DATA SCIENCE & MACHINE LEARNING (part - 4)

techworldthink • March 06, 2022

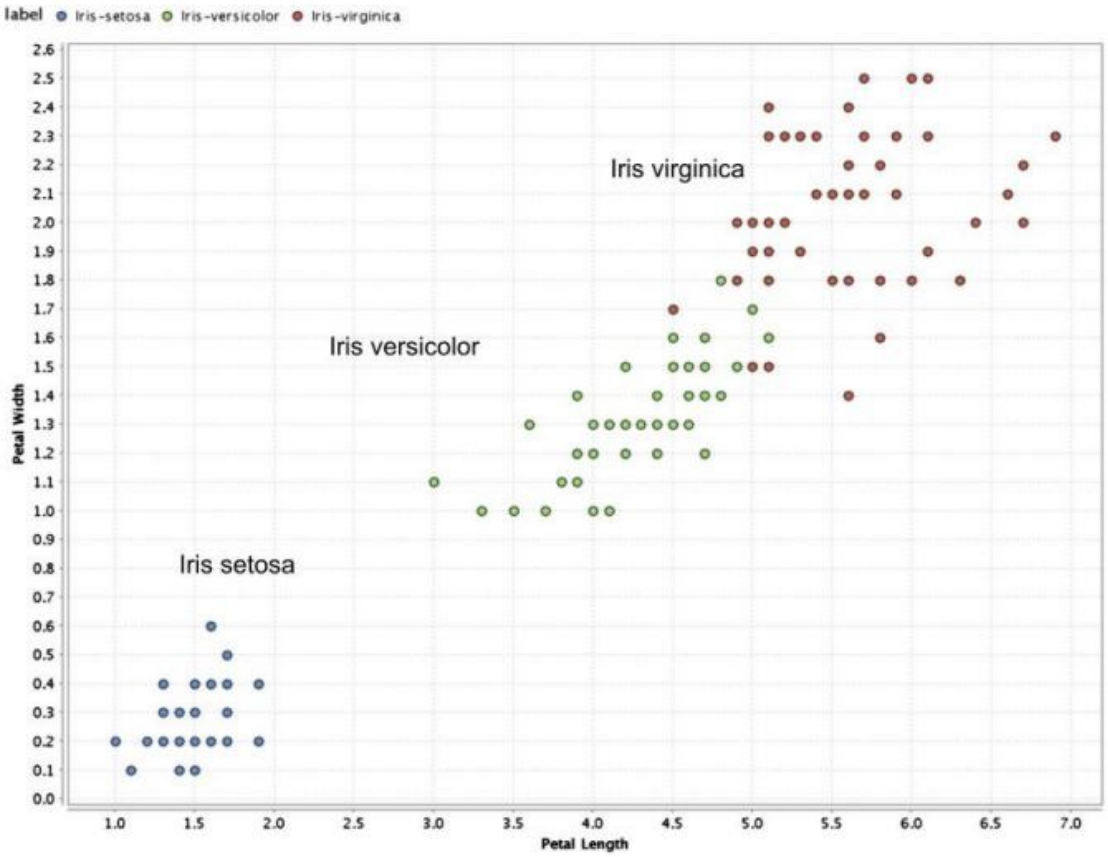
11. Explain the various methods for visualising multivariate data.

Multivariate exploration is the study of more than one attribute in the dataset simultaneously. This technique is critical to understanding the relationship between the attributes.

The multivariate visual exploration considers more than one attribute in the same visual.

