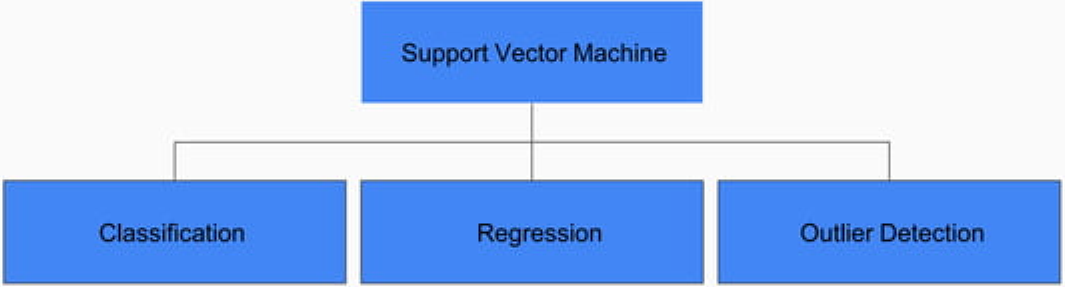
Agenda

- Introduction to SVM
- Advantages & Disadvantages of SVM
- Classification Algo using SVM
- Linearly Separable Data
- Non linearly separable Data
- Understanding Kernels
- Gamma & C for RBF kernel
- Unbalanced Data
- Regression using SVM
- Kernels in Regression
- Novelty Detection
- Additional
- Applications

Introduction to Support Vector Machine

- First developed in the mid-1960s by Vladimir Vapnik.



Advantages of SVM

- Effective in high dimensional spaces.
- Where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

Disadvantages of SVM

- Very sensitive to hyper-parameters
- Different kernels needs different parameters
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation

Classification using SVM

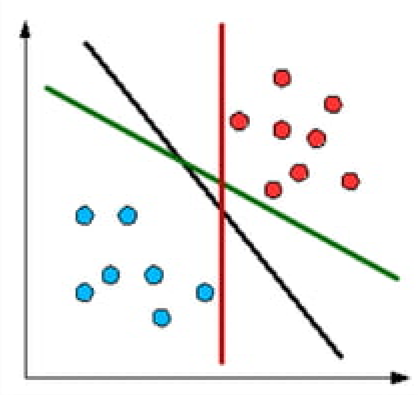
Algo

- This algorithm looks for a linearly separable hyperplane, or a decision boundary separating members of one class from the other.
- If such a hyperplane does not exist, SVM uses a nonlinear mapping to transform the training data into a higher dimension. Then it searches for the linear optimal separating hyperplane.
- With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.
- The SVM algorithm finds this hyperplane using support vectors and margins.

Linearly Separable Data

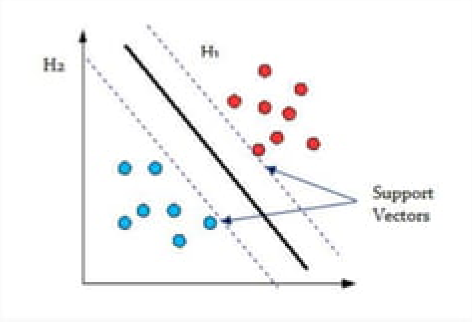
- Among the infinite straight lines possible to separate the red from blue balls, find the optimal one.
- Intuitively it is clear that if a line passes too close to any of the points, that line will be more sensitive to small changes in one or more points.
- Hyperplane Generalization

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n = 0$$



Maximum Margin Classifier

- A natural choice of separating hyperplane is optimal margin hyperplane (also known as optimal separating hyperplane) which is farthest from the observations.
- Finding the hyperplane that gives the largest minimum distance to the training examples, i.e. to find the maximum margin.

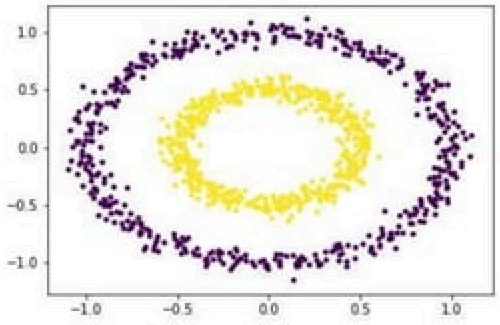


Soft-margin Classifier

- Maximum margin classifier may not practically exist.
- Extending maximum margin to a soft-margin, a small amount of data is allowed to cross margins or even the separating hyperplanes.
- Support Vector Machine maximizes the soft margin.
- C parameter leads to larger penalty for errors & thus inversely proportional to soft margin.

