Iris Flower Dataset

Table of Contents

Introduction

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Predicted attribute: class of iris plant.

This is an exceedingly simple domain.

This data differs from the data presented in Fishers article (identified by Steve Chadwick, spchadwick '@' espeedaz.net).

- The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature.
- The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa" where the errors are in the second and third features.

GOALS

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).

Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, We developed a some discriminant model to distinguish the species from each other.

Importing Packages

1. **Numpy -** Implementing multi-dimensional array and matrices.
2. **Pandas -** For data manipulation and analysis.
3. **Matplotlib -** Plotting library for Python programming language and it's numerical mathematics extension NumPy.
4. **Seaborn -** Provides a high level interface for drawing attractive and informative statistical graphics.
5. **Scikit-learn -** Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Data Loading and Data Description

1. This dataset has **150** rows and **6** columns.
2. Summary of data types in this dataset:
    - **Numeric:** 4 (Float), 1 (Integer)
    - **Object:** 1
3. None of the variables having null and zero values.

Data Understanding and Exploration

1. **Id** column has no significance thus deleted **Id** column.
2. **Species** column is target variable and having three different species mentioned below:
   a. **Iris-setosa**
   b. **Iris-versicolor**
   c. **Iris-virginica**
3. There are no columns having high correlation with each other.
4. Below mentiones variable are measured in Centimeters.
   a. **SepalLengthCm**
   b. **SepalWidthCm**
   c. **PetalLengthCm**
   d. **PetalWidthCm**

Metrics for Evaluating ML Algorithms

Model evaluation aims to estimate the generalization accuracy of a **model** on future (unseen/out-of-sample) data.

Different performance metrics are used to evaluate different Machine Learning Algorithms. Here we have used **Classification Metrics** with following tools.

1. **Classification Report**
    a. **Precision – What percent of your predictions were correct?**
    b. **Recall – What percent of the positive cases did you catch?**
    c. **F1 score – What percent of positive predictions were correct?**
2. **Confusion Matrix**
3. **Classification Rate/Accuracy**

Label Encoding

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important preprocessing step for the structured dataset in supervised learning.

Here we encoded **Species** column & Splitting the dataset in independent and dependent variables.

Feature Scaling - Standard

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

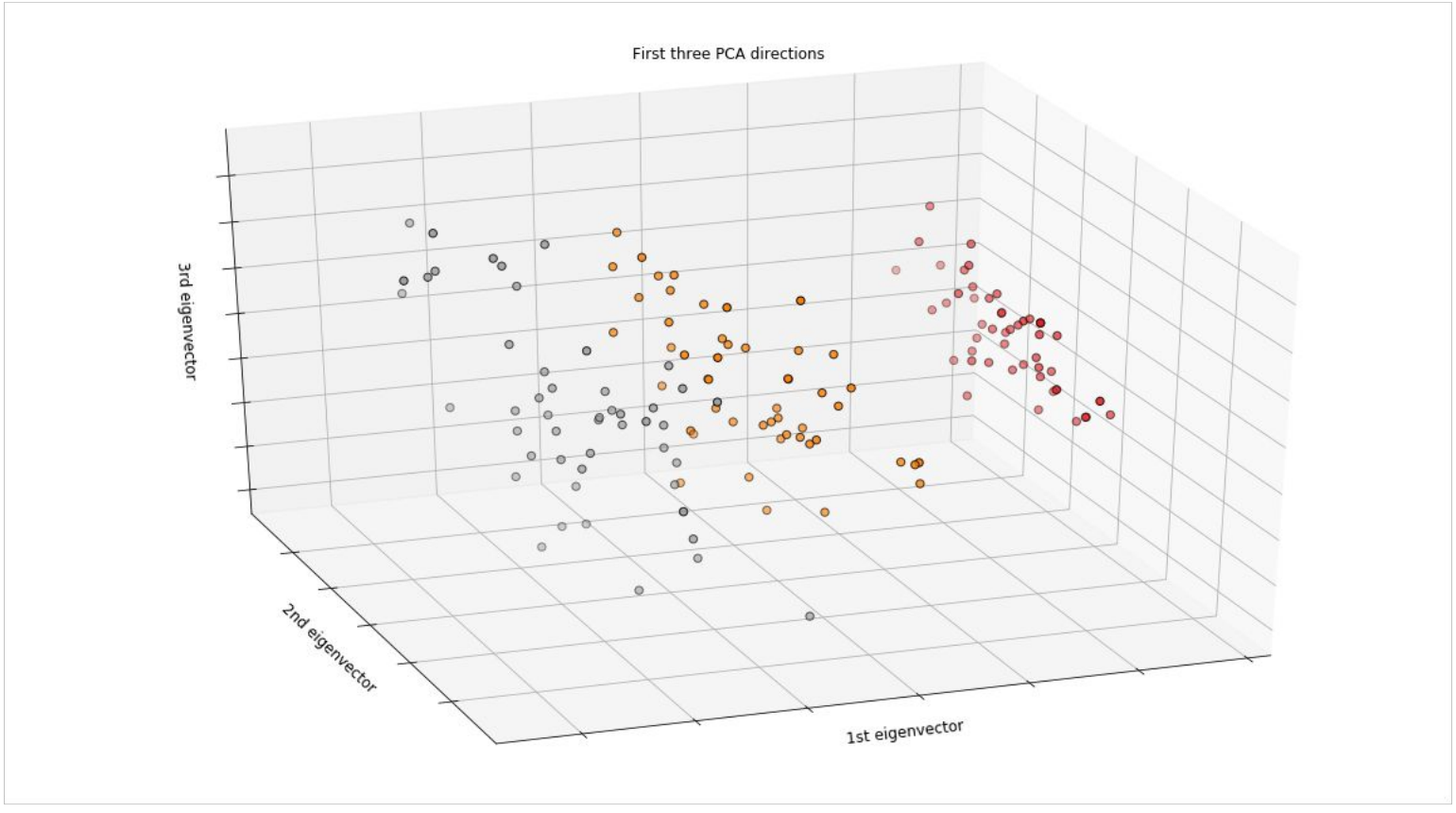
Consider the two most important ones:

