Assignment 0
Date assigned: Jan. 9, 2019
Date due: Jan. 16, 2019

1) [10 points] You have developed a fast algorithm for retrieving 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?

This is not machine learning, because the algorithm does not learn from the data and is explicitly programmed. The algorithm does not improve over time, take insight from past searched, reveal any unseen data, or make estimations. It is not used as a tool to understand large data, it can only find what is there independent to other data.

2) [10 points each] For each of the following scenarios, state which type of machine learning would be appropriate: classification, regression, clustering, or a Pattern discovery. Briefly (1-2 sentences) justify your answers.

    a) 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 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.

Clustering would be the best machine learning type to use here because clustering is used to dividing a dataset into homogeneous groups and sometimes for Segmentation analysis. In the case of this scenario, the model will divide the 1000 students into alike groups which follows the goals of the clustering model.

    b) 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.

To predict wither a new formation will produce oil the classification machine learning would be appropriate. This is because the model assigns the class labels (oil or no oil) to previously unseen records as accurately as possible using past data. Classification best fits the goal to predict wither a formation will produce oil.

    c) The florist in your neighborhood has a pretty good idea what kinds of flowers arrangements her steady customers like, but she'd like to be more scientific and send personalized coupons for her customers. She has excellent records of all their past purchases, as well as complete histories of which arrangements they have viewed on the shop's website

Pattern type of machine learning works best in this situation, because this is an example of market basket analysis. The pattern type machine learning will reveal useful associations within the data to help reveal which customers are more likely to be interested in specific coupons.

d) 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.

Regression best fits this scenario because the model should predict value of response variable on previously unseen records as accurately as possible using past data. Regression is a model that can predict response variable as a function of the values of the patient's records.

e) A coffee shop manager in UW Bothell started to collect transactions record to find information that can help predict whether a customer is likely to be lost to a competitor coffee shop on Campus.

Classification will work best because the goal of the coffee shop manager is to classify a status of new customers as easily lost or not easily lost. This aligns with the goals of classification to assign class labels to previously unseen records as accurately as possible.

3) [30 points] A retailer company want to reduce cost of mailing by targeting a set of customers likely to buy a new cell-phone product.

1. What is the unit of observation?

The unit of observations is the customers because it is used to describe the smallest entity with measured properties of interest for a study.

2. What are the examples and probable features (5 minimum)?

The customer's features that may be useful for this scenario include annual income, age, days since last phone purchase, percent of income spent on technology/smart devices, city of residence, number of children, job, hobbies.

3. What is the type of each feature?

Annual income, age, days since last phone purchase, percent of income, number of children are all numerical type. City of residence, job, and hobbies is categorical/nominal.

4) [10 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.)

I work at University of Washington Bothell and check out video equipment (cameras, lighting kits, microphones, etc…) to students. When a student does not return the equipment within the one-week deadline a hold is put on their account until they return it. It would be useful to know how many days are expected to pass before a student returned their equipment or if it will be overdue. We could then know when we would get items of equipment back in stock and if we will run out soon.

For this example, I believe that regression could help predict the number of days before the equipment is returned. The students would be the examples and features could be the number of past holds, residential distance from school, number of days on campus (class schedule), involvement on campus, age, class standing, GPA, how often they check out equipment, history of checkouts and returns (days), other non-IT related holds/tickets (library books, parking), major.

Another machine learning type that is applicable, but less informative would be to just find out if they are likely return the equipment late. This falls closer to the classification type and would use the same example (students) and features.