Machine Learning

Dr Changjiang He, Dr Kuo-Ming Chao
Computer Science| School of Art
University of Roehampton

Lesson 3.3

Other Supervised Learning Methods

Contents

- Naïve Bayes
- Polynomial Regression
- Quantile Regression
- Bayesian Linear Regression Model
- …
- K-nearest Neighbours
- Random Forest
- Support Vector Machine
- …

Naïve Bayes

- Naive Bayes classifier is a probabilistic machine learning based model for classification. The crux of the classifier is based on the Bayes theorem.

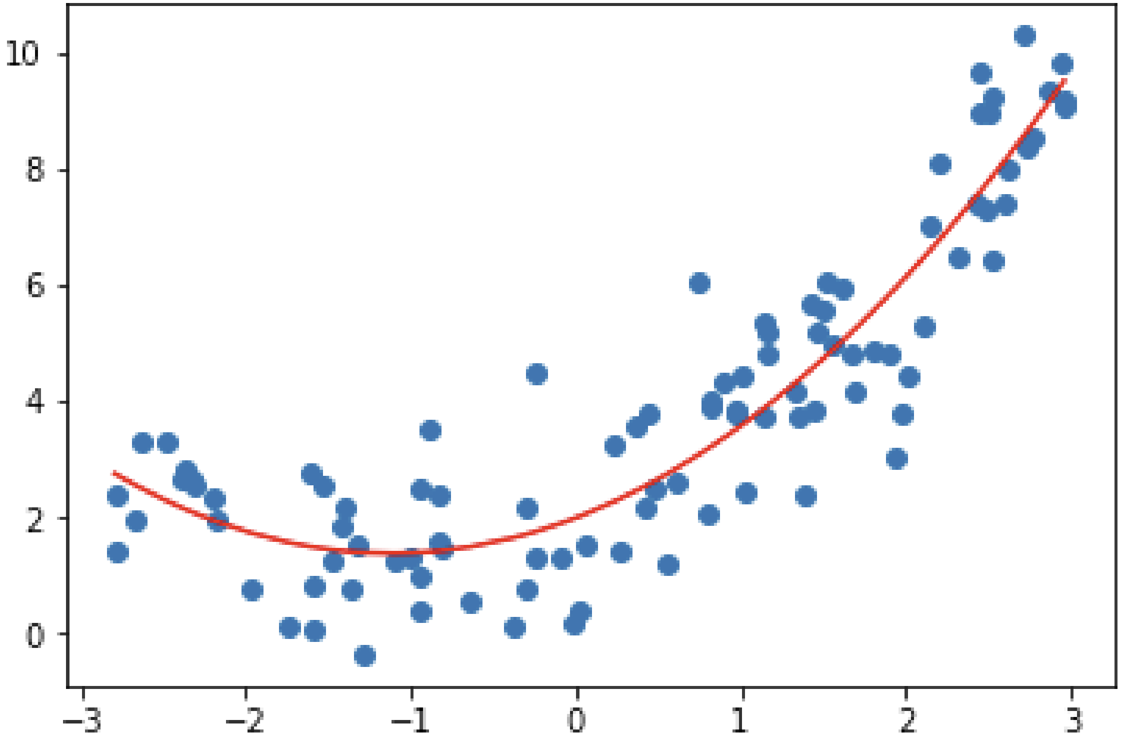
- They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Polynomial Regression

- The technique of polynomial regression analysis is used to represent a non-linear relationship between dependent and independent variables.

- It is a variant of the multiple linear regression model, except that the best fit line is curved rather than straight.

