02450 Introduction to Machine Learning and Data Mining E21

Report 01 - Groupe 186

ABSTRACT

The objective of this report is to apply the methods we have learned in the first section of the course,"Data: Feature extraction, and visualization" on our own data set to get a basic understanding of our data prior to the further analysis (project report 2)

Bashar Bdewi S183356 Malaz Alzarrad S180424

Dataset Description: (Malaz)

Iris dataset is a multivariate dataset of three classes of Irises and it was collected by the American botanist Edgar Anderson (1935) and introduced by the British statistician and geneticist Ronald Fisher in his article published in 1936 ”The Use of Multiple Measurements in Taxonomic Problems” introducing linear-discriminant-function technique. Fisher’s paper is referenced frequently to this day for being such a classic in the field. Basically, discriminant analysis aims to produce a simple function that, based on four measurements, will classify a flower specie correctly. Instead of ’guessing’ it was the beginning of creating “predictors” to classify the samples in the dataset. The Iris data set is a best known and understood dataset and one of the most used to analyse data sets in statistics, data visualization, machine learning, etc The iris dataset is available online from University California Irvine’s (UCI) machine-learning repository of datasets (http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data).

Previous use and Summarize: (Malaz)

The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification and clustering examples and to test algorithms. Information about the original paper and usages of the dataset can be found in the UCI Machine Learning Repository – Iris Data Set (http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data).

Problem Definition: (Malaz)

This data set consists of the physical parameters of three species of flower — Versicolor, Setosa and Virginica. The numeric parameters which the dataset contains are Sepal width, Sepal length, Petal width and Petal length. In this data we will be predicting the classes of the flowers based on these parameters. The data consists of continuous numeric values which describe the dimensions of the respective features. We will be training the model based on these features.

Data issues (Malaz)

We import and get to know the data, so we find at this database contains 5 attributes. We obtained this dataset from Kaggle.

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. species: – Iris Setosa – Iris Versicolour – Iris Virginica

After Checking the data type we can divide these features in two groups: quantitative and categorical

-Quantitative features – Continuous : sepal length - sepal width - petal length - petal width .

-Categorical features - Discrete: species

Our target feature is categorical (classification problem).

Now we Check the data type of each column in python, we get so that there's no missing values.