- **Min-Max Normalization**
- **Standardization**

Dimensionality Reduction

Dimensionality reduction is a really important concept in Machine Learning since it reduces the number of features in a dataset and hence reduces the computations needed to fit the model. PCA is one of the well known efficient dimensionality reduction techniques. in this tutorial we will use PCA which compresses the data by projecting it to a new subspace that can help in reducing the effect of the curse of dimensionality. Our dataset consists of 4 dimensions(4 features) so we will project it to a 3 dimensions space and Plot in one 3d graph.

Dimensionality Reduction



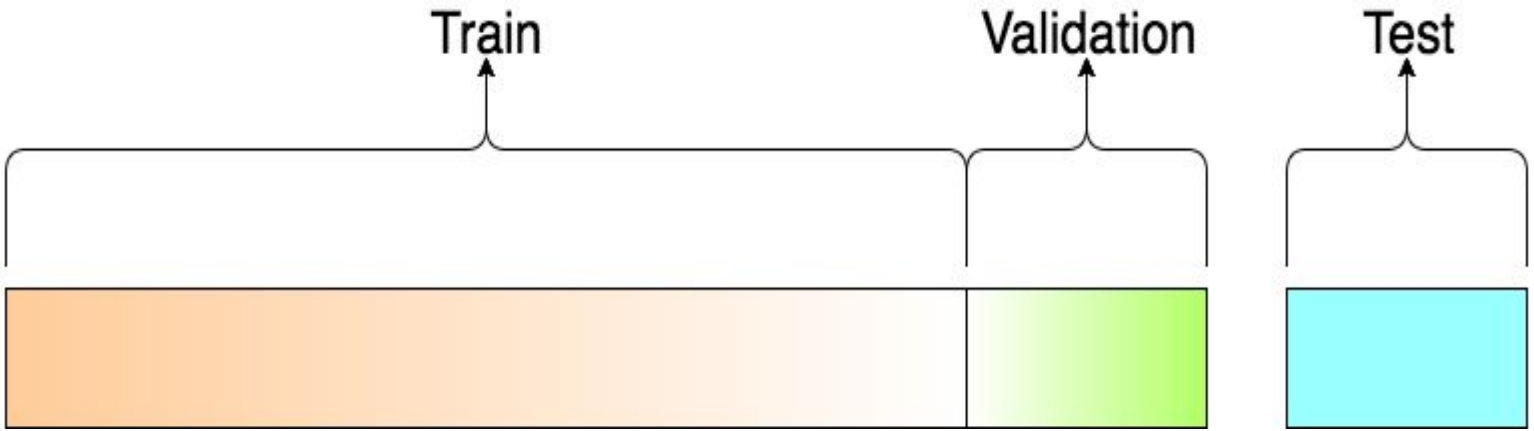
Splitting Data

The train test split function is for splitting a single dataset for two different purposes: training and testing. The testing subset is for building your model. The testing subset is for using the model on unknown data to evaluate the performance of the model.

The data we use is usually split into training data and test data.