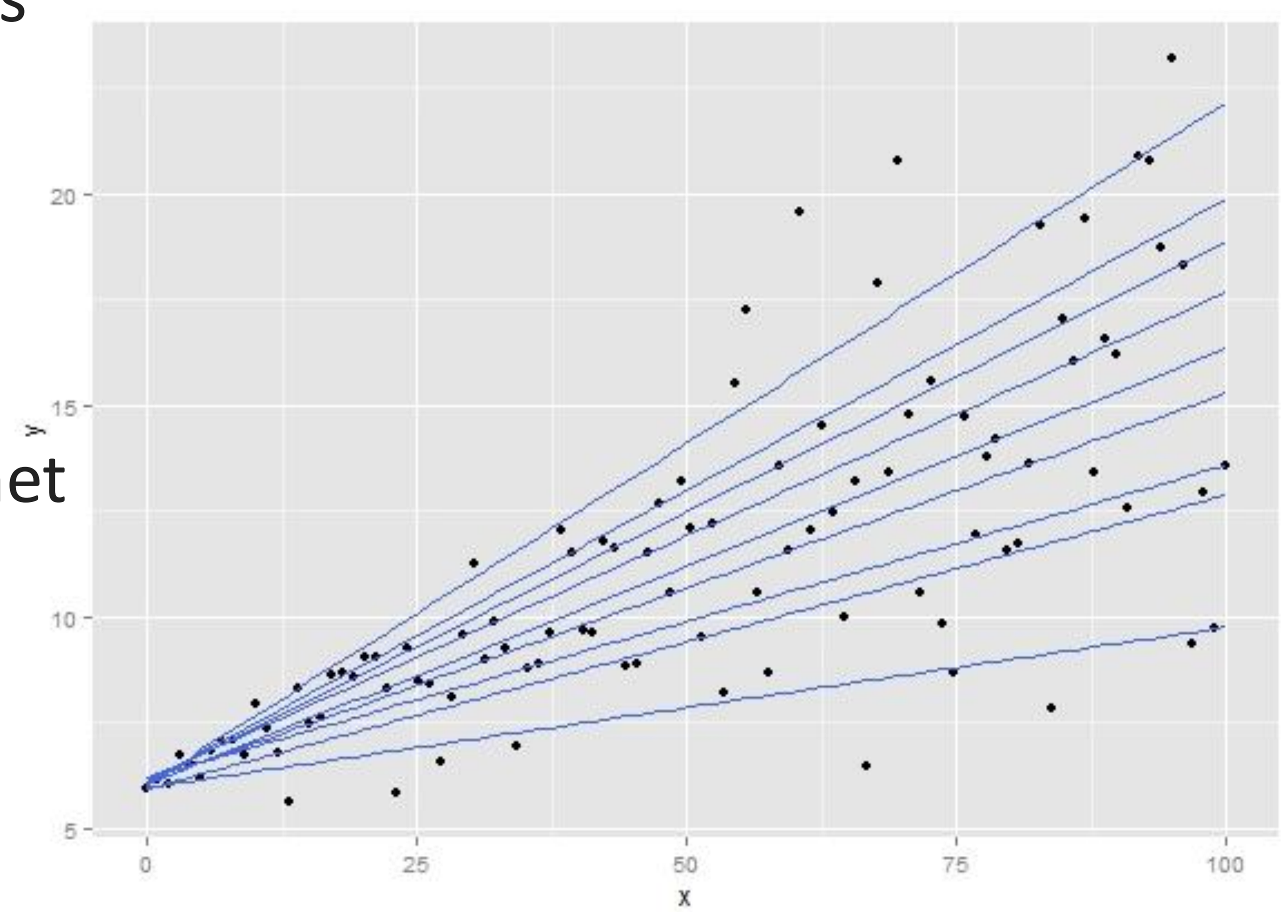


Quantile Regression

- The quantile regression approach is a subset of the linear regression technique.