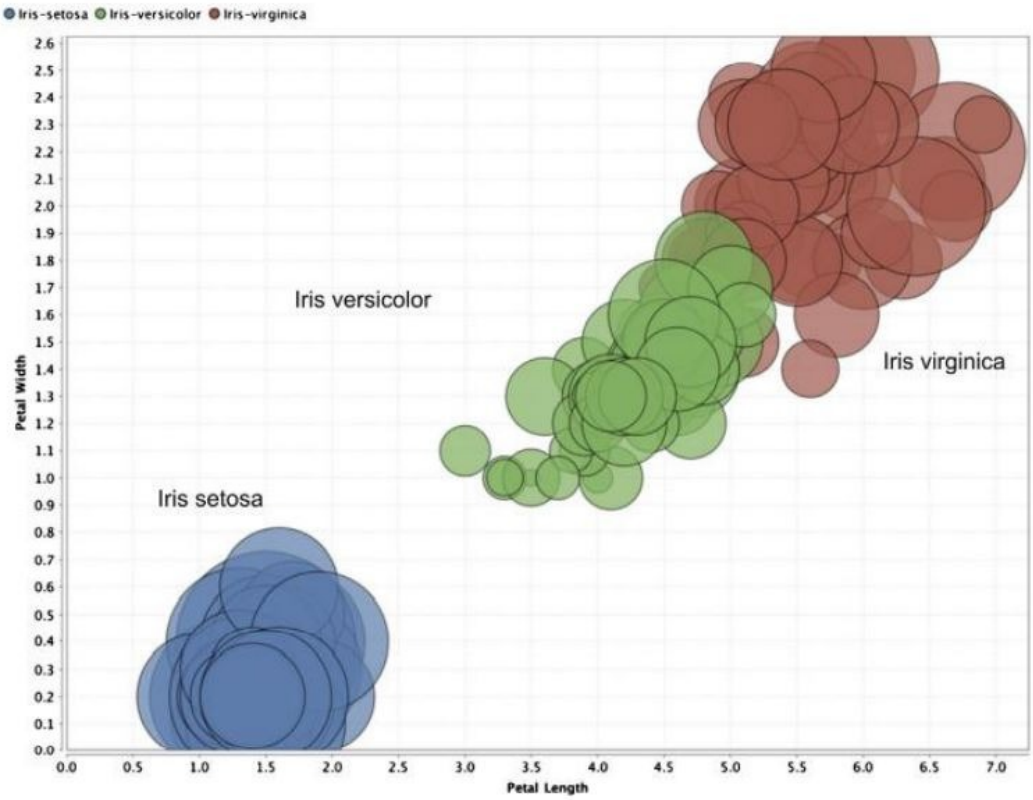
Scatterplot

A scatterplot is one of the most powerful yet simple visual plots available. In a scatterplot, the data points are marked in Cartesian space with attributes of the dataset aligned with the coordinates. The attributes are usually of continuous data type. One of the key observations that can be concluded from a scatterplot is the existence of a relationship between two attributes under inquiry. If the attributes are linearly correlated, then the data points align closer to an imaginary straight line; if they are not correlated, the data points are scattered. Apart from basic correlation, scatterplots can also indicate the existence of patterns or groups of clusters in the data and identify outliers in the data. This is particularly useful for low-dimensional datasets.



