Patrick Austin

CS 491: Data Mining, Homework 1

25 September 2017

1-1 a. Sampling and feature subset selection are similar insofar as they are both preprocessing techniques used to select a subset of the total data to be analyzed. That is, both techniques aim to limit the amount of data that is to be analyzed to some subset of the whole dataset.

Where sampling and feature selection differ is the strategy used to select the subset of data. Sampling techniques operate by taking samples of the whole set of data according to some strategy, such as random sampling or stratified sampling, with the intent being that the sample will retain the properties of interest of the whole set of data. These representative samples can then be analyzed to produce a result that is very close to what the outcome would have been for analyzing the whole of the data, but at a much lower cost in terms of time and hardware usage.

In contrast, feature selection reduces the amount of data being considered by simply excluding some of the features of the data from consideration. Often, some features can be excluded intuitively, by common sense or knowledge of the domain of the problem. There are also more complex approaches, embedded, filter, and wrapper selection, that can be used to find less obvious features that can be excluded.

So, in summary, both sampling and feature selection seek to reduce the total dataset under consideration by reducing it to a subset of the original. Sampling does this by choosing a subset of the data entries (i.e. records in a piece of record data), and feature subset selection does it by choosing a subset of the data fields (i.e. attributes in a piece of record data).

b. Feature subset selection and dimensionality reduction are both techniques for preprocessing a dataset to reduce its complexity. This reduction enables datamining algorithms to achieve better results more quickly, with fewer hardware demands. Moreover, both operate by reducing the dimensionality of the data, the number of data attributes under consideration.

However, feature subset selection and dimensionality reduction employ different approaches to reach this goal. As discussed in part a., feature selection excludes data attributes from consideration entirely- it determines a subset of the original attributes that should be considered and removes the others completely.

In contrast, dimensionality reduction is specifically used to refer to techniques that create new attributes that are combinations of the original attributes. For example, a PCA dimensionality reduction technique uses linear algebra to create new attributes that are linear combinations of a number of original attributes, seeking to capture the maximum amount of variation in the data possible. These new attributes are then considered. Unlike in feature subset selection, attributes were not simply removed from consideration; instead, they were combined to make new attributes.

In summary, both techniques reduce the dimensionality of the data under consideration. But where feature subset selection does this by removing features outright, dimensionality reduction does this by creating a smaller number of attributes that are combinations of the originals.

c. If decimal scaling is used to normalize, then the new value of x is 480 / 104 = **.048**, and if min-max normalization is used the new value of x is

( 480 - (-100) ) / ( 9990 - (-100) ) \* ( 1 - (-1) ) + (-1) =

( (580/10090) \* 2 ) - 1 = **-.885**.

1-2 a. If asked to sample 10% of 10 million Twitter accounts via stratified sampling, I would require more information to be specified about the groups or types of accounts that need to be represented in the sample. With that information about the required groups of objects specified, I could then make sure that the sampling process chooses an equal number of accounts from each of the groups, or draw a number of accounts from each group that is proportional to the size of the group in the overall dataset.

For example, if asked to make sure that the owners of the accounts come from all the different countries in the world, I could attempt to use profile or IP data to sample an equal number of accounts originating in each country. Or, I could draw a number of accounts for each country proportional to the number of users from that country in the overall data.

If asked to sample the 10 million accounts using sampling without replacement, a type of simple random sampling, I would simply choose 10% of the accounts from the original dataset for the sample at random, removing each chosen sample from the original dataset after it is chosen so it cannot be chosen again. That is, after randomly choosing the first account for the sample from the original 10 million, I would draw the next account for the sample at random from the remaining 9,999,999 accounts. This sampling method would prevent the possibility of the same account being put into the sample more than once.

b. If asked to work with the location data associated with the Twitter accounts and forced to account for the lack of location data in many user profiles, I would be faced with several options or approaches for addressing the problem.

Eliminating accounts that have no profile data available from consideration is one option. The viability of this approach would depend on the proportion of accounts lacking profile data. If only a few lack the data, then it may be possible to do good analysis with this approach. If many lack the data, it’s hard to imagine getting good results. My suspicion is that the proportion of accounts lacking this field is quite high, so I would consider other options instead. Eliminating location attributes from consideration does not seem viable to me, since we have been asked to work on location data.

Another option is to ignore missing values during analysis. Objects can be compared and considered in similarity in terms of the values that are available, ignoring the ones that are missing. However, since this approach also works best when the number of missing values is low, and I suspect the number of missing values would be high, I will instead consider another approach.

A final option is to estimate missing values. Depending on the specifics of the data available for these accounts, I am most optimistic about this approach. For example, if user IP information is available in the data, or user tweets can be checked for location metadata, then it seems likely to me that good estimates could be made in many cases without much ambiguity. For another approach, a user’s location could be estimated using the location of people he or she follows- the logic being that if I follow a large number of people who are definitely from Reno, then I am probably from Reno as well. There are likely to be many other approaches that could be considered as well, depending on available data. While I would need more specifics to endorse this approach with certainty, I suspect that Twitter user data is best suited to this approach among these options.

c. When considering these 10 million Twitter accounts, it may prove useful to have some means to determine highly similar accounts. For example, this may allow the removal of bots and alternate accounts from the sample, allowing analysis to more closely focus on legitimate individual users.

Again, a full accounting of options and best approaches would require somewhat more information about the properties of these accounts that are available in the dataset. If geolocation or IP data is available, then perhaps an argument could be made for considering accounts much more likely to be highly similar if they utilize the same IP or tweet from highly similar locations. Likewise, a measure of similarity could look at tweet contents, searching for similar phrases or use of hashtags.

Another important tool would be to make use of the users an account follows. Two accounts that follow all the same accounts should be very likely to be considered highly similar. In fact, this is a property that has been used in the real world to visually document networks of Twitter bots.

In all likelihood a robust approach to finding similar Twitter accounts would consider multiple or even all of these characteristics. More information about what is really meant by a similar account would also help narrow down the particulars. Should a father’s Twitter account be considered highly similar to his daughter’s due to geotemporal similarities, despite quite different content and behavior? Should two individuals interested in politics be considered similar due to both following numerous politicians, despite living on opposite ends of their country? What if they Tweet about highly similar topics, in highly similar ways? What if they don’t? These and other questions would have to be considered in more detail.

1-3 For axiom 1a., the difference of two sets will always be greater than or equal to 0 since a set must have 0 or more elements, so d(x,y) must be greater than or equal to 0.

For axiom 1b., d(x,y) only equals 0 when the sizes of x - y and y - x are both 0. This will only happen when x = y, because in this case the sets will be the same and so their difference will be the empty set, with a size of 0. Thus d(x,y) only equals 0 if x equals y.

For axiom 2, consider that d(x,y) = size(x - y) + size (y - x) and d(y,x) = size(y - x) + size(x - y). These have the same terms; their positions have merely been flipped. Thus

d(x,y) = size(x - y) + size(y - x) = d(y,x), so the axiom holds.

For axiom 3, see the next page.