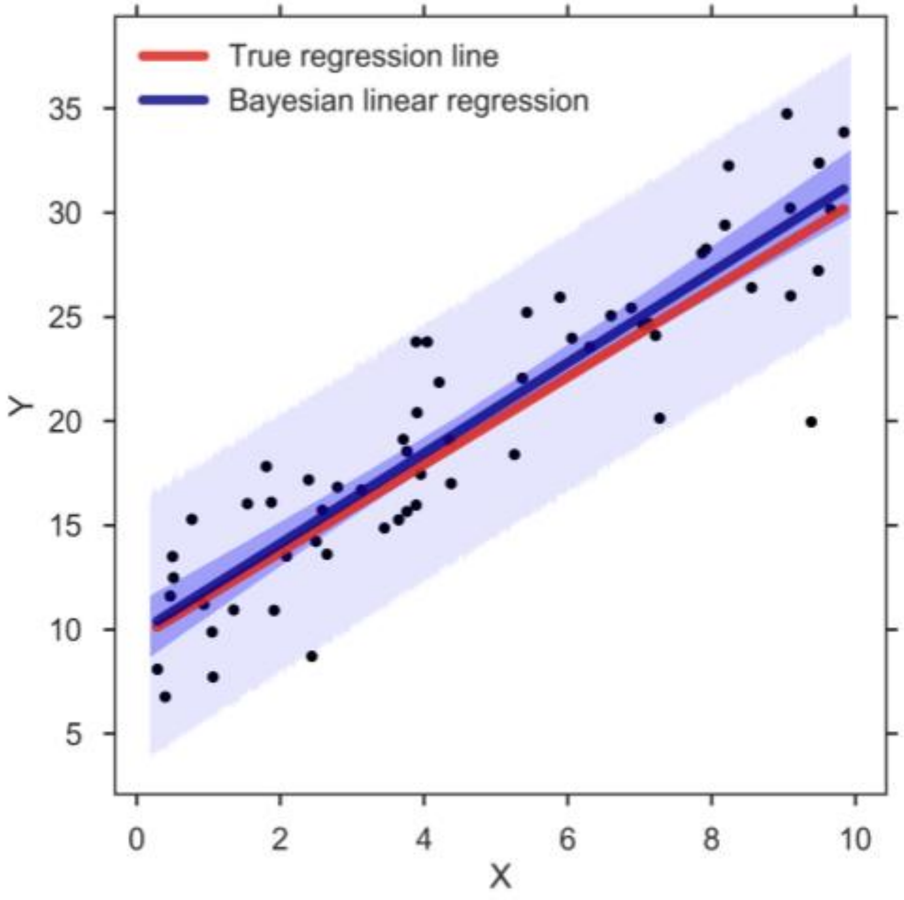
- It is employed when the linear regression requirements are not met or when the data contains outliers.

- In statistics and econometrics, quantile regression is used.



Bayesian Linear Regression Model

- Bayesian linear regression uses Bayes' theorem to calculate the regression coefficients' values.
- Rather than determining the least-squares, this technique determines the features' posterior distribution.
- As a result, the approach outperforms ordinary linear regression in terms of stability.



Explore More Regression Methods by Yourself

- Principal Components Regression
- Partial Least Squares Regression
- Elastic Net Regression
- …

