AI Notes

Notes on *Chapter 2: Data, Measurements, and Data Preprocessing*

Author: Abdullah Yassine

September 1, 2025

Contents

1 Data Types

Data object is an entity. Objects can be things like employees at a company, customers at a business, and so on. The **attributes** that describe those objects are a feature of the object. They describe something of the object. Salary is an attribute of an employee, the number of items bought is an attribute of a customer, and so on.

Attributes can have different ranges of values. We explore them here.

1.1 Nominal attributes

Nominal attributes are *names of things*. They cannot be ordered, but are categorized. Things like the colors of an eye, types of t-shirts, and so on.

Those attributes can also be turned into numbers, like 0 for black, 1 for red, and so on. But mathematical operations on these numbers are meaningless. It does not mean anything when you take the mean for example. You can take the mode which we take about later.

1.2 Binary attributes

A **binary attribute** only have two values, 0 or 1, or True or False.

A binary attribte can be **symmetric** if there is no preference on who gets the 1 or 0, like male or female.

An **assymetric** binary attribute is where it's not equally important. Like *positive* or *negative* of a disease.

1.3 Ordinal attributes

An **ordinal attribute** is where you can rank them, but the magnitude of difference between successive values aren't known. This is things like small, medium, or large, but we can't tell how much larger a large is from a medium. Because you can order them, we can take the mode and median, but not the mean.

1.4 Numeric attributes

A **numeric attribute** is measurable quantity.

Interval-scaled attributes are measured on equal-sized scale. They can be zero, positive, or negative, and they do have an order. We can also quantify the difference between values here. For example, temperates of 10 and 50 Celsius can have a difference of 40, but you cannot have ratios here or that they don't have meaning.

Ratio-scaled attributes have an inherent zero-point scale, so you can take ratios and they do have meaning. If you take temperature in Kelvin, this is where this comes. Other things include counts of things like *years of experience* or *number of words* are ratio-scaled.

1.5 Discrete vs. continous attributes

A **discrete attribute** can have limited number of values or can be countably infinite. It can numeric or otherwise. Things like *color of eye*, *sizes of drinks*, and so on. They can numeric, for example, 0 and 1 for binary attributes. Countably infinite just means the range of values can be infinite, but those values can be "listed" one by one, like *customer_id* or zip codes.

Continous is just infinite, if it's not discrete.

2 Statistic of Data

2.1 Measuring the central tendancy

Central tendancy is what measures the middle of things so we can have a picture of the middle of the distribution. First, we have the **mean**:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_n}{N}$$

Sometimes, we want to give weights to some of the elements signifying their importance. If so, we use the following **weighted average**:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

Sometimes, the mean is not the best tool because it's very sensitive to outliers. We can use **trimmed mean** and just ignore those outliers.

The next tool we have is the **median** which is the value that seperates the left half from the right half given the list is ordered.

The next tool we have is **mode**, which is the value that occurs the most often in the distribution. Sometimes, a distribution can have multiple modes, so we call the distribution **multimodal**.

Another tool is the **midrange** which can be used to assess the central tendancy of the dataset. It's the average of the largest and smallest values of the list. Look at the following

picture to see the difference between symmetric and asymmetric or skewed data:

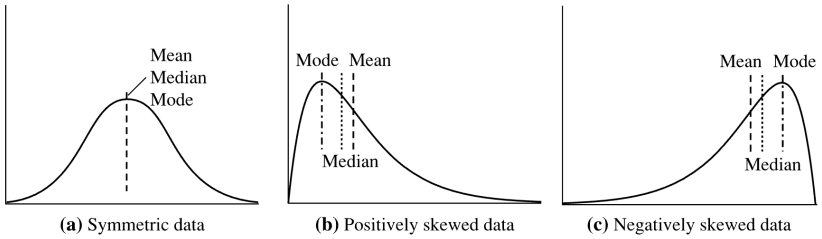

Figure 1: Difference between symmetric and assymetric

3 Measuring the Dispersion of Data

We now try to measure the spread of data. We begin with **range**, which is the just the difference between the largest value and the smallest value in the set.

Then, we have **quantiles** which are points taken at regular intervals that essentially divide the set into equal parts.

The 4-quantiles are the ones that divide the data set into four equal parts, they are also called **quartiles**. We also have the **percentiles** which divide the data into 100 equal-sized parts.

Back to the quartile, the **first quartile** is the 25% percentile, which cuts off the lowest 25%. You also have the **third quartile**, which cuts off the lowest 75%. The distance between the first and third quartiles is simply the range of the middle half of the data. This is called **interquartile range** (IQR):

$$\text{IQR} = \text{Q}_3 - \text{Q}_1$$

If the data set is odd, then you exclude the median element and divide up the set into the left half and the right half then calculate median again. If even, you must include them again and divide into left and right half while including them.

Another way of identifying **outliers** is to see if the suspected outlier is above $1.5 \times \text{IQR}$ of the third quartile or below the first quartile.

Thus, we have the **five-number summary** which consists of *minimum* (smallest value), Q_1, median, Q_3, *maximum*, and this is represented by **boxplots**.

Boxplots have following properties:

- Ends of boxes is quartiles.

- Middle part is median.
- The lines are extended to the smallest and largest observations.

If outliers are beyond $1.5 \times$ IQR, we draw them as points as individuals. When that happens, we draw the actual $1.5 \times$ IQR as well as the individual points:

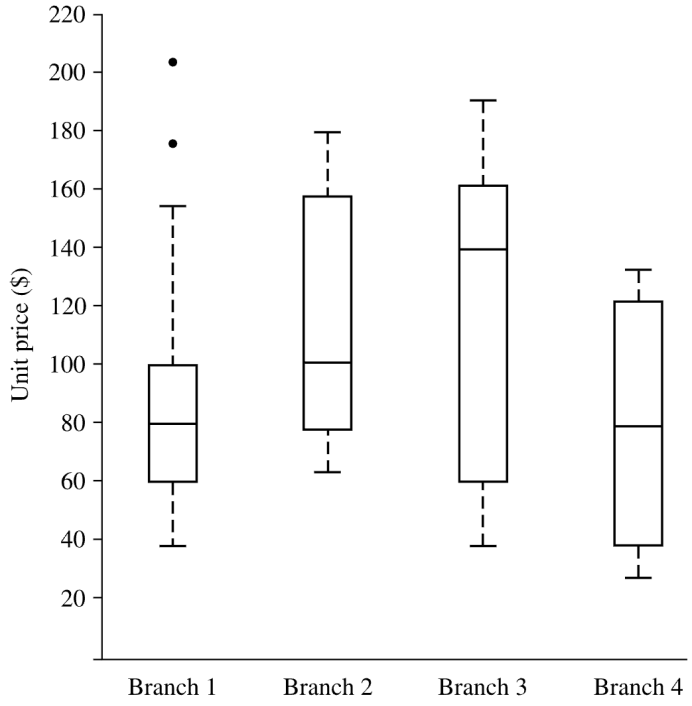

Figure 2: Example of boxplot

3.1 Variance and standard deviation

Variance is calculated as following:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

Standard deviation (σ) is just the square of variance. SD is zero when there is no spread, otherwise it's positive.

4