**1) [5 points]** Give a **concrete example** of Data Object in the real word. Think about the attributes/features and **list all** attributes you can think of. A nice draw of a simple table will be awesome.

For this example, I will use a house as the data object as a concrete example of a real-world object. In this scenario, the attributes would include:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Price** | **Lot size (ft.)** | **House Size (ft.)** | **Car Space** | **Baths** | **Rooms** | **Pool** | **Zip** | **Walk score** | **Bike score** |
| $500,000 | 5000 | 3500 | 2 | 3.5 | 4 | No | 98027 | 30 | 70 |
| $400,000 | 4500 | 2400 | 1 | 2 | 3 | Yes | 98040 | 98 | 100 |

Some other attributes might be latitude/longitude, photos, address, State, City, description, type (townhouse, condo, etc.), view, year built, equipment, utilities.

**2) [5 points]** Briefly (1-2 sentences) justify your answers to the below questions.

1. You have developed an algorithm to automatically and efficiently retrieve addresses and phone numbers from a very large database, using a person’s name as the search key. Is this machine learning? Why or why not?

**ANSWER:** No, this is most likely not machine learning as this example is more likely a simple query search from the database using SQL or a programming language search function. Machine learning would require the machine to learn about the features within the dataset and use statistics to generate predictions, or classifications for the data.

1. You have developed an algorithm to automatically and efficiently retrieve all similar images (could be ranked by similarity scores or probabilities) to a query image from a very large image database. Is this machine learning? Why or why not?

**ANSWER:** No, this is not a machine learning problem as this would just be a SQL query but just for images. In this query, the algorithm would check pixel by pixel to see the similarities and then return the closest pictures that match the similarity scorecard.

**3) [10 points, 2.5 each]** For each of the following scenarios, state which type of machine learning would be appropriate: classification, regression, clustering, or association rule analysis. Briefly (1-2 sentences) justify your answers.

1. You are a manager of a local Safeway store, and sales data indicates that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. You want to make more benefit from this by using promotional pricing or product placements.

**ANSWER:** Association rule analysis – for this example, I believe that the model would be association rule analysis as the store wants to determine what customers buy based on data they have showing prior customer transactions. This is typical for associative rule analysis as it receives many transactions and shows the relationship between certain data objects and other data objects in terms of transactions (ex for transactions a, b, and c: {a, b -> c}).

1. A dietician has been trying to understand how people’s dietary choices affect the amount of weight they gain or lose, but isn’t seeing obvious patterns. For a recent 6-month period, he has good records for 150 of his clients on their consumption of 12 different foods, along with the change in their weight over that period.

**ANSWER:** Regression – for this example, regression makes sense as the data model would be able to take in many different features and find certain trends that a human would be unable to see. In this case, using regression to see predictions for future clients.

1. You are having an argument with your friend about how many social groups there are at your school. You believe there are about half-a-dozen natural groups based on tastes in things like music, clothes, athletics, and politics, while your friend thinks everyone’s tastes are pretty random. You discover you can access (publicly available) individual records from a poll where 1000 students scored their preferences on 20 forms of arts and entertainment.

**ANSWER:** Clustering – for this example, clustering makes a lot of sense as you want to see the similarity between the certain friends in groups. By seeing which friends connect more often, using clustering would show you who hangs out with who the most due to their closeness/similarity.

1. You work at an oil company, and they are interested in predicting whether wells drilled in several new formations will produce oil or not. They give you a large quantity of data from past drilling efforts (geographic location, depth of well, type of rock, age of formation, etc.), along with the success or failure of each drilled well.

**ANSWER:** Classification - with this example, classification seems to make the most sense as the company would want to know if they should drill or not. By giving an classification machine learning algorithm data that is important, it could determine if the odds of drilling will result with oil.

**4) [4 points]** Describe a situation in your life where machine learning might be of benefit. For example, this could be something you deal with at work, at school, or on the internet (e.g. a social networking site). Say as much as you can about the problem to be solved, the data or information you might collect, and the type of machine learning you think is applicable. (There isn’t necessarily a correct answer to this question; I just want you to start being able to recognize opportunities to apply machine learning.)

**ANSWER:** I feel that a machine learning algorithm for what my pay would be at a company would be incredibly useful when I find a job. The problem is what my salary would be given the amount of time, title, my work statement, and other factors, what my pay would be at the company. This would help me negotiate my salary, prepare for what would maximize my salary, and career path decisions. I would see that regression would be the most likely machine learning method for this situation.

**5) [6 points, 2 each]** Term project proposal

1. Which dataset and problem are you planning to work on for the term project? Why you are interested in this problem?

Initially, I was thinking of doing a machine learning classification model on wine to determine what would qualify the quality of wine (on a scale of 1-10). Additionally, I had an idea of doing Zillow housing prices in Seattle based on data from Zillow by using regression. However, now that others also have this project, I want to check with them to see what we can do to prevent doing the same project. If I am unable to complete these two ideas, I would then work on a classification model for predicting if people have more than $50K in income or less based on census data. All three of these topics are interesting to me as I think they would be challenging but doable within the time frame of the class. Additionally, I think it would be fun to see what makes a “quality wine” or find the price of any house in the area. Finally, I think the census data really is encompassed with “big data” and could provide really cool insights.

For b) and c), I will answer those with the wine example as I did the most research prior to class on that subject.

1. What kind of useful information/knowledge/pattern do you expect to discover from that dataset?

For the wine example, I would expect to be able to see what quality the wine actually is and that certain patterns exist within the data that I would be unable to see. For example, would the acidity of the wine or age impact the quality? I would think that this information would impact that score and it would be interesting to see what the model does with the features that exist for the example.

1. Why you think your work will be valuable? (What is the potential future application of your work, or who will be your potential customers?)

I think this work will be valuable since to me, I know nothing about wine and it would be great to be able to pick out great wines for my friends. Additionally, since the data looks really clean, it would be great to start my efforts in machine learning with an example I could understand a little easier. For my potential customers, I feel that this machine learning model could help them with buying hidden gems and then being able to buy cheap and sell high. Finally, I think getting this working will look great on my resume/cover letter and it could be a great example to have on GitHub (if that is possible).