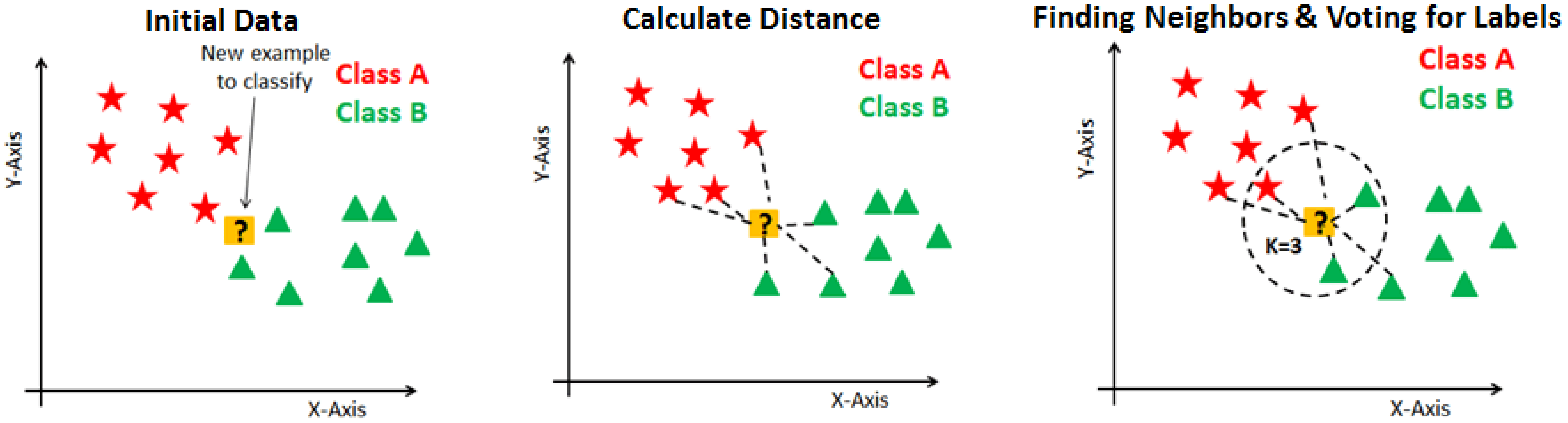
K-nearest Neighbours

- K-nearest neighbours (k-NN) is a pattern recognition algorithm that uses training datasets to find the *k* closest relatives in future examples.

- When k-NN is used in classification, you calculate to place data within the category of its nearest neighbour.

- If *k* = 1, then it would be placed in the class nearest 1. *K* is classified by a plurality poll of its neighbours.