Here we are checking if there is any inconsistency in the dataset

And as we see there are no null values in the dataset, so the data can be processed.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             150 non-null    int64
 1   SepalLengthCm  150 non-null    float64
 2   SepalWidthCm   150 non-null    float64
 3   PetalLengthCm  150 non-null    float64
 4   PetalWidthCm   150 non-null    float64
 5   Species        150 non-null    object
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB
```

We take a Peek at the Data

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

Id is the unique identifier for each flower. In this machine learning project, it will not help with our model' so we drop the Id column.

Her is Statistical summary using .describe()

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

Let's interpret the above statistical desciption of our dataset:

The descriptoin shows we have data with super low std (standard deviation).

the range of the SepalLengthCm is: 4.300000 - 7.900000

the range of the SepalWidthCm is: 2.000000 - 4.400000

the range of the PetalLengthCm is: 1.000000 - 6.900000

the range of the PetalWidthCm is: 0.100000 - 2.500000

Visualization: (Malaz)

The essence of a graph is the clear communication of quantitative information. The ACCENT principles emphasize, or accent, six aspects that determine the effectiveness of a visual display for portraying data (ApprehensionClarity-Consistency-Efficiency-Necessity-Truthfulness).
Exploratory Data Analysis (EDA) is a pre-processing step to understand the data. There are numerous methods and steps in performing EDA, however, most of them are specific, focusing on either visualization or distribution, and are incomplete. Here will understand, explore, and extract the information from the data to answer the questions or assumptions. So let's see various visual representations of the data to understand more about relationship between various features.

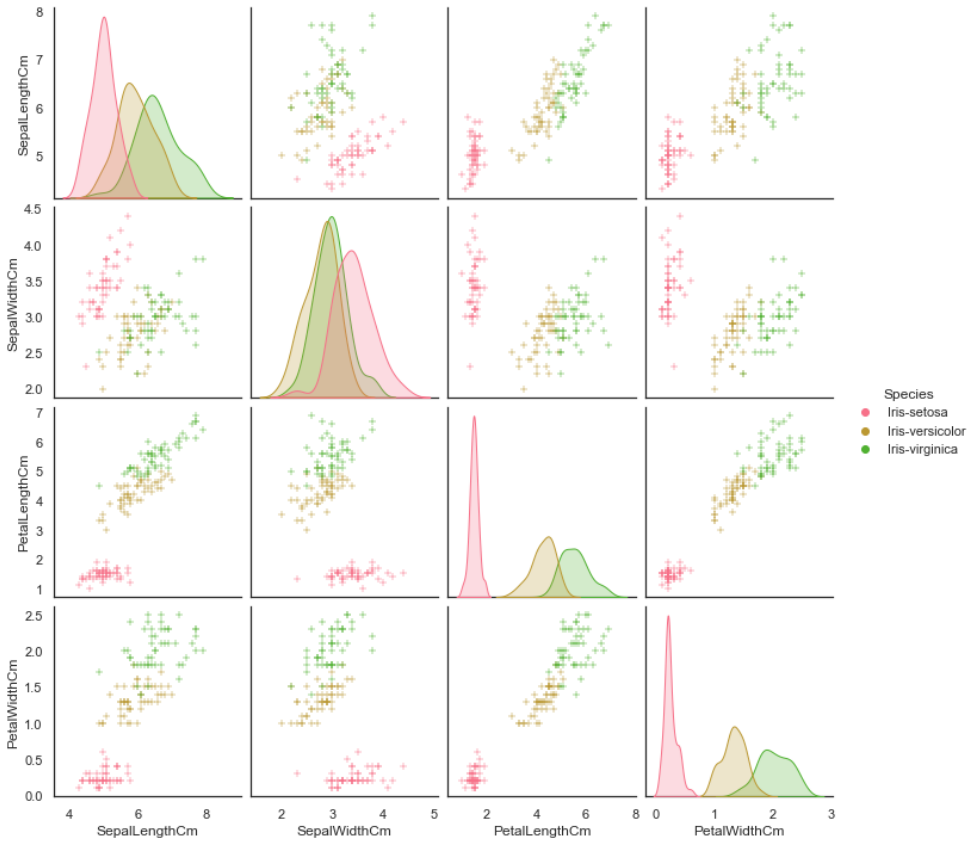
- There are 150 observations with 4 features each (sepal length, sepal width, petal length, petal width).

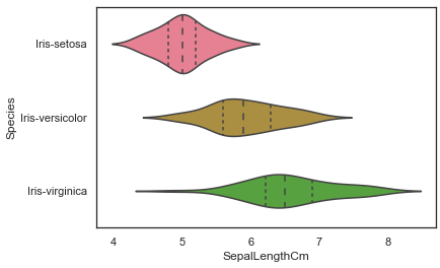
- There are no null values, so we don't have to worry about that.

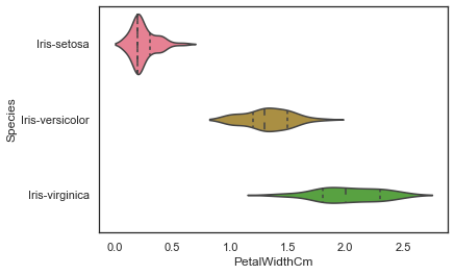
- There are 50 observations of each species (setosa, versicolor, virginica)

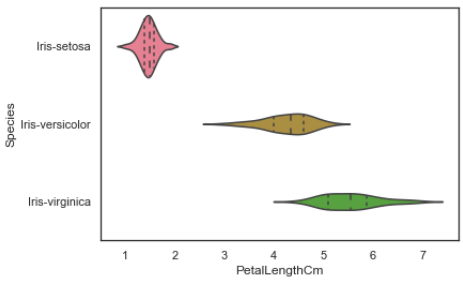
- After graphing the features in a pair plot, it is clear that the relationship between pairs of features of a iris-setosa (in pink) is distinctly different from those of the other two species.

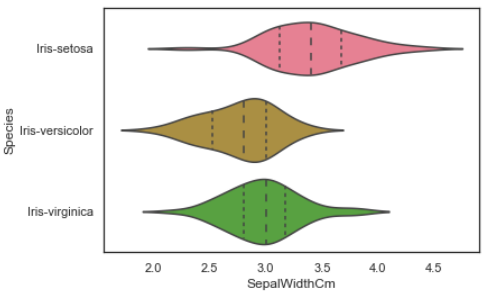
- There is some overlap in the pairwise relationships of the other two species, iris-versicolor (brown) and iris-virginica (green)


SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Species
Iris-setosa
Iris-versicolor
Iris-virginica


Iris-setosa
Iris-versicolor
Iris-virginica
Species
SepalLengthCm


Iris-setosa
Iris-versicolor
Iris-virginica
Species
PetalWidthCm


Iris-setosa
Iris-versicolor
Iris-virginica
Species
PetalLengthCm


Iris-setosa
Iris-versicolor
Iris-virginica
Species
SepalWidthCm

PCA (finds the principal components of data) (Bashar)

It is often useful to measure data in terms of its principal components rather than on a normal x-y axis. They're the underlying structure in the data. They are the directions where there is the most variance, the directions where the data is most spread out.

