CS659: HW1

02/15/2015

1. Given the data set found in the file z

(a) Use the first 99 records to compute the covariance matrix $\sum$

Answer: Denote the first column and second column as X and Y, then $Cov(X,Y) = \sum(X_i - \bar{X})(Y_i - \bar{Y})/N$,

$$Cov(X,Y) = \begin{bmatrix} 1.729 & 0.837 \\ 0.837 & 1.599 \end{bmatrix}$$

(b) The centroid of those 99 records, which is computed as a record whose attributes values are the average of the other record values

Answer: The centroid is (3.965 4.960).

(c) Using the result of a), compute

(i) The mahalanobis distance of records 100, 101 and 102 to the centroid computed in b)

Answer: The mahalanobis distance between two nodes x and y is $D = \sqrt{(x-y)^T S^{-1}(x-y)}$ where S is the co-variance metrix.
Thus the mahalanobis distance for records 100, 101, and 102 are: 3.409, 4.982, 5.449.

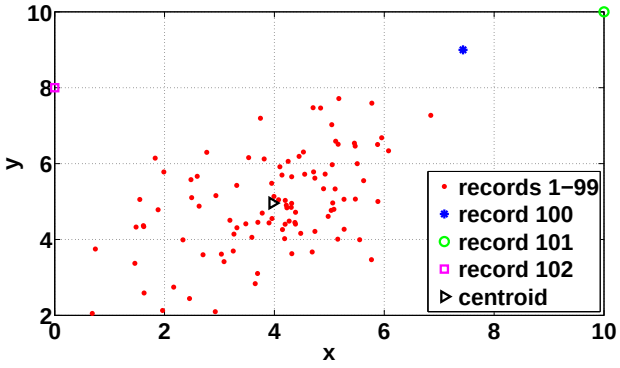
(ii) Repeat i) but compute the Euclidean distance

Answer: The Euclidean distance between two nodes x and y is $D = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.
Thus the Euclidean distance for records 100, 101, and 102 are 5.323, 7.863 and 4.997 respectively.

(iii) Compare and explain the differences between the results in i) and ii) submit a graph and your brief comments.

Answer: From the figure, we can see that record 101 have smaller Mahalanobis distance than record 102, but larger Euclidean distance than record 102. The reason is because Euclidean distance does not consider covariance of the attributes while Mahalanobis distance does. And the variance along the axis of record 101 is larger than the variance along the axis of record 102.
The figure is shown in Figure 1. From the figure, we can see that record

Figure 1: Visualization of z.txt.



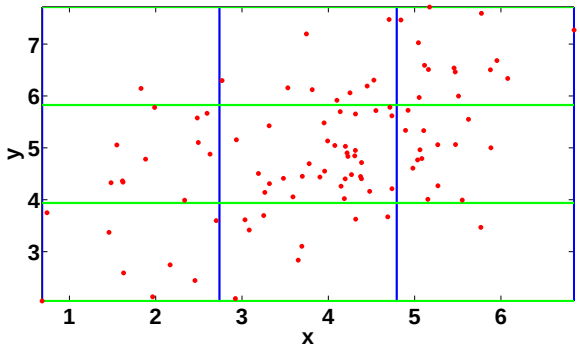
(d) Use the first 99 records, produce a data set with nominal(categorical) attributes, where each of the two attributes has 3 values. Use equi-width, and equi-depth for this task.

Answer: We use (x,y) to denote a data point.

For **equal-width**, the x values are divided into 3 bins, they are [0.683950 2.738300],(2.738300, 4.792650],(4.792650, 6.847000] respectively; while the y values are divided into the following 3 bins [2.049900 3.937867],(3.937867, 5.825833],(5.825833, 7.713800]. The results are shown in Figure 2, with blue line shows the division of x, and green line shows the division of y.

For **equal-depth**, the values are divided into bins such as that each bind contains the same number of instances. The x values are divided into the following 3 bins, [0.683950 3.610804],(3.610804, 4.698594],(4.698594, 6.847000] The y values are divided in to the following 3 bins, [2.049900 4.408182], (4.408182 5.567526],(5.567526, 7.713800]. The results are shown in Figure 3, with blue line shows the division of x, and green line shows the division of y.

Figure 2: Equal-width



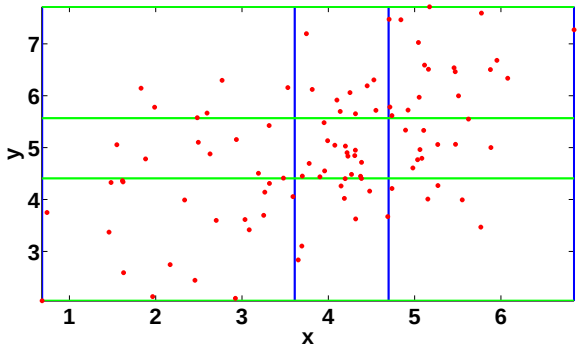
2. Using weka.