K-nearest Neighbors



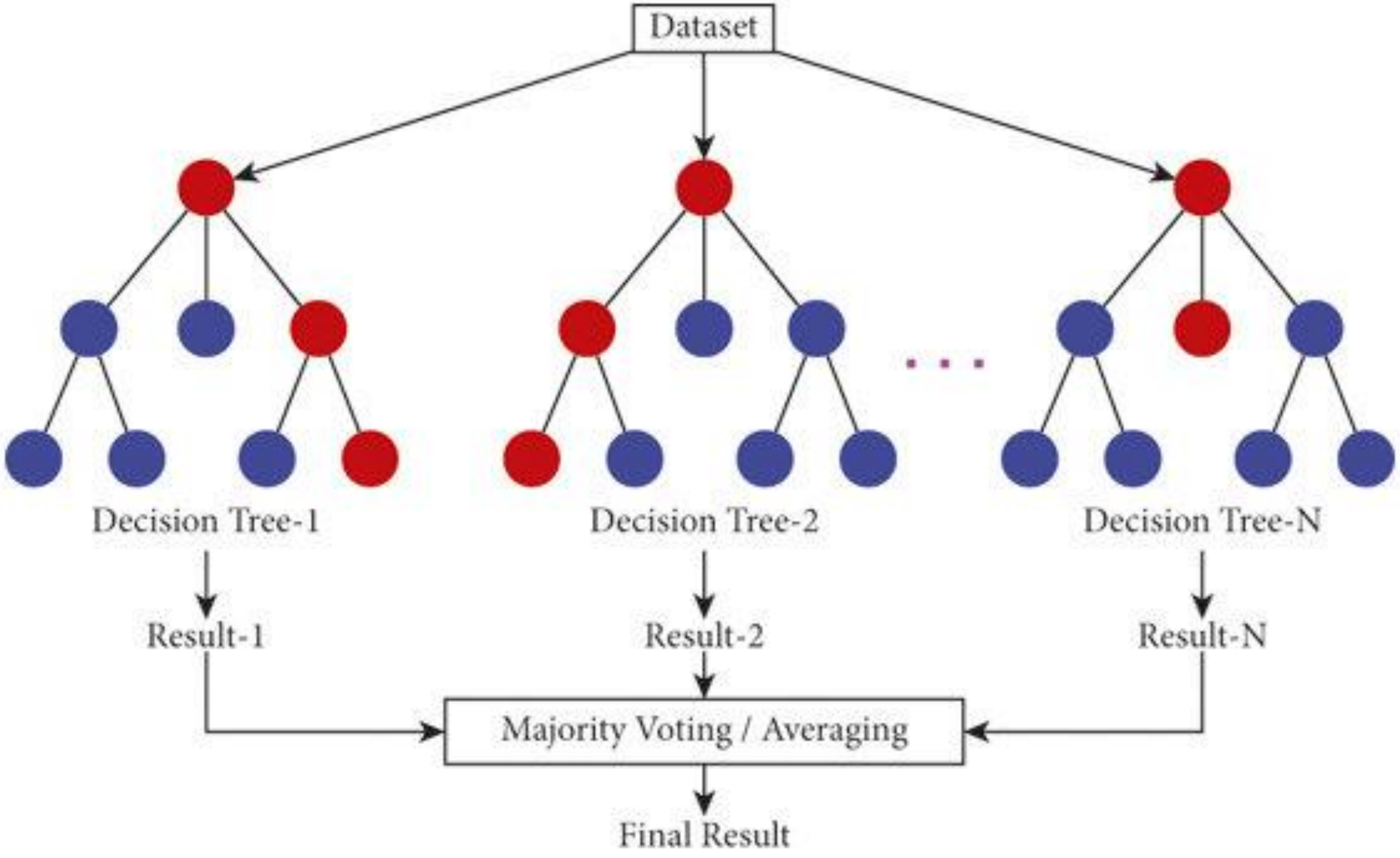
Random Forest

- The random forest algorithm is an expansion of decision tree, in that you first construct a multitude of decision trees with training data, then fit your new data within one of the trees as a “random forest.”

- It, essentially, averages your data to connect it to the nearest tree on the data scale.

- Random forest models are helpful as they remedy for the decision tree’s problem of “forcing” data points within a category unnecessarily.

