Ching-Hsiang Chu

CSE 5243 Intr to Data Mining

chu.368@osu.edu

Homework 1

1. **IRIS dataset**
   a. Data observation
   This dataset has 5 attributes and 150 examples. Beside the attribute "class", all other four attributes are length or width, whose attribute level are Ratio.

```
>summary(iris_)
 sepal_length    sepal_width     petal_length    petal_width     class
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Iris-setosa    :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Iris-versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   Iris-virginica :50
Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

   b. Data preprocessing
   For this simple dataset without any missing value or various data type, we do not really need to perform any preprocessing. Also, the distribution of values of each attribute are fairly smooth. (i.e., not being skewed)
   The only thing we need to do is to discard the last attribute, which is the "class" we will be predicting later and do not need to be considered into the distance calculation in this homework. Moreover, we can normalize the data, so that every attribute contributes same amount of weight to the distance calculation.
   c. Distance calculation
   We apply two simple distance functions based on *Minkowski Distance*:

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r\right)^{\frac{1}{r}}$$

   <1> City block distance (L1 norm, r=1)
   <2> Euclidean distance (L2 norm, r=2)
   Specifically, if we have n examples, two $n \times n$ distance matrices will be built to store the distance between each pair of examples. Then we pick k smallest number for each example (not include the example itself) as the output.
2. **Income dataset**
   a. Data observation
   This dataset has 15 attributes and 3250 (or 24420) examples. The table below summaries these attributes (not include the "class").

Attr. Name	**Type**	**Values**	**Note**
Age	Ratio	17-90	
Workclass	Nominal	9 factors	68% are "Private"
Fnlwgt	Unknown	19847-1161363	unknown meaning
Education	Nominal	16 factors	
Education_cat	Nominal	16 numbers	
Marital_status	Nominal	7 factors	79% are either "Never-Married" or "Married-civ-spouse"
Occupation	Nominal	15 factors	
Relationship	Nominal	6 factors	
Race	Nominal	5 factors	

Gender	Nominal	2 factors	
Capital_gain	Ratio	0-99999	91.9% are zero
Capital_loss	Ratio	0-4356	95.9% are zero
Hour_per_week	Ordinal/Ratio	1-99	
Native_country	Nominal	39 factors	89.1% are "United-State"

b. Data preprocessing

i. Reduce Dimension

Based on our observation, we could discard "Fnlwgt" attribute since the meaning of this attribute is not clear and it is significant skewed. Moreover, "Education" and "Education_cat" illustrate same idea, so we can combine them. Finally, we only have 12 attribute to deal with.

ii. Missing values

In the given dataset, number of missing values are not a huge number. Considering we do not have good enough domain knowledge to make accurate prediction for those missing values, here we simply ignore those examples with missing values. After doing this, there remains 3020 examples, which is still 93% of data. Therefore, we use these examples to perform the distance calculation.

iii. Transformation

To make calculation work, we need transform the dataset to numbers. Since we will be predicting if the salary of given example is higher than 50K or not, so the assigned number here are based on this purpose. (i.e., if the given circumstance might earn more money, higher number will be assigned) Meanwhile, we also take a closer look at each attribute to see if we can further reduce their complexity.

<1>workclass: "Federal-gov", "Local-gov", "Never-worked", "Private", "Self-emp-inc", "Self-emp-not-inc", "State-gov", "Without-pay"

We can group these into four groups, higher number means it's probably paid more. (just my htpothesis)

0: {"Never-worked", "Without-pay"},
1: {"Self-emp-inc", "Self-emp-not-inc"},
2: {"Local-gov", "State-gov", "Federal-gov"},
3: {"Private"}

<2> education: As an example has higher education background (degree), higher score is assigned. Note that we group few values togerther:
{"Masters", "Prof-school"}, {"Some-college", "HS-grad"}, {"Assoc-acdm", "Assoc-voc"}

<3> Marital_Status: As an example is not single at the moment, higher score is assigned. (Probably has higher salary)

0: {"Widowed"}
1: {"Divorced", "Separated", "Never-married", "Married-spouse-absent"}}
2: {"Married-AF-spouse", "Married-civ-spouse"}

<4> Relationship: Similar to marital status, we made assumption that a person with family may earn more money.

0: {"Other-relative", "Not-in-family", "Unmarried"}
1: {"Own-child", "Wife", "Husband"}

<5> **Native_Country**: We put the the corresponding GDP ranking[1] for them, which can somehow show how different they are in term of the money they can make.

<6> Occupation, Race, Gender: Do not have good hypothesis to measure how different they are, so we simply assign different numbers to them.

iv. Normalization

We apply the simple normalization scheme to make the value of each attribute locate between 0 and 1.

$$Norm = \frac{data - \min{(data)}}{\max{(data)} - \min{(data)}}$$

Note that we do not normalize the attributes, whose values do not have meaning (i.e., Occupation, Race, Gender) since it is meaningless. We will describe how to calculate distance for these attributes in next section.

Based on previous observation, **capital_gain** and **capital_loss** has majority of zero the and also are skewer. We took following steps, to prevent those outliers dominate the normalized values.

<1> Ignore zeroes, calculate the median of rest of examples, said *med*.

<2> Any value larger than the *med*, we replace it to 1.

<3> The rest of examples perform the usual normalization calculation as shown above.

In this way, we basically divide it into three parts: <1> zero, <2> low (0~1), <3> high (1).

c. Distance calculation

As in previous IRIS dataset, we apply same distance functions for Income dataset. As mentioned in previous section, there are attributes (i.e., Occupation, Race, Gender), whose values do not have special meaning but just to be used to differentiate with each other. For these attributes, the distance between any two examples is simple 0 or 1. Others are just similar to what we described for IRIS dataset.

---

[1] "GDP (Official Exchange Rate)". CIA World Factbook. Retrieved August 24, 2015.