Training Dataset: The sample of data used to fit the model.
Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.
Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Splitting data

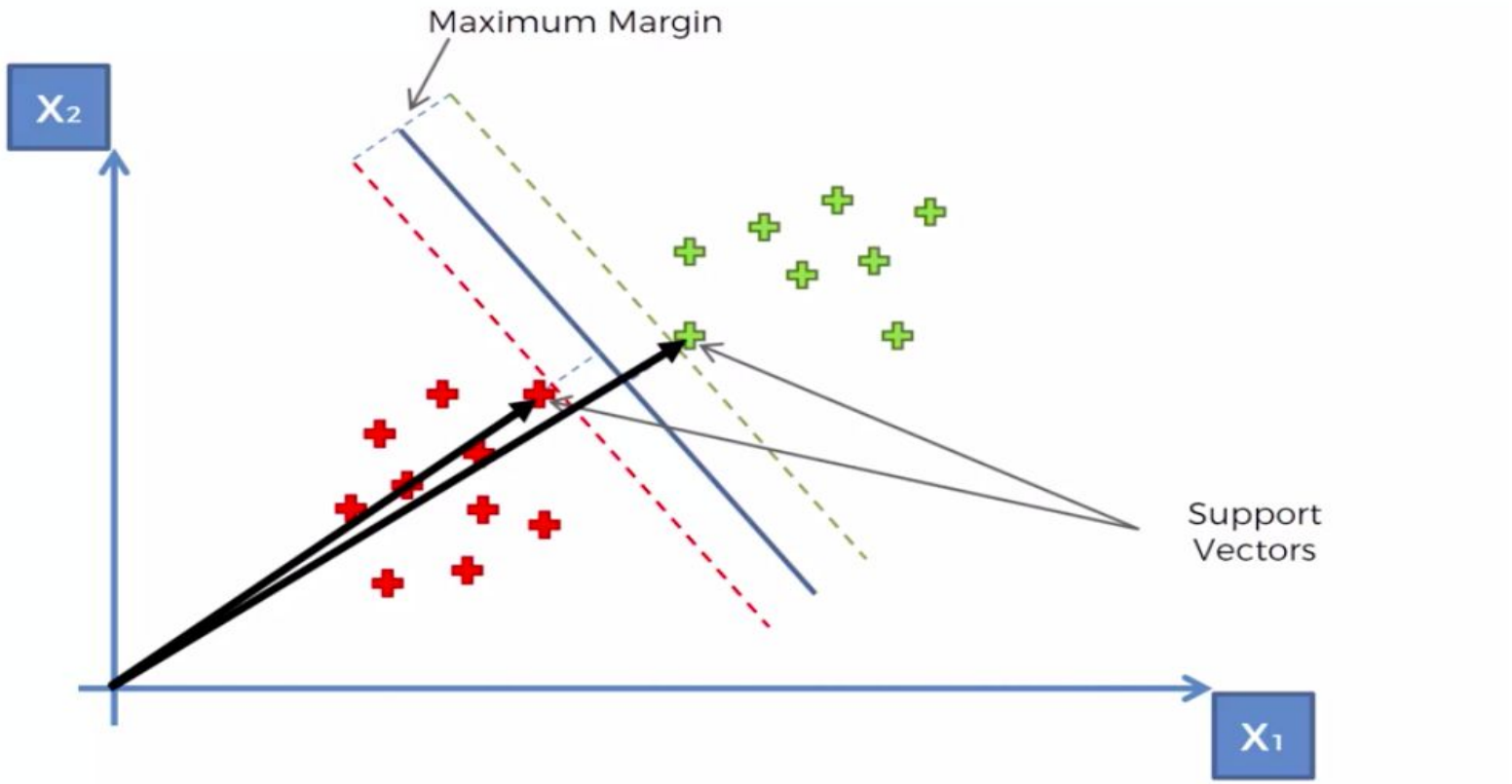


Building the Model - SVM

A support vector machine (SVM) is a type of supervised machine learning classification algorithm. SVMs were introduced initially in **1960s** and were later refined in **1990s**. However, it is only now that they are becoming extremely popular, owing to their ability to achieve brilliant results. SVMs are implemented in a unique way when compared to other machine learning algorithms.

Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes. All the data points that fall on one side of the line will be labeled as one class and all the points that fall on the other side will be labeled as the second. Sounds simple enough, but there's an infinite amount of lines to choose from. How do we know which line will do the best job of classifying the data? This is where the LSVM algorithm comes in to play. The LSVM algorithm will select a line that not only separates the two classes but stays as far away from the closest samples as possible. In fact, the "support vector" in "support vector machine" refers to two position vectors drawn from the origin to the points which dictate the decision boundary.

Building the Model - SVM



Accuracy of SVC with different kernels

Accuracy of the SVC Clasification with **Linear kernel** and no other adjust is:
0.9666666666666667

Accuracy of the SVC Clasification with **Polynomial kernel** and no other adjust is: **0.9333333333333333**

Accuracy of the SVC Clasification with **Radial Basis kernel** and no other adjust is:
0.9333333333333333

Accuracy of the SVC Clasification with **Sigmoid kernel** and no other adjust is: **0.8**