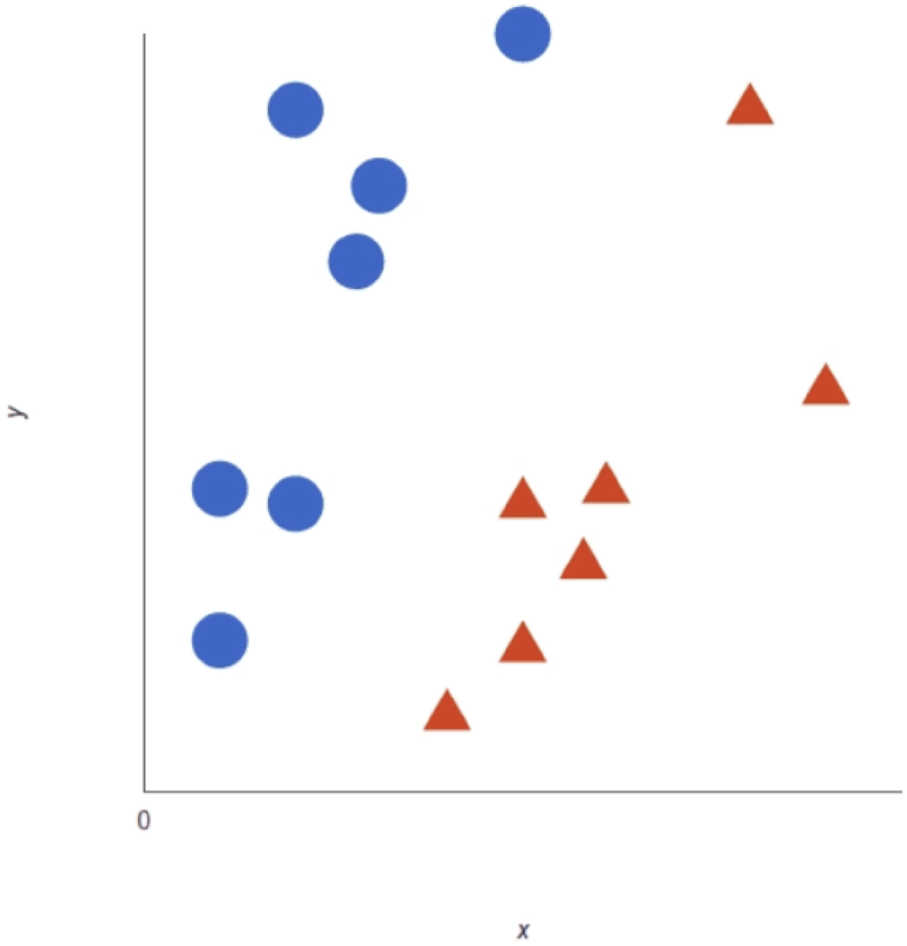
Random Forest



Support Vector Machine

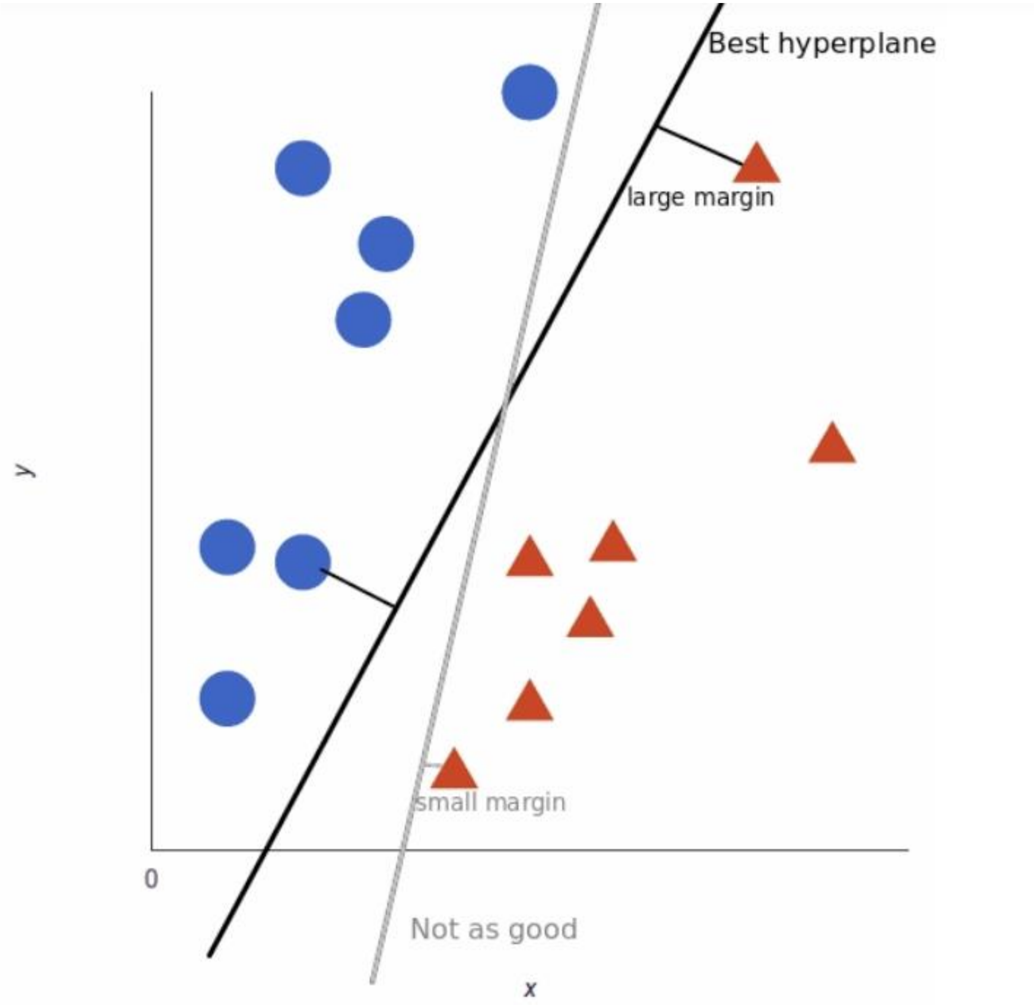
- A support vector machine uses algorithms to train and classify data within degrees of polarity, taking it to a degree beyond *X/Y* prediction.