(a) Data Preparation and integration

Answer: The steps I use for the data integration.

(i) Set merged file relation as *hungarian-cleveland-heart-disease*

Figure 3: Equal-depth



(ii) Check attributes. Note that attributes name "chest_pain" and "cp" from these two files refer to same attributes.

(b) Description data summarization

(i) Study for example the age attributes. What is the mean? Its standard deviation? Its min and max?

Answer: Mean is 51.146
standard deviation is 9.08
min is 28, max is 77.

(ii) Provide the **five-number summary** of this attribute. Is this figure provide in Weka?

Answer: The five-number summary is: [min,first quartile, median,third quartile,max] = [28, 44.0, 52.0, 58.0, 77]
The five-number summary is not provided in Weka.

(iii) Switch to the **Visualize** tab. What is the term used in the textbook to name the series of boxplots represented? By selecting the maximum jitter, and looking the **num** column, can you determine which attributes seem to be the most linked to heart disease? Paste the **boxplot** representing the attribute you find the most predictive of heart disease (Y) as a function of **num**(X). Does any pair of different attributes seem correlated?

Answer: It has the name **scatter plot matrix**.
The attribute most linked to heart disease is **slope**. The boxplot is shown in Figure 4.
Attribute **exang** and **slope** seem correlated as shown in Figure 5.

(c) Data Preparation and cleaning

(i) Remove the missing values with the method of your choice, explaining which filter you are using and why you make this choice. Document what you did in your homework solution.

Answer:
I use filter **RemoveWithValues**, in the setting of this filter, **splitPoint** is used to filter numeric values, instances with values smaller than given value will be selected; **invertSelection** is used to invert the selection; **matchMissingValues** is sued to match

Figure 4: boxplot of num and slope attributes

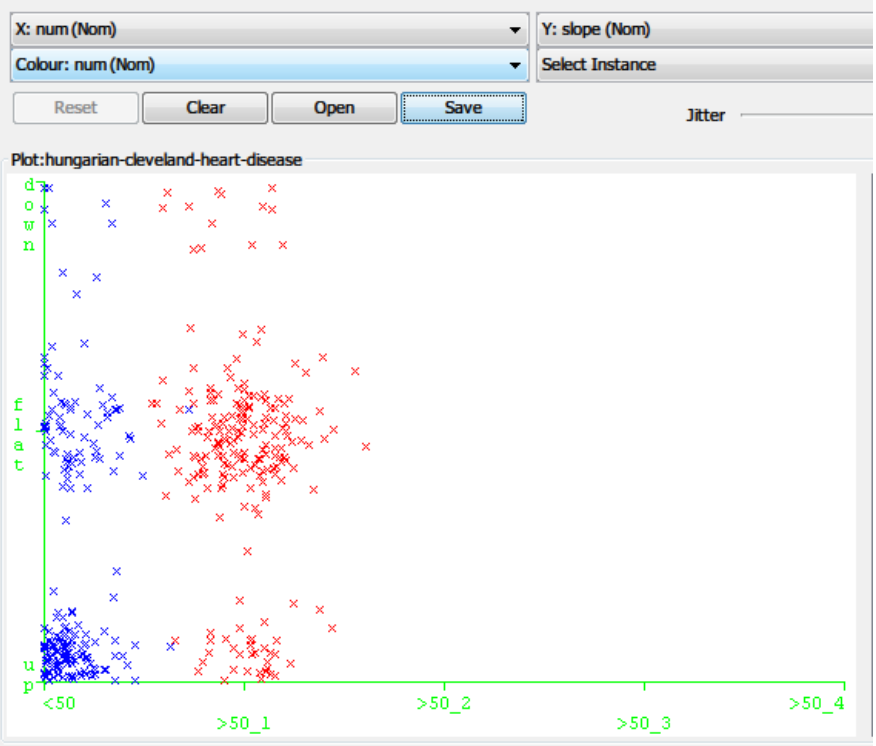
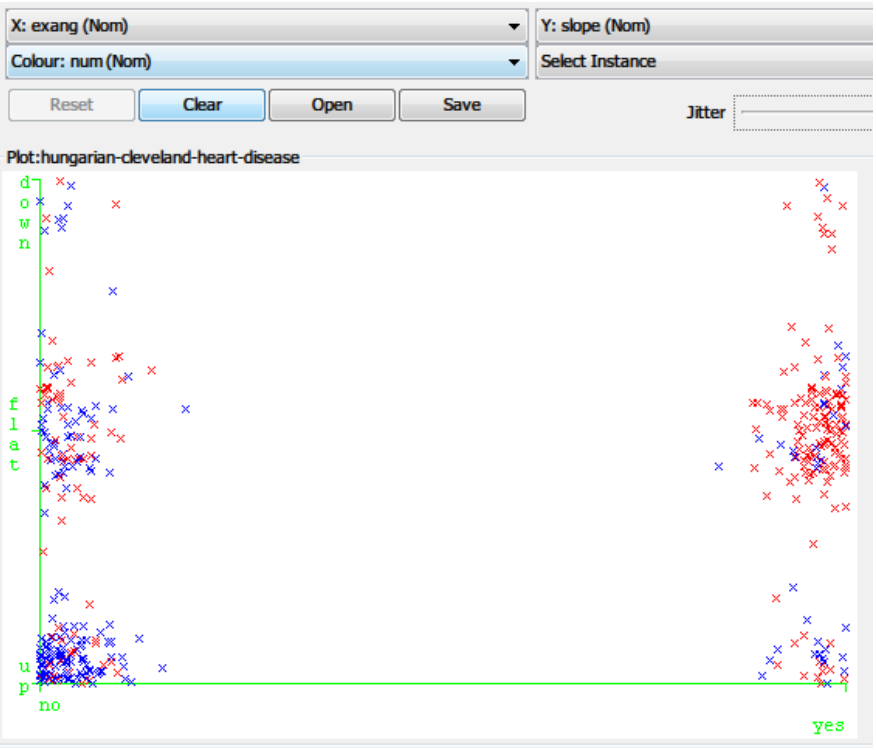


