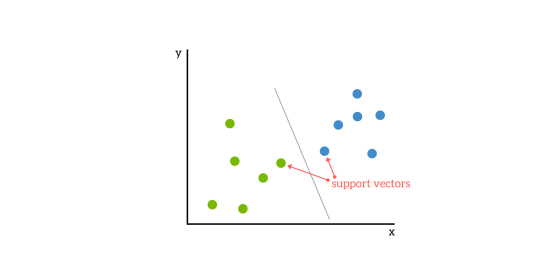
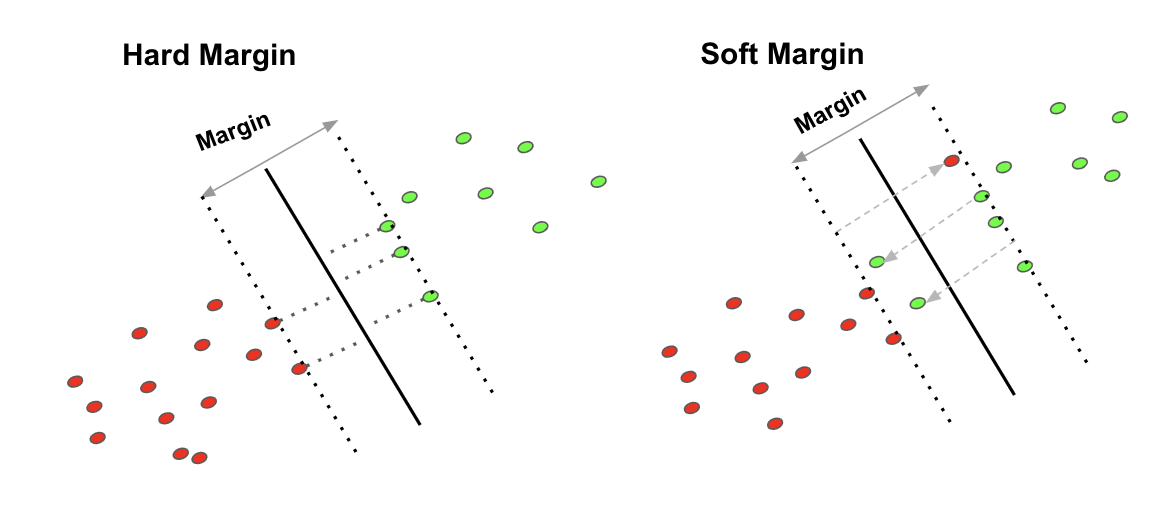
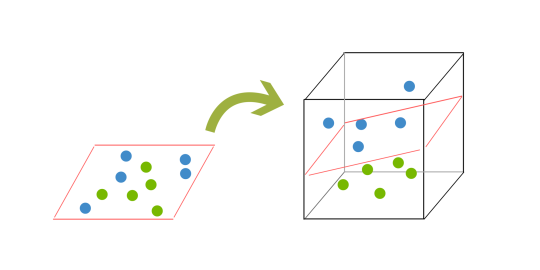
SVM:



* SV are data points closest to hyper plane
* If these data points are removed, the position of hyperplane changes
* Farther the data point from the hyperplane, more confident we are that data is correctly classified



* Dist. b/w hyperplane and nearest datapoint form either set is called margin
* Aim is to choose a hyperplane with greatest possible margin b/w hyperplane and any point in train set so that new data is correctly classified
* There are two kinds of margins
  1. Hard margin:
     + Strict in dividing the hyperplane.
     + Doesn’t allow for miss classification
     + Suitable for linearly separable data
  2. Soft margin:
     + In real time data is not linearly separable
     + So, we allow margin violation to occur i.e., miss classification is allowed



* ‘C’ is the parameter used to define the soft margin. It adds a penalty for each misclassified data point.
  1. Small ‘c’---🡪low penalty---🡪larger margin at expense of misclassification
  2. Large ‘c’---🡪high penalty---🡪smaller margin (overfit)
  3. Penalty is directly proportional to decision boundary
* When clear separation of datapoints is not possible, we transform data into higher dimension. now the separable is plane since its 3d. this mapping of data into a higher dim. Is called kernelling
* Kernels:
  1. Linear SVM
     + Rarely used.
     + In real time there is no linear separable data
  2. RBF (gaussian) radial base function
     + Circular bounded regions
     + Prone to overfit
     + Capture noise
  3. Polynomial
     + Number of higher dimensional features increases exponentially
     + Therefore, computationally expensive
  4. Sigmoid
* Gamma:
  1. Gamma parameter of RBF controls distance of influence of a single training point.
  2. Low gamma indicates similarity radius which results in more points grouped together
  3. High gamma indicates points need to be very close to each other in order to be considered in the same group or class
  4. High gamma ------🡪 overfit
* Gamma vs C:
  1. For linear kernel, optimizing ‘C’ parameter would suffice.
  2. For RBF kernel, both ‘C” and gamma need to be optimized simultaneously.
     + If gamma is large effect of ‘c’ becomes negligible
     + If gamma is small, c affects the model just like it affects the linear model
       - 0.0001 < gamma < 10
       - 0.1 < c < 100



* SVM chooses ‘A’ as hyperplane because it gives more importance accuracy of classification than maximizing margin



* SVM has feature to ignore outliers and find hyperplane that has max. margin. We can say that SVM classification is robust to outliers.
* Less effective on noisier datasets with overlapping classes
* Training time is high for larger datasets