Support Vector Machines

In honour of Breast Cancer Awareness Month, October 2020, along with a Case Study on Breast Cancer Cell Classification (on the basis of whether they are MALIGNANT OR BENIGN)

About Support Vector Machines

- Supervised Classification Machine Learning model.

- They work in 3 simple steps:
  - Start with data in a relatively low dimension
  - Move the data into a higher dimension
  - Find a Support Vector Classifier that separates the higher dimension data into two groups

They are great because...

- ★ They can handle outliers
- ★ They allow a few misclassifications in order to predict a better result (Bias-Variance tradeoff)
- ★ They can handle overlapping classifications
- ★ They can fit data within an n-dimensional plane (hyperplane)

Important Terms

★ **Margin** - Shortest distance between the observations and the threshold

★ **Soft Margin** - When we allow misclassifications within the margin, then this is the distance between the observations and the threshold

★ **Support Vectors** - The observations on the edge and within the soft margin

Important Terms, continued

★ **Support Vector Classifier -** (AKA Soft Margin Classifier) When a soft margin is used to determine the location of a threshold, then an SVC is being used to classify observations

★ **Support Vector Machines -** Used when our data cannot be classified using a simple, 2D plane

Important Terms, continued

- ★ **Hyperplane -** Any plane with ***n*** number of dimensions

- ★ **Kernels -** These are mathematical functions used to transform the data so that they can fit in an n-dimensional plane. There are three types of kernels commonly used:
  - Linear Kernels
  - Polynomial Kernels
  - Radial Kernels

Linear and Polynomial Kernels

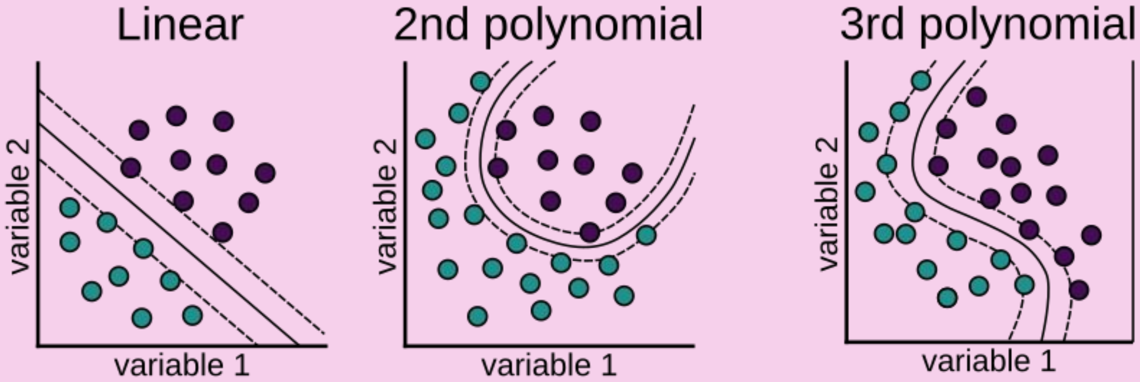
★ ***Linear Kernels*** - These are used to make simple SVCs without any sort of data transformation.

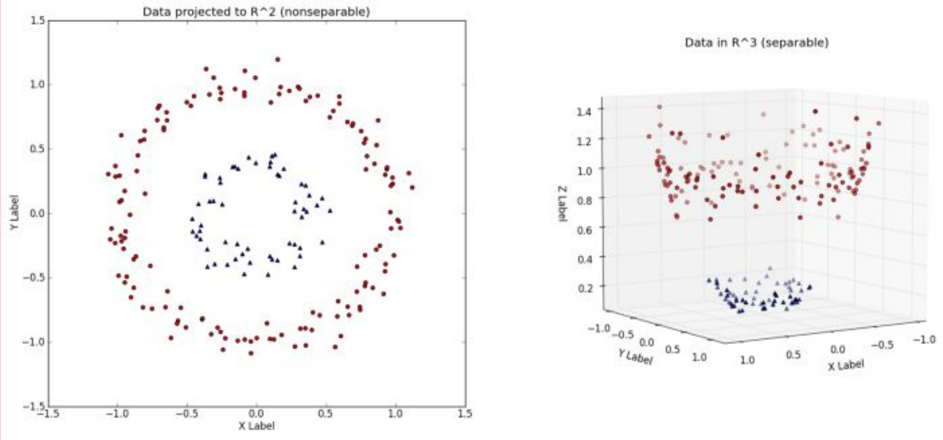
★ ***Polynomial Kernels*** - These increase the dimension of data to fit it on a higher plane, by setting the degree of the polynomial ***d*** and the relationships between each pair of observations and using these to find a good SVC.

Radial Kernels

- AKA ***Radial Basis Function Kernels***
- Find a SVC in infinite dimensions and behaves like a weighted k-NN model.
- Closest observations (nearest neighbours) have a lot of influence on how the new observations are classified.
- Since this only calculates the relationships between every pair of points as if they are in higher dimensions, they don't do any transformations on the data.