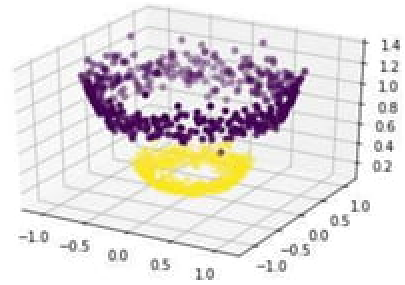
Data non-linearly Separable

- Can't really draw a line to seperate yellow from purple



Data non-linearly Separable

- If such a hyperplane does not exist, SVM uses a nonlinear mapping to transform the training data into a higher dimension.
- Transformation , Z = X^2 + Y^2
- Now, we see a plane exists separating them both



Kernels

- We don't have to do the transformation manually.
- This is done by kernel tricks

Kernel Tricks

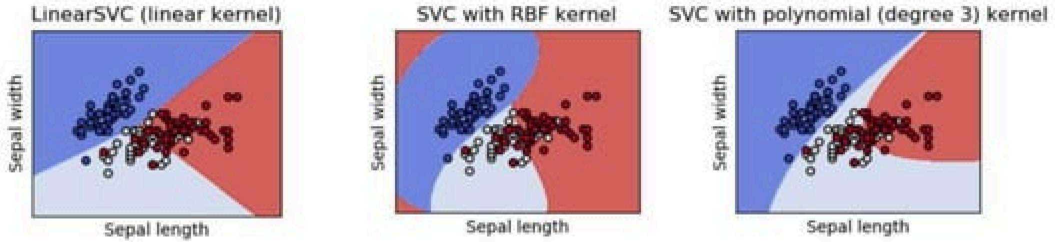
Linear

Poly

RBF

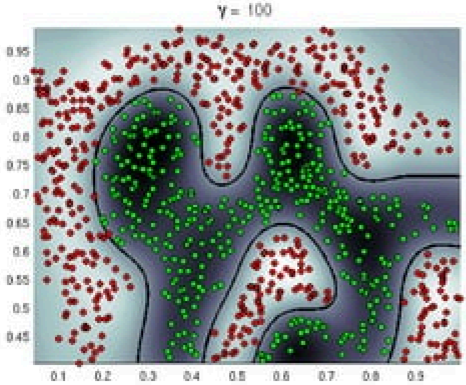
Custom

Comparing Kernels - Classification



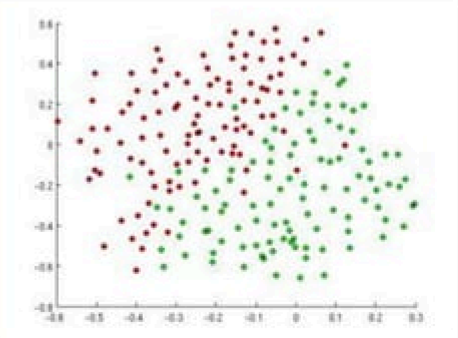
RBF

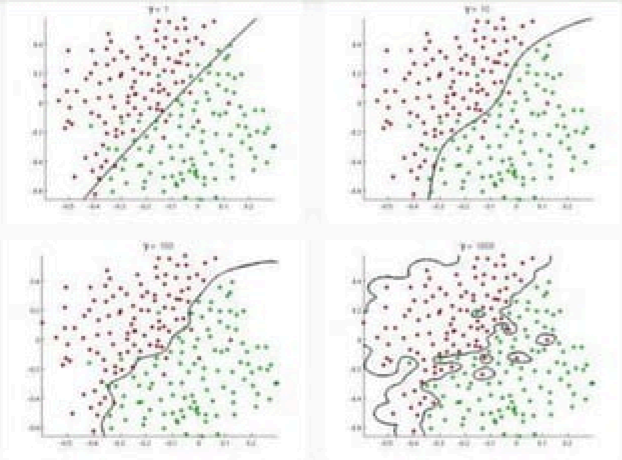
$$K\left(x^{(i)}, x^{(j)}\right) = \phi(x^{(i)})^T \phi(x^{(j)})$$
$$= \exp\left(-\gamma \left\| x^{(i)} - x^{(j)} \right\|^2\right), \qquad \gamma > 0$$



Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'.

RBF & Gamma



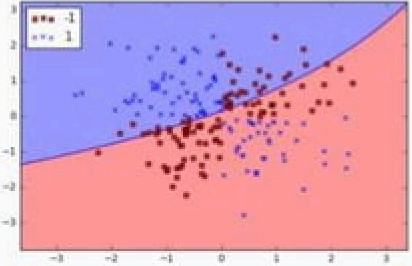


RBF & C

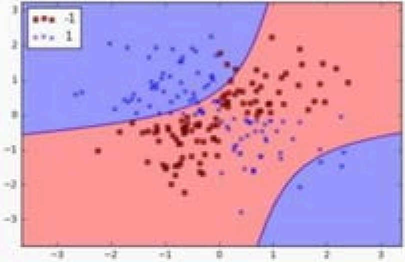
- The C parameter trades off misclassification of training examples against simplicity of the decision surface.
- A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors.

