# Linear Regression for Classification

## Problem Statement

The use of iris data set for the prediction of species is a classic example for classification problem. This classification problem needs to be solved by the Linear Regression which is a supervised learning problem. A linear regression algorithm needs to be developed that can predict the species of input provided to the algorithm with almost certainty (close to 100% accuracy).

## Data

The Iris Data set contains 5 columns – Sepal Length, Sepal Width, Petal Length, Petal Width and Species.

We use the first 4 columns as our features – Sepal length, Sepal width, Petal length and Petal width. All the columns for features are float values. The values in these columns make up the A matrix.

Last column is going to be used as the labels. It states the species of the plant which have the corresponding features. The data set contains 3 species – Iris setosa, Iris virginica and Iris versicolor.

Example of data:

Data visualization:

## Strategy

The strategy for implementing a solution is to use k-folds cross validation to create k number of bins in the data and then train and test data on each of these bins which will give **Beta** for each the bins. To reduce the overfitting, we will find the **Beta Mean** values which will give a good linear regression for classifying these records.

**Fit Method:** The fit method is the method use for obtaining the **Beta** values from the matrix **A**, from records in the features for each bin and **Y**, from the label encoded values for the species.