- For a simple visual explanation, we'll use two tags: *red* and *blue*, with two data features: *X* and *Y*, then train our classifier to output an *X/Y* coordinate as either *red* or *blue*.



Support Vector Machine

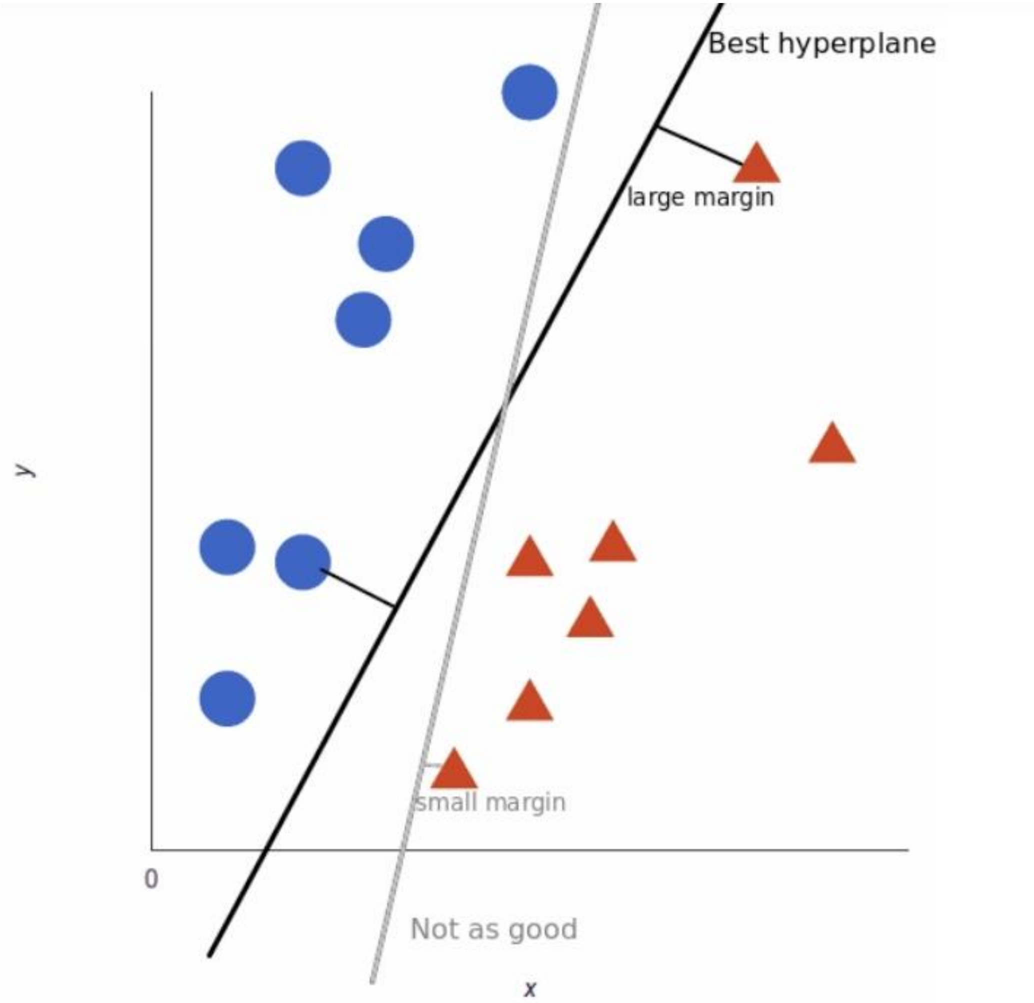
- The SVM then assigns a hyperplane that best separates the tags.
- In two dimensions this is simply a line.
- Anything on one side of the line is *red* and anything on the other side is *blue*.