Figure 5: boxplot of exang and slope attributes



instances with missing values. Note that it is independent of **invertSelection**, i.e., **invertSelection** will not affect the selection result of **matchMissingValues**.

I takes the following steps to filter the missing value instances:

- step 1: Choose **RemoveWithValues** filter, select the attributes which contains the most missing instances (attribute ca, index 12), set **matchMissingValues** as true, the final setting is, "RemoveWithValues -S 0.0 -C 12 -L first-last M". After this step, there are only 301 instances remaining.
- Step 2: continue to select attributes to match missing values until there are no missing values in the instances. I select attribute **thal** next, note that since it is not numeric value, all its instances will be selected if the split point is set as 0.0, thus we need to set the **invertSelection** as true to invert the selection. Finally, there are 297 instances remaining.

(ii) Past a screenshot of the first 10 rows of this clean dataset.

Answer: It is shown in Figure 6.

Figure 6: First 10 rows of clean dataset

```
47,male,asympt,150,226,f,normal,98,yes,1.5,flat,0,reversable_defect,>50_1
63,male,typ_angina,145,233,t,left_vent_hyper,150,no,2.3,down,0,fixed_defect,<50
67,male,asympt,160,286,f,left_vent_hyper,108,yes,1.5,flat,3,normal,>50_1
67,male,asympt,120,229,f,left_vent_hyper,129,yes,2.6,flat,2,reversable_defect,>50_1
37,male,non_anginal,130,250,f,normal,187,no,3.5,down,0,normal,<50
41,female,atyp_angina,130,204,f,left_vent_hyper,172,no,1.4,up,0,normal,<50
56,male,atyp_angina,120,236,f,normal,178,no,0.8,up,0,normal,<50
62,female,asympt,140,268,f,left_vent_hyper,160,no,3.6,down,2,normal,>50_1
57,female,asympt,120,354,f,normal,163,yes,0.6,up,0,normal,<50
63,male,asympt,130,254,f,left_vent_hyper,147,no,1.4,flat,1,reversable_defect,>50_1
53,male,asympt,140,203,t,left_vent_hyper,155,yes,3.1,down,0,reversable_defect,>50_1
57,male,asympt,140,192,f,normal,148,no,0.4,flat,0,fixed_defect,<50
```

(d) Data preparation: transformation

(i) **Attribute construction** - for example adding an attribute representing the sum of two other ones. Which **Weka filter** permits to do this?

Answer: AddExpression filter.

(ii) Normalize an attribute. Which **Weka filter** permits to do this? Can this filter perform Min-max normalization? Z-score normalization? Decimal normalization? Provide detailed information about how to perform these in Weka.

Answer: **Normalize** filter can be used to normalize an attribute. It **can** perform min-max normalization and decimal normalization, but **cannot** perform Z-score normalization. **Standardize** filter can perform Z-score normalization. For decimal normalization, we select filter **Normalize**, and set the scale as 2, set the translation as -1, this will create the normalize range [-1,+1]. For the other normalization, we can apply the filters directly.