RBF & C

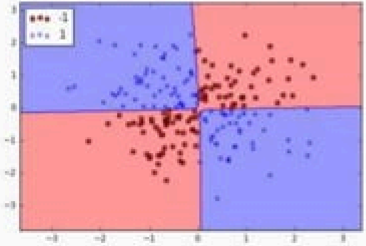
XOR Data



C = 1

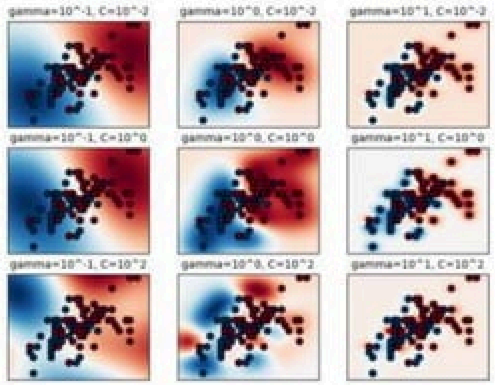


C = 10



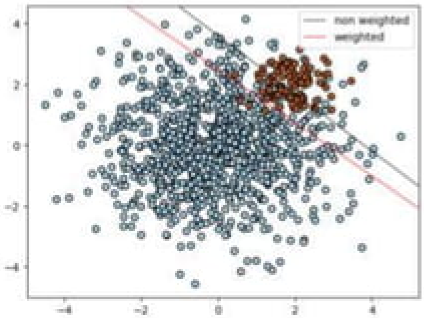
C = 10000

RBF - Gamma & C


gamma=10^-1, C=10^-2
gamma=10^0, C=10^-2
gamma=10^1, C=10^-2
gamma=10^-1, C=10^0
gamma=10^0, C=10^0
gamma=10^1, C=10^0
gamma=10^-1, C=10^2
gamma=10^0, C=10^2
gamma=10^1, C=10^2

Unbalanced Data

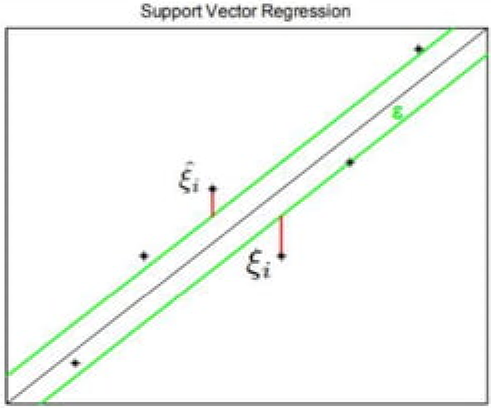
- Find the optimal separating hyperplane using an SVC for classes that are unbalanced.
- Parameter - class_weight



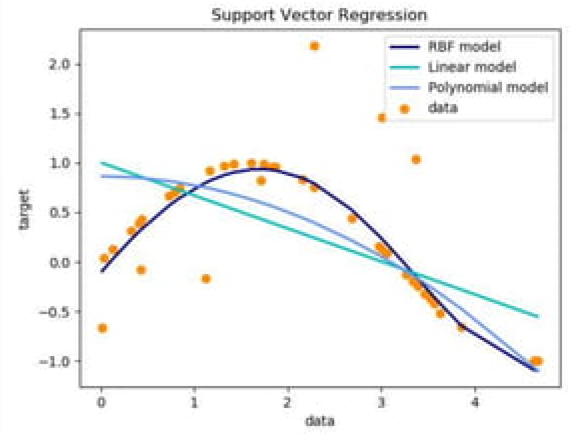
Regression using SVM

Foundations

- y = f(x) + noise
- This can be achieved by training the SVM model on a sample set, i.e., training set, a process that involves sequential optimization of an error function.



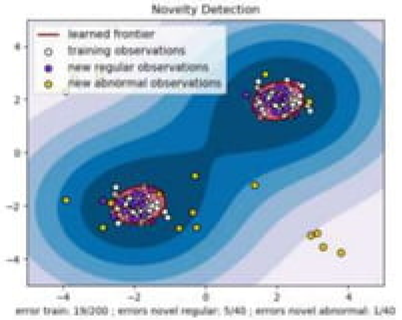
Comparing Kernels - Regression



Novelty Detection using SVM

One Class SVM

- The training data is not polluted by outliers, and we are interested in detecting anomalies in new observations.
- One-class SVM is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set.



Additional : Custom Kernel

- You can also use your own defined kernels by passing a function to the keyword kernel in the constructor.

```
>>> import numpy as np
>>> from sklearn import svm
>>> def my_kernel(X, Y):
...     return np.dot(X, Y.T)
...
>>> clf = svm.SVC(kernel=my_kernel)
```

Applications

Image Classification

Text Classification