Support Vector Machine

- In sentiment analysis, for example, this would
be *positive* and *negative*.

- In order to maximize machine learning, the best hyperplane is the one with the largest distance between each tag.

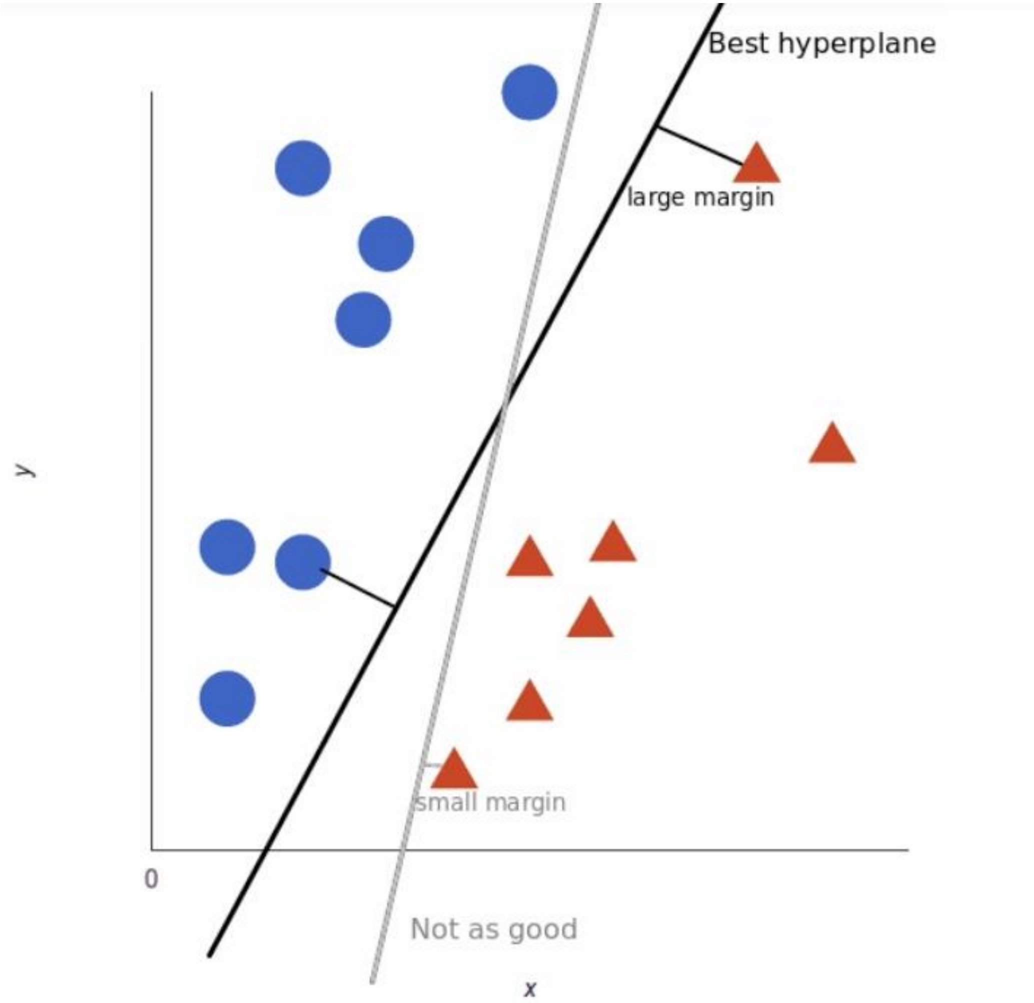


Support Vector Machine

- However, as data sets become more complex, it may not be possible to draw a single line to classify the data into two camps.

- Imagine the above in three dimensions, with a *Z-axis* added, so it becomes a circle.

- SVM allows for more accurate machine learning because it’s multidimensional.


Best hyperplane
large margin
y
small margin
0
Not as good
x