Bubble Chart

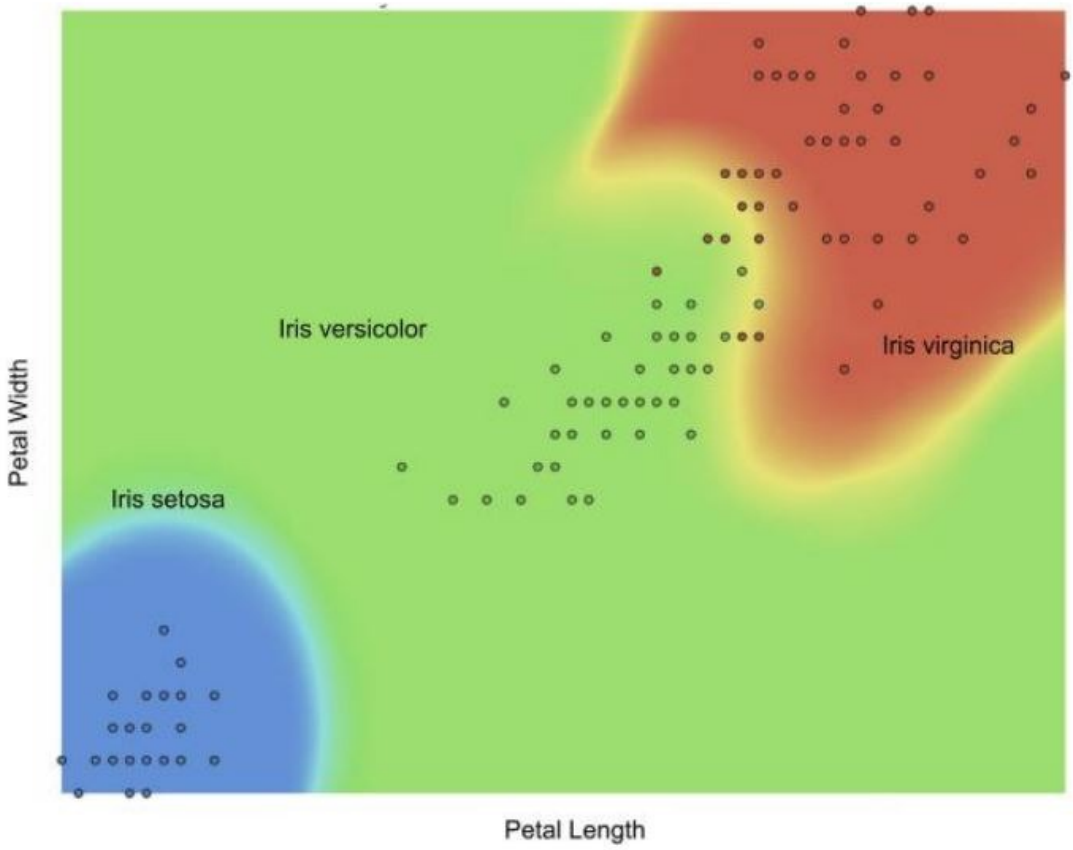
A bubble chart is a variation of a simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point. In the Iris dataset, petal length and petal width are used for x and y-axis, respectively and sepal width is used for the size of the data point. The color of the data point represents a species class label



Density Chart

Density charts are similar to the scatterplots, with one more dimension included as a background color. The data point can also be colored to visualize one dimension, and hence, a total of four dimensions can be visualized in a density chart.

here, petal length is used for the x-axis, sepal length for the y-axis, sepal width for the background color, and class label for the data point color.

