Introduction to Machine Learning and Data Mining

**Part 1 Data: Types, Features and Visualization**

Introduction

This chapter is about:

* Defining machine learning and data mining.
* Machine learning methods.
* Implement the right machine learning methods based on the situation and then task.

What is machine learning and data mining?

* Data mining
  + The process of extracting useful information from a huge amount of data. It is used by humans to discover new, accurate, and useful patterns in data or meaningful relevant information for the ones who need it.
* Machine learning
  + The process of discovering algorithms that have improved courtesy of experience derived data. It is the algorithm that permits the machine learning without human intervention. It’s a tool to make machine smarter, eliminating the human element.
* Machine learning tasks
  + Supervised learning
    - Task is to predict a quantity based on other quantities.
      * Classification
        + Here we are given observed values x and have to predict a discrete response y, we are given discrete observations of some object and have to determine what class the object belong to.
      * Regression
        + In regression we are given observed values x and have to predict a continuous response y.
  + Unsupervised learning
    - The machine uses unlabeled data and learns on itself without any supervision.
      * Clustering
        + Clustering algorithm means you’re going to give the algorithm a lot of input data with no labels and let it find any groupings in the data it can.
      * Density estimation
        + In density estimation we try to quantify how likely a given future observation is given past observations, the probability distribution of the data set.
      * Anomaly detection
        + Anomaly detection is figuring out which observations significantly deviate from other observations.
  + Reinforcement learning
    - Corresponds to the case where a computer has to control a robot based on sensory input and a reward signal.

Data and attribute types

* What is data?
  + Data, or a dataset, is a collection of electronically stored information.
* Dataset
  + Temporal: if for instance the observation is made sequentially over time.
  + Varying number of features: if each observation has a varying complexity. For instance, the text files have a varying number of words, and the 3d models a varying number of surface points.
  + Self-referential structure: For instance, in the social graph, the edges refer to the same group of people.
* Attributes
  + Nominal: If the variable is not ordered and only uniqueness matters. An example is the country of origin or the id.
  + Ordinal: If the variable is ordered.
  + Interval: If the variable is ordered and the relative magnitude of the variable has a physical meaning.
  + Ratio: If the value 0 of the variable has a specific, physical meaning.
* Data issues
  + Irrelevant or spurious attributes: The ID column is irrelevant as it only depends on the ordering of the data.
  + Outliers: The safety training of -100 must be due to some kind of error. We call such observations outliers.
  + Missing data: The Miles/Gallon attribute for car 4 is missing.

Problems

2.1

The correct answer is B: For both Age and PV there are a natural zero and we can apply all the operators <, >, =, not equal, \*, / thus these attributes are ratio.

2.2

The correct answer is C: All the attributes are ratio since 0 means absence of what is being measured.

2.3

The correct answer is D.