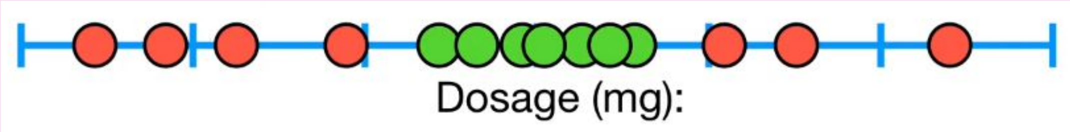
Support Vector Machine





Example: Medicine Dosage works on Patients

- ★ Medicine Dosage that work on Patient.
- ★ Dosage doesn’t work when the quantity is high or low. Represented by Red Dots
- ★ Dosage only works when the dosage is optimal which is in the middle represented by the Green dots.



Polynomial Function

Function = $(a * b + r)^d$

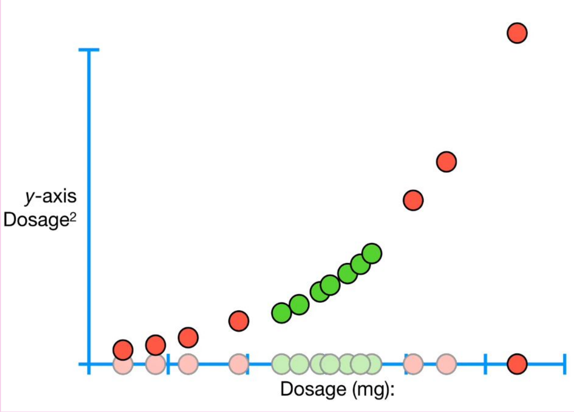
Degree(d) = 2, Constant (r)= 1

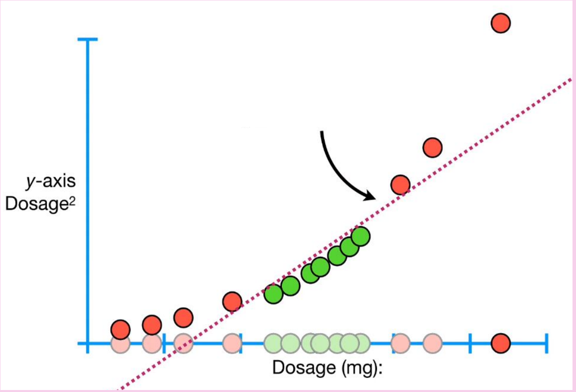
Function = $(a*b+1)*(a*b+1)$

$= (a^2b^2+2ab+1)$

Dot Product =
$(\sqrt{2}a, a^2, 1).(\sqrt{2}b, b^2, 1)$

Which gives the coordinates in higher dimension.