(iii) Normalize all real attributes in the data set using the method of your choice - State which one you choose.

Answer: I select the filter **Normalize**, and using the default setting with scale set as 1 and translation set as 0.

(iv) Save the normalized dataset into heart-normal.arff, and paste here a screenshot showing at least the first 10 rows of this dataset.

Answer: It is shown in Figure 7.

Figure 7: First 10 rows of heart-normal.arff

```
0.375,male,asympt,0.528302,0.228311,f,normal,0.206107,yes,0.241935,flat,0,reversable_defect,>50_1
0.708333,male,typ_angina,0.481132,0.244292,t,left_vent_hyper,0.603053,no,0.370968,down,0,fixed_defect,<50
0.791667,male,asympt,0.622642,0.365297,f,left_vent_hyper,0.282443,yes,0.241935,flat,1,normal,>50_1
0.791667,male,asympt,0.245283,0.23516,f,left_vent_hyper,0.442748,yes,0.419355,flat,0.666667,reversable_defect,>50_1
0.166667,male,non_anginal,0.339623,0.283105,f,normal,0.885496,no,0.564516,down,0,normal,<50
0.25,female,atyp_angina,0.339623,0.178082,f,left_vent_hyper,0.770992,no,0.225806,up,0,normal,<50
0.5625,male,atyp_angina,0.245283,0.251142,f,normal,0.816794,no,0.129032,up,0,normal,<50
0.6875,female,asympt,0.433962,0.324201,f,left_vent_hyper,0.679389,no,0.580645,down,0.666667,normal,>50_1
0.583333,female,asympt,0.245283,0.520548,f,normal,0.70229,yes,0.096774,up,0,normal,<50
0.708333,male,asympt,0.339623,0.292237,f,left_vent_hyper,0.580153,no,0.225806,flat,0.333333,reversable_defect,>50_1
0.5,male,asympt,0.433962,0.175799,t,left_vent_hyper,0.641221,yes,0.5,down,0,reversable_defect,>50_1
```

(e) Data preparation: reduction
How to perform sampling with Weka filters? Can it perform the two main methods: **Simple Random Sample Without Replacement**, and **Simple Random Sample with Replacement**. Report your findings.

Answer: Filter **Resample** can perform sampling. It can perform both of these two main methods. The setting **noReplacement** is set true when perform sampling without replacement, and set false when perform with replacement. The attribute **sampleSizePercent** is used to set the size of sample.

3. For the Clevelandheart data, use Weka to perform Principal Component Analysis

(a) Keep only attribute(in the new space) that represent up to 70% of the total variance

Answer: Apply "PrincipalComponent" filter, and set *variance covered* field as 0.7

(b) Remove the rest of the attributes and save the resulting data set.

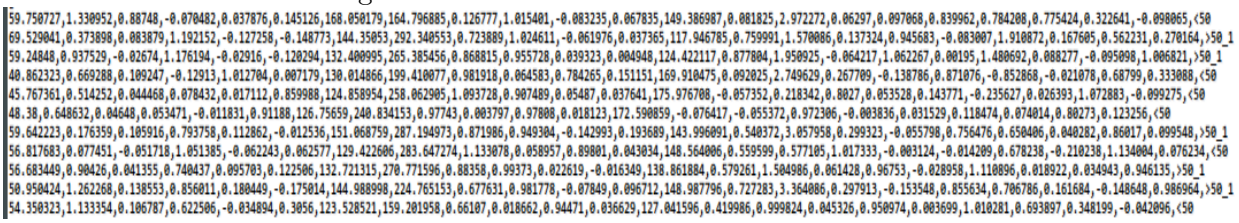
(c) Reconstruct the approximate original dataset.

Answer: We use **"Select attributes"** option, and use **PrincipleComponents** filter, and set standard deviation as 0.7, and set **"transformBackToOrignal"** as true. The first 10 rows of reconstructed data are shown in Figure 8.

(d) Compute the distance (Enclidian) between every vector in the original Cleveland set and the approximate one. Report the average and standard deviation of the set of distances. Report the three records with highest distance between approximation and the original.

Answer: Let X_1 be the vector of the original data, and X_2 be the vector of the approximate one. Here X_1 is the reconstruct dataset that represent 100% of the total variance. Then we compute the Euclidean distance between each corresponding row vector. We have:

Figure 8: First 10 rows of reconstruct data set



- average distance: 35.7906245546
- standard deviation:20.9852626932
- the three records with highest distance: 175.71812427, 134.38525847,115.11248262.