PCA finds a new set of dimensions (or a set of basis of views) such that all the dimensions are orthogonal (and hence linearly independent) and ranked according to the variance of data along them. It means more important principle axis occurs first. (more important = more variance/more spread out data)

PCA works :

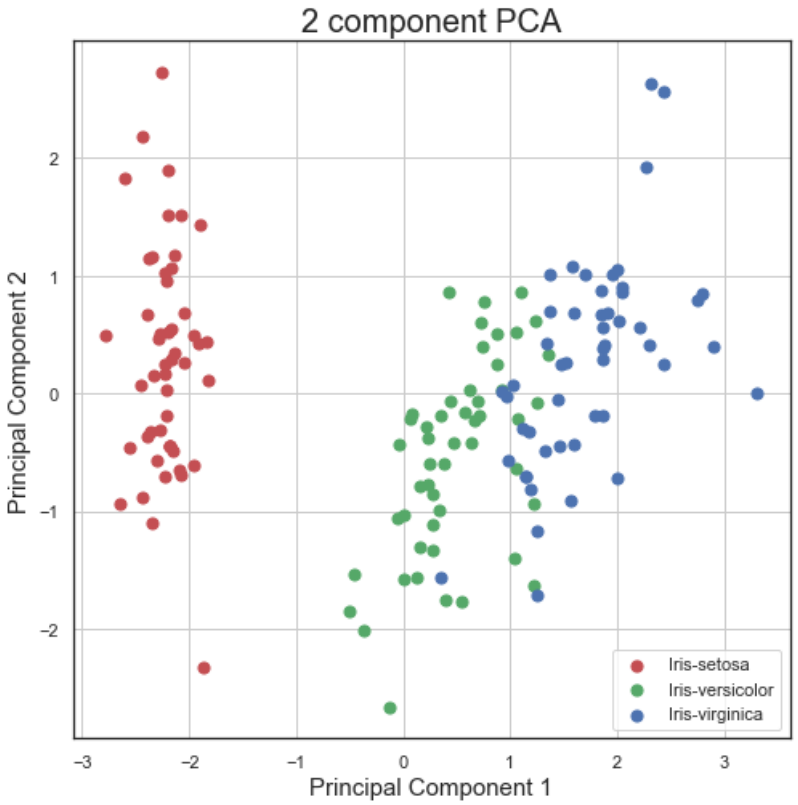
- Calculate the covariance matrix X of data points.
- Calculate eigen vectors and corresponding eigen values.
- Sort the eigen vectors according to their eigen values in decreasing order.
- Choose first k eigen vectors and that will be the new k dimensions.
- Transform the original n dimensional data points into k dimensions

For a lot of machine learning applications it helps to be able to visualize our data. Visualizing 2 or 3 dimensional data is not that challenging. However, even the Iris dataset used in this part of the is 4 dimensional. We can use PCA to reduce that 4 dimensional data into 2 or 3 dimensions so that we can plot and hopefully understand the data better.

PCA is effected by scale so you need to scale the features in your data before applying PCA. Use **StandardScaler** to help us standardize the dataset's features onto unit scale (mean = 0 and variance = 1) which is a requirement for the optimal performance of many machine learning algorithms.

The original data has 4 columns (sepal length, sepal width, petal length, and petal width). We projects the original data which is 4 dimensional into 2 dimensions. We should note that after dimensionality reduction, there usually isn't a particular meaning assigned to each principal component. The new components are just the two main dimensions of variation.

Now we are plotting 2 dimensional data. Notice on the graph below that the classes seem well separated from each other.



The explained variance tells us how much information (variance) can be attributed to each of the principal components. This is important as while we can convert 4 dimensional space to 2 dimensional space, we lose some of the variance (information) when we do this.

By using the attribute **explained_variance_ratio_**, we can see that the first principal component contains 72.77% of the variance and the second principal component contains 23.03% of the variance. Together, the two components contain 95.80% of the information.

```
In [2]: pca.explained_variance_ratio_
Out[2]: array([0.72770452, 0.23030523])
```

Explaining what we have learned : (Malaz and Bashar)

Malaz

I have learned a lot about data especially Data preparation, Data cleaning, Correcting data, Data Analysis and it was a good start to Machine Learning and data mining techniques to do our project . I learned how to start applying standard machine learning process, beginning with understanding what the data represented followed by handling it and how to perform Data Analysis using ACCENT Principles for effective graphical display and Exploratory Data Analysis (EDA) to understand how the features relate to the outcome.

Bashar

I have learned how to choose our dataset and scaling data, standardization and normalization . How to visualize data and get knowledge from it. I have learned one of the most dimensionality reduction techniques in machine learning and data science technology (PCA).

Conclusion

We have just implemented some of the common Machine Learning and described and explored the classic Iris dataset with data visualizations.

We covered: Loading the dataset- Summarizing the dataset - Visualizing the dataset.-Evaluating some algorithms