Imagine that you are an evolutionary biologist studying irises and that you have collected measurements on a large number of iris samples. Your goal is to identify different species within this collection.

Here is a classical iris dataset first analyzed by Fisher. We will upload the data in the form of a DataFrame -- similar to a spreadsheet -- where the columns are different measurements (or features) and the rows are different samples. Below, we show the first 5 lines of the dataset.
In [1]:
# Julia version: 1.5.1
using CSV, DataFrames, Statistics, Plots, LinearAlgebra, StatsPlots

In [2]:
df = CSV.read("iris-measurements.csv")
first(df, 5)

Out[2]: 5 rows × 5 columns

<table>
<thead>
<tr>
<th>Id</th>
<th>PetalLengthCm</th>
<th>PetalWidthCm</th>
<th>SepalLengthCm</th>
<th>SepalWidthCm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Int64</td>
<td>Float64</td>
<td>Float64</td>
<td>Float64</td>
<td>Float64</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.4</td>
<td>0.2</td>
<td>5.1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.4</td>
<td>0.2</td>
<td>4.9</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1.3</td>
<td>0.2</td>
<td>4.7</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1.5</td>
<td>0.2</td>
<td>4.6</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1.4</td>
<td>0.2</td>
<td>5.0</td>
</tr>
</tbody>
</table>

There are 150 samples.

In [3]:
nrow(df)

Out[3]: 150

Here is a summary of the data.

In [4]:
describe(df)

Out[4]: 5 rows × 8 columns

<table>
<thead>
<tr>
<th>variable</th>
<th>mean</th>
<th>min</th>
<th>median</th>
<th>max</th>
<th>nunique</th>
<th>nmissing</th>
<th>eltype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>Float64</td>
<td>Real</td>
<td>Float64</td>
<td>Real</td>
<td>Nothing</td>
<td>Nothing</td>
<td>DataType</td>
</tr>
<tr>
<td>1</td>
<td>Id</td>
<td>75.5</td>
<td>1</td>
<td>75.5</td>
<td>150</td>
<td>Nothing</td>
<td>Int64</td>
</tr>
<tr>
<td>2</td>
<td>PetalLengthCm</td>
<td>3.75867</td>
<td>1.0</td>
<td>4.35</td>
<td>6.9</td>
<td>Nothing</td>
<td>Float64</td>
</tr>
<tr>
<td>3</td>
<td>PetalWidthCm</td>
<td>1.19867</td>
<td>0.1</td>
<td>1.3</td>
<td>2.5</td>
<td>Nothing</td>
<td>Float64</td>
</tr>
<tr>
<td>4</td>
<td>SepalLengthCm</td>
<td>5.84333</td>
<td>4.3</td>
<td>5.8</td>
<td>7.9</td>
<td>Nothing</td>
<td>Float64</td>
</tr>
<tr>
<td>5</td>
<td>SepalWidthCm</td>
<td>3.054</td>
<td>2.0</td>
<td>3.0</td>
<td>4.4</td>
<td>Nothing</td>
<td>Float64</td>
</tr>
</tbody>
</table>

Let’s first extract the columns into vectors and combine them into a matrix, and visualize the petal data. Below, each point is a sample. This is called a scatter plot (https://en.wikipedia.org/wiki/Scatter_plot).

In [5]:
X = reduce(hcat,
            [df[:, PetalLengthCm], df[:, PetalWidthCm],
             df[:, SepalLengthCm], df[:, SepalWidthCm]]);
Observe a clear cluster of samples on the bottom left. This may be an indication that these samples come from a separate species. What is a cluster? Intuitively, it is a group of samples that are close to each other, but far from every other sample.

Now let's look at the full dataset. Visualizing the full 4-dimensional data is not so straightforward. One way to do this is to consider all pairwise scatter plots.
We would like a method that automatically identifies clusters -- whatever the dimension of the data. We will discuss a standard way to do this: \( k \)-means clustering.

This topic has two main goals:

1. To review basic facts about Euclidean geometry, vector calculus, probability, and matrix algebra.
2. To introduce a first data science problem and highlight some relevant, surprising phenomena arising in high-dimensional space.

We will come back to the iris dataset in an accompanying tutorial notebook.
Optional reading

You may want to review basic linear algebra and probability. In particular, take a look at

- Chapter 1 in [Sol] and Sections 1.2.1-1.2.4 in [Bis]

where, throughout this course, we will refer to the following textbooks available online:


Material for this lecture is covered partly in:

- Sections 1.4, 9.1 and 11.1.1 of [Bis]

References

Parts of this topic's notebooks are based on the following references.