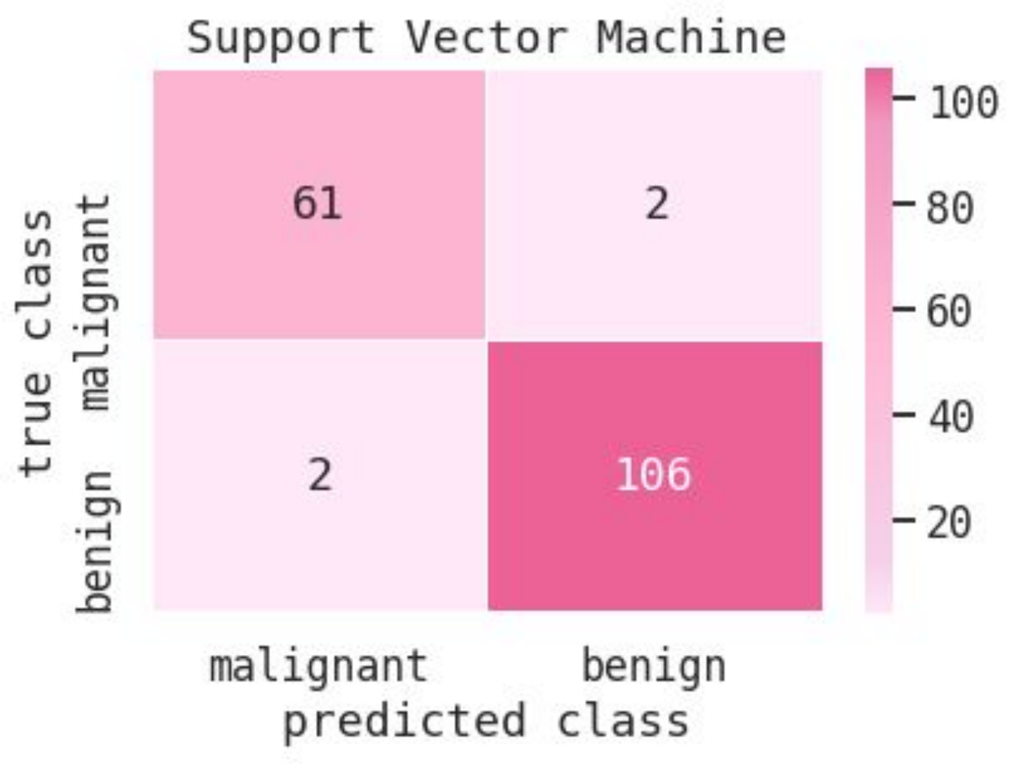
Case Study

Breast Cancer Cell Classification

- ***Target Variables*** - Malignant, Benign
- ***Features*** - There are 30 input features of cancer cells defined by this dataset
- ***Shape*** - There are 568 rows and 30 columns of data
- Training and testing data have a 70:30 split
- Optimal Parameters for training the SVM were:
  - ***Kernel*** - RBF
  - ***C*** - 0.5
  - ***Gamma*** - 0.0001

Result - Confusion Matrix

- ★ TN - 61 / 63
- ★ TP - 106 / 108
- ★ FN - 2 / 63
- ★ FP - 2 / 108



- ★ Link to Program

Result - Classification Report

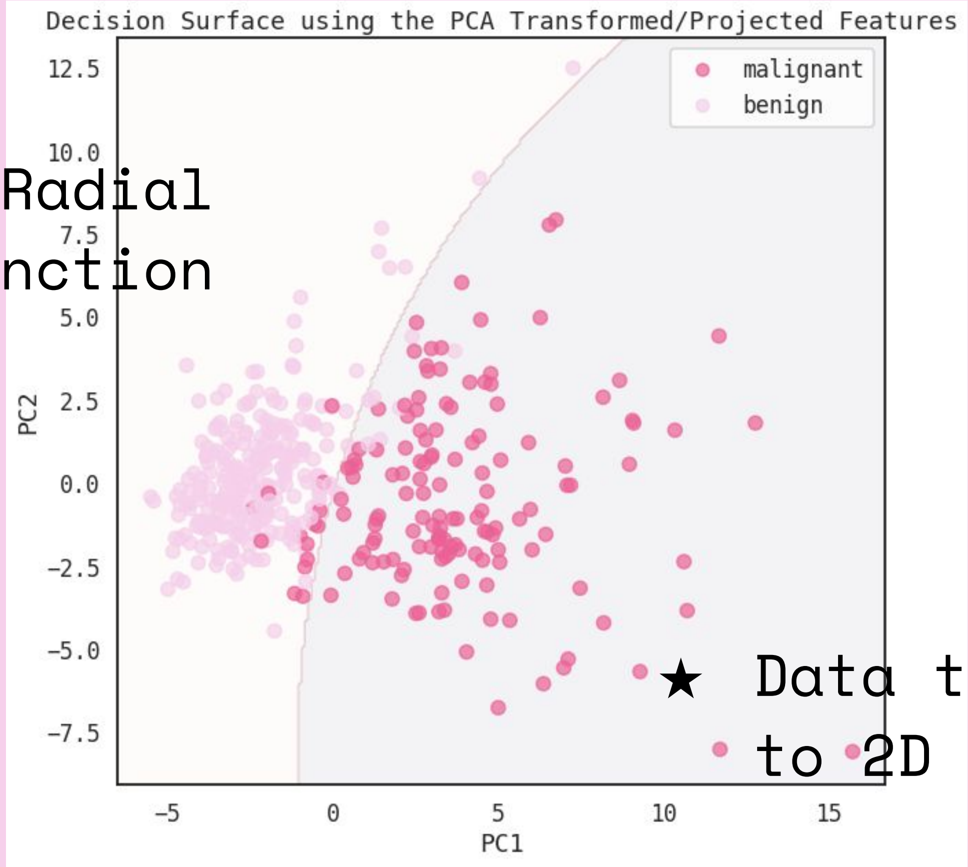
- ★ accuracy score: 97.66%
- ★ precision score: 98.15%
- ★ recall score: 98.15%
- ★ f1 score: 98.15%

- ★ MSE (Mean Square Error) = 2.34%

classification report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	63
1	0.98	0.98	0.98	108
accuracy			0.98	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.98	0.98	0.98	171

Result - Decision Boundary - PCA



★ Using a Radial Basis Function Kernel

★ Data transformed to 2D using PCA

References + Resources Used

- ★ Support Vector Machines, Clearly Explained!!! - StatQuest with John Starmer (YouTube)
- ★ Python Data Science Handbook - Jake VanderPlas
- ★ CyberPunk Style Matplotlib Graphs - Dominik Haitz (Matplot Blog)
- ★ Support Vector Machine Detailed Analysis - Niraj Verma (Kaggle)
- ★ UCI Machine Learning Repository (Dataset)