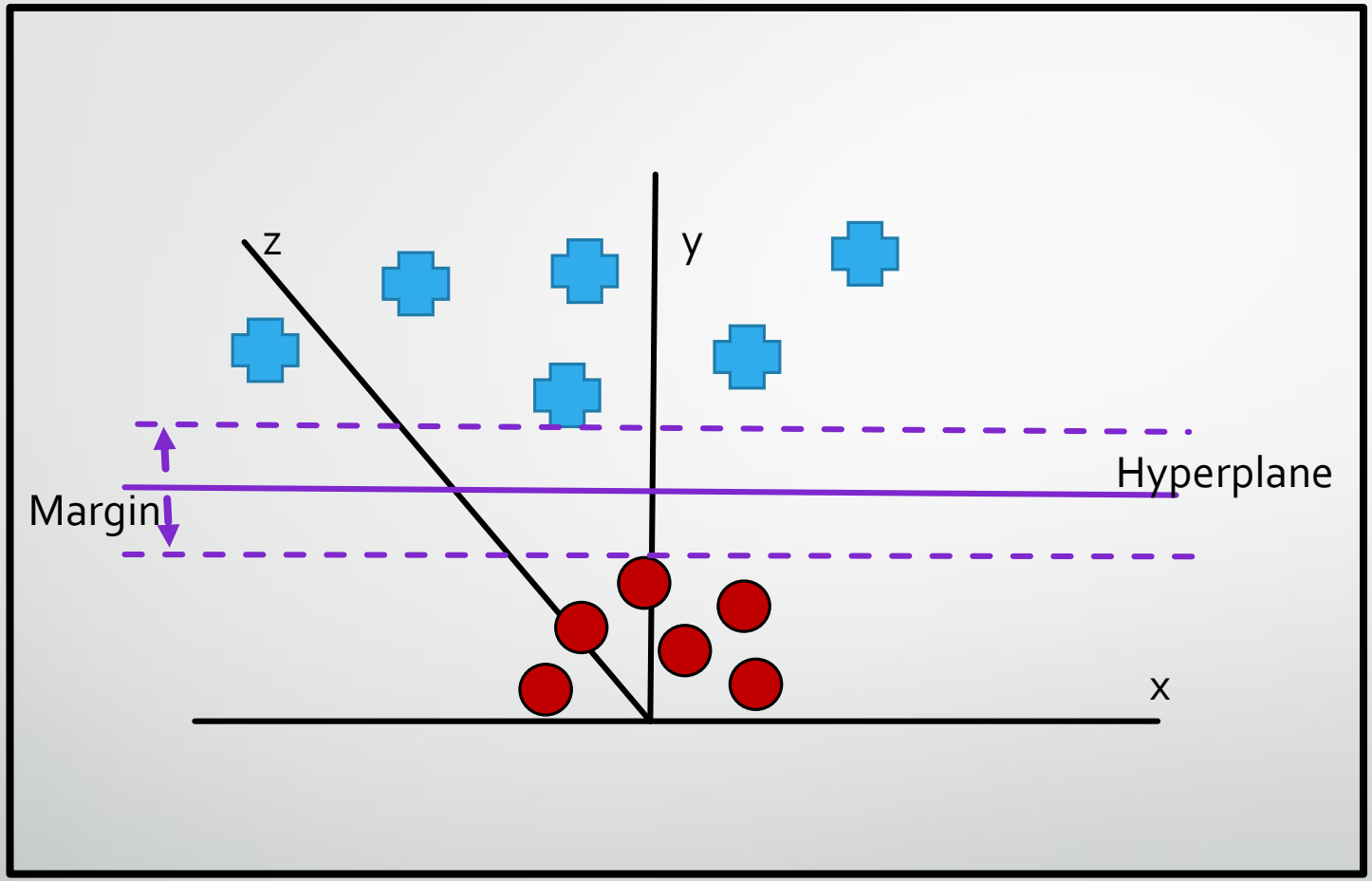


- Fig 9. Non-linear SVM (adapted from javatpoint.com)

3.6 Non-linear SVM Example

- We need to find a way to separate the data.
- In order to do so we introduce a 3rd dimension z which is also called a rotational axis in the Cartesian coordinate system.
- Mathematically $z = x^2 + y^2$
- Introducing this into our scenario and rotating the axes to accommodate z gives us fig 10.

3.6 Non-linear SVM Example



- Fig 10. 3 Dimension view of data point with hyperplane (adapted from javatpoint.com)

3.6 Non-linear SVM Example

- Plotting the data in 3 dimensional space clearly shows us where to place the best hyperplane to use.
- Fig 10 shows the data in 3 dimensional space and also the position of the best hyperplane considering a proper value of C.
- The data can also be plotted back into 2 dimensional space and will produce the same hyperplane in the form of a circle.

Summary

- Logistic regression is the type of regression used in scenarios where the outcome or dependent variable is binary (0 or 1) or discrete categorical (yes/no).
- With logistic regression we are interested in a dichotomous outcome, hence the need to convert the data points based on outcome to probabilities between 0 and 1. The sigmoid function allows us to achieve this.
- Multinomial logistic regression can be used in scenarios where classification is in multiclass, for example when a car can be classified as either a salon(sedan), pickup or station wagon.
- The support vector machine (SVM) is a supervised learning algorithm that is used for both classification and regression problems by introducing a margin that better helps in the process.
- Support vectors are used to determine the margin used by SVM.
- The hyperplane of SVM is simply the boundary that is defined as being the best among the many possible planes that could be used.
- SVM can be classified as being either linear SVM or non-linear SVM.
- Soft margins allow for generalization of the data and avoids overfitting which hard margins do. It is therefore important to fit the margins in such a way that it is not too generalized and not too overfit.

References

- Mining, E. (2019). *Machine Learning for Beginners: A Complete and Phased Beginner's Guide to Learning and Understanding Machine Learning and Artificial Intelligence* . Independent.
- Theobald, O. (2021). *Machine learning for absolute beginners: A Plain english introduction*. nakładem autora.
- Llullaku, S. S., Hyseni, N., Bytyçi, C. I., & Rexhepi, S. K. (2009). Evaluation of trauma care using Triss method: The role of adjusted misclassification rate and adjusted W-statistic. *World Journal of Emergency Surgery*, *4*(1), 2. https://doi.org/10.1186/1749-7922-4-2
- Misra, R. (2021, November 26). *Support vector machines - soft margin formulation and kernel trick*. Medium. Retrieved April 18, 2022, from https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe
- *Support Vector Machine (SVM) algorithm - javatpoint*. www.javatpoint.com. (n.d.). Retrieved April 18, 2022, from https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
- *The top 5 differences between Jaguars and leopards*. The Wildcat Sanctuary. (2015, April 21). Retrieved April 18, 2022, from https://www.wildcatsanctuary.org/the-top-5-differences-between-jaguars-and-leopards/#:~:text=The%20jaguar%20is%20stockier%20and,shorter%20than%20the%20leopard's%20tail.&text=Though%20jaguars%20and%20leopards%20both,rosettes%20have%20spots%20inside%20them.