Anna Gillies Theory Answers

1.1 In your own words, what does the role of a data scientist involve?

1.1 The role of a data scientist involves the compilation of data from multiple sources to test hypotheses, ideas, and create models to influence conclusions and decisions

1.2 What is an outlier? Here we expect to see the following:

a. Definition

b. Examples

c. Should outliers always be removed? Why?

d. What are other possible issues that you can find in a dataset?

1.2a. An outlier is a data point that significantly differs from other observations within the dataset
1.2b. An example of this could be in the following dataset: 200, 400, 600, 800, 12394832. In this example, the last point would be an outlier in the set. Outliers are often considered an outlier if they lie either before the interquartile range or above the third quartile range.

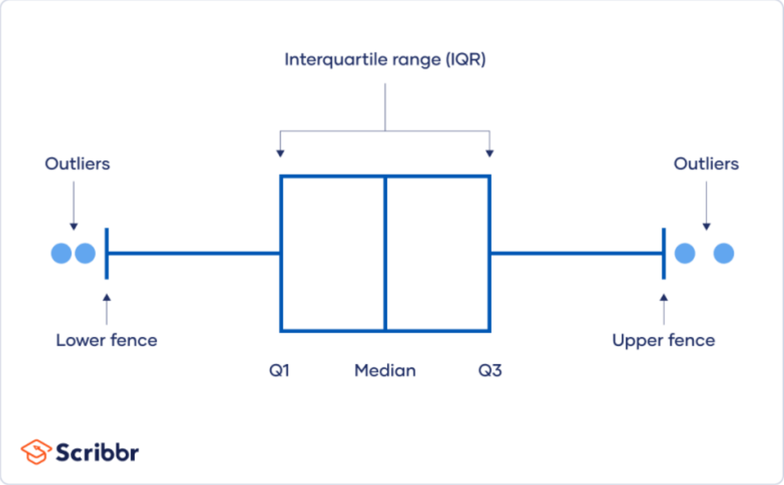

Image credit: https://www.scribbr.co.uk/stats/statistical-outliers/

1.2c. Outliers should not always be removed. Outliers should be removed when they are as a result of misentered data, when it's irrelevant to the analysis and skews data inappropriately, or when it has a significant impact to performance. Outliers should typically be retained when it's a genuine data-point, if it provides insightful points, or if it is a genuine predictive value. Outliers should always be carefully evaluated and assessed when deciding on their treatment.
1.2d. Other issues you can find in your dataset include : missing data, duplicate records, incorrect data, and inconsistent data.

1.3 Describe the concepts of data cleaning and data quality. Here we expect to see the following:

a. What is data cleaning?

b. Why is data cleaning important?

c. What type of mistakes do we expect to commonly see in datasets?

1.3a. Data cleaning is the process of identifying and removing/correcting errors and inconsistencies in a dataset to improve its quality and prepare it for analysis.
1.3b. Data cleaning is important as it: improves the quality of data, reduces the risk of machine model errors, improves the accuracy of analysis, and enhances decision making
1.3c. In datasets, the most common mistakes are: missing data, duplicate records, inconsistent data, and incorrect data.

1.4 Discuss what is Unsupervised Learning - Clustering in Machine Learning using an example. Here we expect to see the following:

a. Definition.

b. When is it used?

c. What is a possible real-world application of unsupervised learning?

d. What are its main limitations?

1.4a. Clustering is an unsupervised machine learning technique used to identify and group similar data points within large datasets without focusing on a specific outcome. Also known as cluster analysis, clustering is typically employed to organise data into more comprehensible and manageable structures. Unsupervised learning is where, without human intervention, the machine analyses and runs itself.

1.4b. Clustering is used in situations when the data does not have pre-defined labels or categorisation. It is used in: exploratory data analysis, anomaly detection, data preprocessing, and pattern recognition

1.4c. A typical example of unsupervised learning - clustering is within customer analysis and segmentation. Organisations can use clustering to segment customers based on their purchasing behaviour to identify spending categories and loyal customers. This can help with customer marketing strategies and campaigns, improving sales.

1.4d. Unlike supervised learning, unsupervised learning can present less accurate results as it is not measured against labeled data. Furthermore, some clustering techniques, such as hierarchical clustering can be intensive and may not be effective with large datasets.

1.5 Discuss what is Supervised Learning - Classification in Machine Learning using an example. Here we expect to see the following:

a. Definition.

b. When is it used?

c. What is a possible real-world application of supervised learning?

d. What data do we need for it? Is there any processing that needs to be done?

1.5a. Supervised learning involves training a model on a labeled dataset, where each training example is paired with an output label. Classification, a type of supervised learning, aims to have the model predict the categorical label of new instances by identifying patterns learned from the labeled training data. Classification can be split in to a further two classifications, binary (two outcomes) or multi (more than two outcomes).

1.5b. Is it used in various instances when the objective is to assign a predetermined category to new observations. For example, classifying spam and non-spam emails for spam filtering.

1.5c. Real world examples include: spam filters, image recognition, and diagnoses. A further example could include training a model on different dog breeds. If the model is taught using a list of dogs with their breeds, the model can then figure out the characteristics to associate with each dog. After the learning period is over, the model can then accurately identify dog breeds.

1.5d. Data needs to be labeled with a label or category. For example, with spam detection, this could involve emails being labelled as 'spam' or 'not spam'. Processing is required, including data cleaning, transformation, and normalisation to ensure accurate inputs to enable accurate outputs.