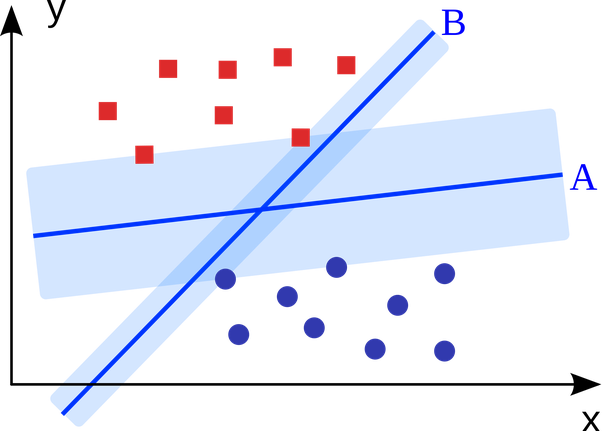
**SUPPORT VECTOR MACHINE (SVM)**

* SVM is binary classification algorithm
* Suppose we have some set of points of two types in N-dimensional space, SVM generates N-1 dimensional hyperplane that separates those points into two groups.

****

**Maximal-margin classifier:**

* The input variables (x) in our dataset (column) forms n-dimensional space. For 2 variables, it forms 2-dimensioanl space.
* The hyper plane is the line that separates the input space into groups based on whether the input belongs to class-0 or class-1.
* The following equation helps SVM to select the hyperplane:

**B0 + (B1 \* X1) + (B2 \* X2) = 0,** where B1, B2 determines the slope of the line and B0 is the intercept and X1 and X2 are input variables.

* By plugging in input values into the line equation, we can calculate whether a new point is above or below the line.
* Above the line, the equation returns a value greater than 0 and the point belongs to the first class (class 0). Below the line, the equation returns a value less than 0 and the point belongs to the second class (class 1).
* A value close to the line returns a value close to zero and the point may be difficult to classify. If the magnitude of the value is large, the model may have more confidence in the prediction
* The distance between the hyperplane and the closest data points is referred as a margin.
* The optimal line that separates those points into two classes is the larger margin. Hence it is referred to as Maximal-margin hyperplane.
* The perpendicular distance between the line and points closest to that line is a margin, which can only be used for line definition and classification. These points are called support vectors as they support or define the hyperplane.

**Soft margin classifier:**

* Real dataset are very complex and cannot be easily separated by the hyperplane. Hence the maximal distance between two classes cannot be achieved as there are more messy data points. Hence some points can be allowed that violates the separating line since we cannot perfectly separate the messy data points. This is called as soft margins.
* In order to allow data points that violate separating the line, additional co-efficient called slack variables are introduced in all dimensions.
* Another parameter called C is introduced, that defines the magnitude of violation allowed in the margin.
* C=0 implies, no violation and it is Maximal-margin classifier.
* Larger C value implies more violation in the hyperplane definition.

**Support vector machine kernels:**

* The support vector machines are implemented using kernels.
* The kernel defines the similarity or distance between the input (x) and the support vectors.
* The inner product between two vectors is the sum of the multiplication of each pair of input values. The equation is as follows: **f(x) = B0 + sum(ai \* (x,xi)).**
* The above equation calculates the inner product between the inputs (x) and all the support vectors (xi).The coefficients for each pair of input B0 and ai can be learned from training the dataset.

**Linear kernel SVM:**

* The dot product is called as the kernel and can be defined as:
  + - K(x, xi) = sum(x \* xi)
* The dot product is the similarity measure used for linear SVM or a linear kernel because the distance is a linear combination of the inputs.

**Polynomial kernel SVM:**

* For higher dimensional dataset, the polynomial kernel is defined as follows:
  + - K(x, xi) = 1 + sum(x \* xi) ^d, where d is the degree of polynomial defined from learning algorithm.
    - When d=1, it is same as linear kernel.

**Radial kernel SVM:**

* The radial kernel is defined as,
  + - K(x, xi) = exp (-gamma \* sum ((x – xi^2)), where gamma can be defined from learning algorithm.
    - **A good default value for radial kernel is 0.1, where 0<gamma<1.**

**Advantages of SVM:**

* SVM can model complex real world problems very easily.
* Training datasets in SVM is very easy.
* Performs well with datasets with any number of attributes.

**Drawbacks of SVM:**

* The biggest limitation of the support vector approach lies in choice of the kernel
* Discrete data handling can be a problem
* The optimal design for multiclass SVM classifier is under research
* Behaves slowly in testing phase