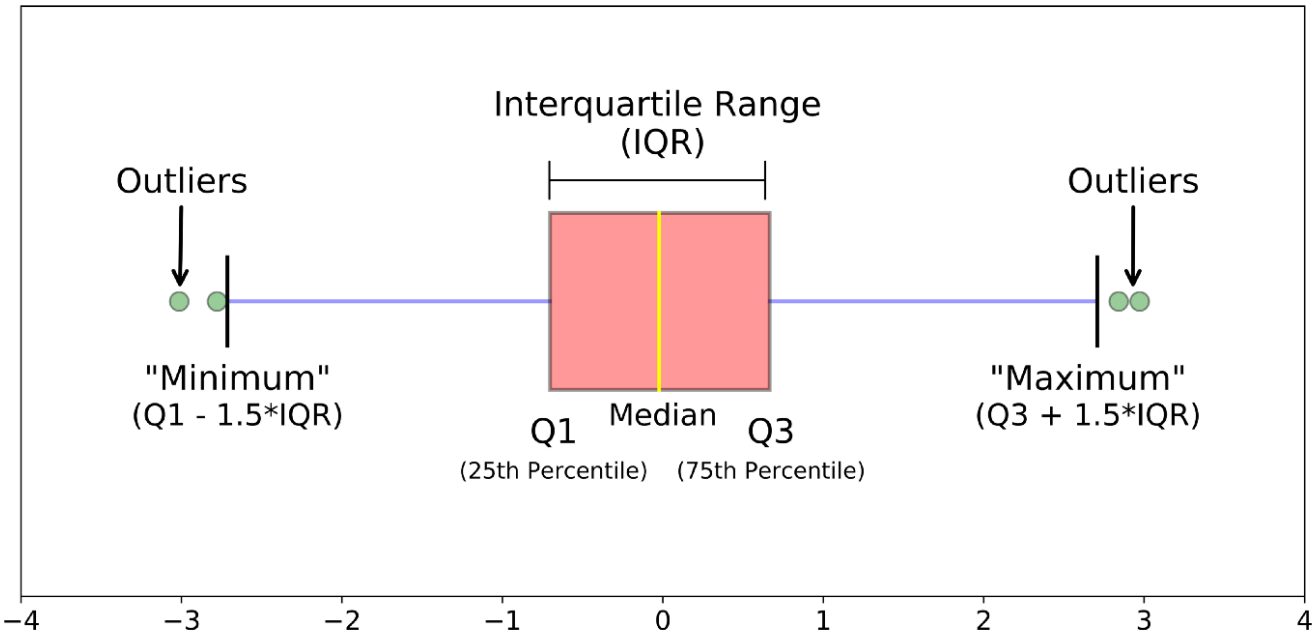


Boxplots

A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

For some distributions/datasets, you will find that you need more information than the measures of central tendency (median, mean, and mode).

You need to have information on the variability or dispersion of the data. A boxplot is a graph that gives you a good indication of how the values in the data are spread out. Although boxplots may seem primitive in comparison to a histogram or density plot, they have the advantage of taking up less space, which is useful when comparing distributions between many groups or datasets.



12. Explain the various processes for preparing a dataset to perform a data science task.

Preparing the dataset to suit a data science task is the most time-consuming part of the process. It is extremely rare that datasets are available in the form required by the data science algorithms. Most of the data science algorithms would require data to be structured in a tabular format with records in the rows and attributes in the columns. If the data is in any other format, the data would need to be transformed by applying pivot, type conversion, join, or transpose functions, etc., to condition the data into the required structure.

Data Exploration

Data Quality

Missing Values

Data Types and Conversion

Transformation

Outliers

Feature Selection

Data Sampling

Data Exploration

Data preparation starts with an in-depth exploration of the data and gaining a better understanding of the dataset. Data exploration, also known as explor atory data analysis, provides a set of simple tools to achieve basic understand ing of the data. Data exploration approaches involve computing descriptive statistics and visualization of data. They can expose the structure of the data, the distribution of the values, the presence of extreme values, and highlight the inter-relationships within the dataset. Descriptive statistics like mean, median, mode, standard deviation, and range for each attribute provide an easily readable summary of the key characteristics of the distribution of data. On the other hand, a visual plot of data points provides an instant grasp of all the data points condensed into one chart

Data Quality

Data quality is an ongoing concern wherever data is collected, processed, and stored. Errors in data will impact the representativeness of the model. Organizations use data alerts, cleansing, and transformation techniques to improve and manage the quality of the data and store them in companywide repositories called data warehouses. Data sourced from well-maintained data warehouses have higher quality, as there are proper controls in place to ensure a level of data accuracy for new and existing data. The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bounds, standardization of attribute values, substitu tion of missing values, etc. Regardless, it is critical to check the data using data exploration techniques in addition to using prior knowledge of the data and business before building models.

Missing Values

One of the most common data quality issues is that some records have missing attribute values. There are several different mitigation methods to deal with this problem, but each method has pros and cons. The first step of managing missing values is to understand the reason behind why the values are missing. Tracking the

data lineage (provenance) of the data source can lead to the identification of systemic issues during data capture or errors in data transformation. Knowing the source of a missing value will often guide which mitigation methodology to use. The missing value can be substituted with a range of artificial data so that the issue can be managed with marginal impact on the later steps in the data science process. Missing credit score values can be replaced with a credit score derived from the dataset (mean, minimum, or maximum value, depending on the characteristics of the attribute). This method is useful if the missing values occur randomly and the frequency of occurrence is quite rare. Alternatively, to build the representative model, all the data records with missing values or records with poor data quality can be ignored. This method reduces the size of the dataset. Some data science algorithms are good at handling records with missing values, while others expect the data preparation step to handle it before the model is inferred. For example, k-nearest neighbor (k-NN) algorithm for classification tasks are often robust with missing values. Neural network models for classification tasks do not perform well with missing attributes, and thus, the data preparation step is essential for developing neural network models.

Data Types and Conversion

The attributes in a dataset can be of different types, such as continuous numeric (interest rate), integer numeric (credit score), or categorical. For example, the credit score can be expressed as categorical values (poor, good, excellent) or numeric score. Different data science algorithms impose different restrictions on the attribute data types. In case of linear regression models, the input attributes have to be numeric. If the available data are categorical, they must be converted to continuous numeric attribute. A specific numeric score can be encoded for each category value, such as poor 5 400, good 5 600, excellent 5 700, etc. Similarly, numeric values can be converted to categorical data types by a technique called binning, where a range of values are specified for each category, for example, a score between 400 and 500 can be encoded as “low” and so on.

Transformation

In some data science algorithms like k-NN, the input attributes are expected to be numeric and normalized, because the algorithm compares the values of different

attributes and calculates distance between the data points. Normalization prevents one attribute dominating the distance results because of large values. For example, consider income (expressed in USD, in thousands) and credit score (in hundreds). The distance calculation will always be dominated by slight variations in income. One solution is to convert the range of income and credit score to a more uniform scale from 0 to 1 by nor malization. This way, a consistent comparison can be made between the two different attributes with different units

Outliers

Outliers are anomalies in a given dataset. Outliers may occur because of correct data capture (few people with income in tens of millions) or erroneous data capture (human height as 1.73 cm instead of 1.73 m). Regardless, the presence of outliers needs to be understood and will require special treatments. The purpose of creating a representative model is to generalize a pattern or a relationship within a dataset and the presence of outliers skews the representativeness of the inferred model. Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

Feature Selection

In practice, many data science problems involve a dataset with hundreds to thousands of attributes. In text mining applications, every distinct word in a document forms a distinct attribute in the dataset. Not all the attributes are equally important or useful in predicting the target. The presence of some attributes might be counterproductive. Some of the attributes may be highly correlated with each other, like annual income and taxes paid. A large number of attributes in the dataset significantly increases the complexity of a model and may degrade the performance of the model due to the curse of dimensionality. In general, the presence of more detailed information is desired in data science because discovering nuggets of a pattern in the data is one of the attractions of using data science techniques. But, as the number of dimensions in the data increase, data becomes sparse in high-dimensional space. This condition degrades the reliability of the models, especially in the case of clustering and classification (Tan, Steinbach, & Kumar, 2005). Reducing the number of attributes, without significant loss in the performance of the model, is

called feature selection. It leads to a more simplified model and helps to synthesize a more effective explanation of the model.

Data Sampling

Sampling is a process of selecting a subset of records as a representation of the original dataset for use in data analysis or modeling. The sample data serve as a representative of the original dataset with similar properties, such as a similar mean. Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling. In most cases, to gain insights, extract the information, and to build representative predictive models it is sufficient to work with samples. Theoretically, the error introduced by sampling impacts the relevancy of the model, but their benefits far outweigh the risks. In the build process for data science applications, it is necessary to segment the datasets into training and test samples. The training dataset is sampled from the original dataset using simple sampling or class label specific sampling. Consider the example cases for predicting anomalies in a dataset (e.g., predicting fraudulent credit card transactions). The objective of anomaly detection is to classify the outliers in the data. These are rare events and often the dataset does not have enough examples of the outlier class. Stratified sampling is a process of sampling where each class is equally represented in the sample; this allows the model to focus on the difference between the patterns of each class that is, normal and outlier records. In classification applications,sampling is used create multiple base models, each developed using a different set of sampled training datasets. These base models are used to build one meta model, called the ensemble model, where the error rate is improved when compared to that of the base models.